遥感影像特征发现的稳健统计模型研究

骆剑承 周成虎
(中国科学院地理研究所信息室，北京 100101)
马江洪
(西安交通大学数学系，西安 710049)

摘 要 高斯混合密度降解模型(GMDD)是一种基于稳健统计理论的层次结构的聚类模型。GMDD 首先假设特征空间是由一组混合的高斯(Gaussian)分布组成，然后通过一定的优化算法来获得特征空间中与预先假设相符合的特征分布，并逐步分离，直到特征空间全部降解为一组混合特征模式的分布集。GMDD 与传统统计聚类模型相比较，主要优点为：特征类别不受限定、抗干扰力强、参数估计与初始无关、考虑密度分布的可变性等。文中初步探讨基于 GMDD 方法的遥感影像特征发现模型(GIFEM)，并提出基于遗传算法的 GMDD 优化模型。

关键词 稳健统计 高斯混合密度 影像特征 遗传算法

0 引 言

几十年来的遥感地学决策分析的综合应用研究，使多平台、多时相、多光谱的遥感影像数据成为“数字化地球”中时空多维数据库的重要信息来源，这些数据和信息本身就反映了地球表层系统中自然和社会活动的双重作用，许多客观地反映地学现象和地学过程的地学知识就蕴涵在以遥感影像数据为代表的海量空间数据集中。地学空间知识发现(GKDD)，就是通过对空间数据进行分析处理，从空间数据库中发现地学知识，为地学决策分析提供更快速、丰富、精确、客观的知识来源。90 年代中期兴起的数据挖掘技术已经在地学知识获取、多波段遥感影像分类和信息提取等方面取得了一定的成功[1-3]。

数据挖掘(Data Mining)是指通过从大型数据库或数据堆中，搜索潜在的模式，去发现预见性的信息，反过来助于决策分析。特征提取属于“知识发现”或数据挖掘的范畴，指从复杂的空间数据集中搜索或抽取隐含的地学特征和模式[4,5]，其目标为：(1)找出模式之间的相关性；(2)将特征空间数据集划分为有意义的子集；(3)对数据集中的模式进行概念描述；(4)通过数据分析发现异常；(5)从数据集中找出规律性来建立预测模型等。

数据挖掘算法，是建立在各种传统方法基础之上的，如相关分析、聚类分析、偏差检测、决策树、最近邻方法、基于规则推理等，而近年来，以仿生物技术的神经计算理论和演化计算理论为基础的人工神经网络(ANN)方法和遗传算法(GA)在数据挖掘的应用中取得了重大进展[4]。ANN 是通过实例来训练网络，使网络获得分布式、并行存储的知识；而 GA 是模拟生命进化机制，以一定的寻优准则，从大量数据集中寻找最优特征。

空间数据挖掘一般是建立在一定的地学模型基础上，根据特征空间的分布规律，一次性地对空间数据进行特征提取或分类。但是，由于空间数据的复杂性和不确定性导致空间数据集之间互相重叠或者由于特征之间的相互干扰，常规数据挖掘方法难以获取细节性和过程性的特征分布结构，直接影响了分析结果的精度和对特征的解释能力。高斯混合密度降解模型(GMDD)是一种基于稳健统计理论的层次聚类方法，其分布模型是假设特征空间是由高斯(Gaussian)密度混合分布组成，通过一定的优化算法逐步降解特征，最后形成一组混合特征模式组成的密度分布集[6,7]。本文是在 GMDD 模型基础上，对空间数据中的特征进行分层提取，提出用遗传算法进行 GMDD 的空间搜索的优化模型，并从遥感影像数据中进行了特征发现的实例分析。

收稿日期：1999-10-27

1　高斯混合密度降解模型

1.1　稳健统计

稳健(Robustness),又称鲁棒性,是指具备一定抗干扰的能力。稳健统计理论(Robust Statistics)是近年来发展起来的一个新的统计学理论。Huber 认为一个稳健系统应该具有 3 个特点[7]:(1)必须有一个比较合理的、高效率的假设模型;(2)模型中出现的小偏差对系统总体性能影响不大;(3)模型中出现大的偏差也不会导致整个系统崩溃。系统崩溃点的定义为:

$$\varepsilon_N^*(T,Z)=\min\left\{\frac{M}{N}:bias(M;T;Z)>\zeta\right\}$$

其中,Z 表示有 N 个数据的点集;T 为估计者;M 表示 Z 中的偏离点(或"坏点")的个数;$bias(M;T,Z)$ 表示系统的偏离程度;ζ 为达到系统崩溃的临界值。理论上,要求一个稳健系统的崩溃点达到 0.5。稳健统计理论在本质上与模糊理论是相互关联的[7],其"尺度"、"权"等特征与模糊集理论中"隶属"是相同的。典型的稳健统计模型有最小体积椭球模型(MVE)、协同稳健估计模型(CRE)、随机概率最小化模型(MINPRAN)、高斯混合密度降解模型(GMDD)等。

1.2　GMDD 概述

在特征空间中,样本的分布是很复杂的,有时并不能用统一的一个参数化分布模型去描述整个空间中的各特征集的分布。混合密度降解模型(Mixture Density Modelling and Decomposition, MDD),属于一种稳健统计模型的聚类模型,从涵义上表达为对特征空间数据集,通过空间搜索逐次寻找最匹配分布特征,而逐步降解,最终获得空间数据集的混合密度分布。如图 1 所示,MDD 的基本过程为:空间数据集 A,每一步通过确定某一类特征的分布函数 F,在空间中挖掘出最符合 F 分布的数据集 X,并从 A 中排除 X,直到 A 都归属为不同空间分布的混合类别 C。

实际上,很难用一个参数化模型来准确定义特征分布。高斯混合密度降解模型(GMDD)是在 MDD 模型基础上,假设特征空间是由一组混合的高斯密度(Gaussian)分布组成的,然后通过梯度下降的迭代算法搜索空间,来获得特征空间密度分布中与预先给定分布最相似的点集,并逐步分离,直到

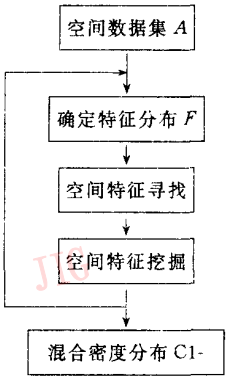

图 1　MDD 基本算法流程

特征空间全部分解为一组由混合高斯密度分布组成的混合特征密度分布集。GMDD 方法与其他统计聚类模型相比较,其统计稳健性的优点包括:(1)特征类别不作限定:不需要预先知道所属类别的个数,通过层次结构的挖掘,加入领域知识,逐步获得类别数;(2)抗干扰能力强:特征分布假设为参数待定的高斯密度分布函数(GPDF),通过空间最优分布搜索获得特征点集,只要存在相似分布点集,就能排除干扰,提取特征模式;(3)参数估计与初始条件无关:特征分布的 GPDF 函数参数预先不确定,与初始状态无关;(4)考虑密度分布的可变性等:在搜索特征模式的过程中,密度分布函数的参数在动态变化,以使特征点集与该分布最相符,因此整个点集的密度分布是混合的,而不是固定的分布。

1.3　GMDD 稳健特征估计函数

GMDD 是一种参数化逐步优化模型,通过搜索空间最优分布的参数,来挖掘该属于密度分布的特征。首先,需要确定要挖掘的特征估计形状函数,这些形状函数都起源于高斯密度分布函数(GPDF)。设 X 为 N 维的特征向量数据集,假设样本 $x^k\in X$ 属于高斯密度分布 $G(x^k,T)$ 的概率为 $\varepsilon(0<\varepsilon<1)$,则属于其他分布 $H(x^k,T)$ 的概率为 $1-\varepsilon$。整个分布函数 f 可由以下形式确定:

$$f(x^k,T)=\varepsilon\cdot G(x^k,T)+(1-\varepsilon)\cdot H(x^k,T)$$

其中 T 表示该特征的高斯密度分布的待定参数集。搜索空间确定最优化的条件是使得 X 集中,各向量的分布值 f 的对数总和 Q:

$$Q=\sum_k \log f(x^k,T)$$

为最大,此时的参数集 T 所表示的分布形状代表了所搜索的特征。

特征空间中基本形状可假设为普通高斯密度分

布的混合,许多复杂的分布或者看似不可能用参数化表达的不规则分布,就可以用参数化高斯密度分布函数混合而成。普通高斯密度分布可表达为:

$$f(x^k)=\frac{(1-\varepsilon)}{(\sqrt{2\pi})^n\sqrt{|C|}}\exp\left(-\frac{1}{2}d^2(x^k)\right)+\varepsilon h(x^k)$$

其中 $d^2(x^k)$ 为马氏(Mahalanobis)距离的平方。C 为协方差矩阵。

$$d^2(x^k)=(x^k-\boldsymbol{m})C^{-1}(x^k-\boldsymbol{m})$$

搜索的参数集 T 为中心点向量 $\boldsymbol{m}$。

1.4 GMDD 搜索算法

GMDD 空间搜索,实质上是空间优化搜索过程,即求参数化分布模型中的参数集 T,其约束条件是使 Q 为最大。传统的优化算法需要通过求导的微分迭代算法,其过程非常复杂,而且容易陷入局部极小。特别对于复杂特征的分布,由于融合了领域知识,用迭代优化算法,就难以获得最优解。

遗传算法(GA)模拟了自然界中进化和遗传发生的繁殖、交配和突变等现象。GA 从一初始种群开始,通过随机选择、交叉和变异等操作,产生一群新的更适应环境的个体,对应于特征空间中,是搜索到越来越优的区域。GA 利用了生物进化和遗传的思想,与传统的优化方法(包括微分法、枚举法、随机搜索法等)相比较,具有不受限制条件的约束、减少收敛于局部极小的可能、计算并行性等特点[8,9]。因此,用 GA 作为 GMDD 空间搜索算法,基本克服了传统优化方法可能带来的求解困难、局部极小、难以融合领域知识等缺陷。用 GA 算法进行 GMDD 空间优化搜索的过程如图 2 所示。

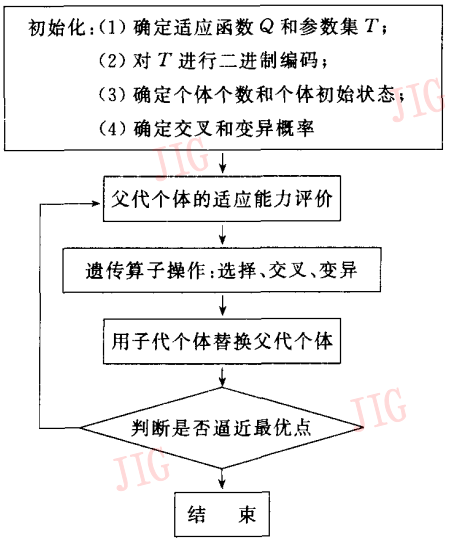

图 2 GMDD-GA 空间搜索优化算法

2 基于 GMDD 模型的影像特征发现

2.1 基于 GMDD 遥感影像特征分布估计的原理

在遥感地学分析模型中,由于遥感信息模糊和不确定性的特点,其特征空间中特征分布模型并不完全符合特定的高斯密度分布,而是分布形状各异、相互交错,难以用一个参数化分布模型来表达这种复杂的分布结构,用传统模型是难以获得特征的最优分布解的。基于 GMDD 影像特征估计模型(GMDD Based Image Feature Estimator Model——GIFEM),根据实际的地学分析模型来确定特征的可能分布,通过 GMDD 逐步降解,在特征空间中寻找到最符合的特征模式,获得特征空间中混合特征密度分布。GIFEM 特征分布模型假设特征空间是由一组混合高斯(Gaussian)密度分布组成的,通过 GA 优化算法进行空间搜索,获得特征空间中密度分布与预先给定的分布最为相似的点集组成的特征,并逐步从影像特征空间分离出去,直到影像特征空间全部降解为一组特征模式的混合密度分布集。

2.2 实例分析

试验区为香港一港岛地区,位于香港特别行政区的南部。该岛的北麓是香港主要城区,最高海拔为太平山,主要地物类型包括森林、草地、裸地、水体、居民地等,各类型间错综分布,十分复杂。选用的遥感资料为 Landsat-TM10 数据,成像时间是 1996 年 3 月 3 日,当时天气情况晴好。图象大小为 400 行×540 列。本次工作是在我们自行开发的遥感地学理解系统(GRS99)上完成的,运行平台为 Pentium Ⅱ 350。通过对研究区域实际情况的了解和对照土地利用图,进行了初步目视解译,把该区域大致分为以下 5 个大类的地物类型:

C1——水体,C2——居民地,C3——裸地,
C4——林地,C5——草地

主要工作原理是对每一个类别内的样本数据进行高斯混合密度分解,过程为:(1)确定类别 I 的特征空间分布 f_i;(2)用 GA 在特征空间中寻找服从 f_i 分布的最优模式 A;(3)从 I 类样本数据集 S 中分离属于 i 的样本集 B,使 B 不再参与特征寻找;(4)判别 S 中是否还存在一定分布的未归属样本,若还有:$i=i+1$,转(1);否则结束。对每一个类别进行以

上的混合密度分解,获得混合特征分布,最后以一定的表达方式来表示影像特征分布。

对照实际土地利用图和目视解译,分别选取了1000个样本数据,来进行土地覆盖类型的影像光谱特征的提取。最后通过规则形式来表达土地覆盖的光谱特征:

IF ((*PCA*123=36.49, 40.62, 48.87, 53.00{*CF*≥0.4}) and (*TM*4=0.00, 5.92, 21.68, 29.56{*CF*≥0.4}) and (*TM*5=0.00, 0.00, 22.51, 39.39{*CF*≥0.4}) and (*TM*7=0.00, 0.00, 9.30, 14.93{*CF*≥0.4})) THEN id is C1 Certainty is very_close_to 0.95

IF ((*PCA*123= 30.01, 37.89, 53.64, 61.52{*CF*≥0.4}) and (*TM*4= 5.59, 14.60, 32.60, 41.61{*CF*≥0.4}) and (*TM*5=0.00, 9.08, 46.60, 65.35{*CF*≥0.4}) and (*TM*7=0.00, 5.64, 27.40, 38.28{*CF*≥0.4})) THEN id is C2 Certainty is very_close_to 0.95

IF ((*PCA*123=42.20, 62.08,101.84,121.72{*CF*≥0.4}) and (*TM*4=29.78, 48.15, 84.91,103.29{*CF*≥0.4}) and (*TM*5=48.73, 85.48,159.00,195.75{*CF*≥0.4}) and (*TM*7=34.21, 52.59, 89.34,107.72{*CF*≥0.4}))THEN id is C3 Certainty is very_close_to 0.95

IF ((*PCA*123=29.62, 34.12, 43.13, 47.64{*CF*≥0.4}) and (*TM*4=22.29, 42.55, 83.05,103.31{*CF*≥0.4}) and (*TM*5=0.00, 18.99, 70.00, 95.51{*CF*≥0.4}) and (*TM*7=0.00, 1.22, 24.48, 36.11{*CF*≥0.4})) THEN id is C4 Certainty is very_close_to 0.95

IF ((*PCA*123=34.89, 43.90, 61.91, 70.91{*CF*≥0.4}) and (*TM*4=22.38, 33.25, 55.01, 65.89{*CF*≥0.4}) and (*TM*5=44.08, 66.97,112.73,135.61{*CF*≥0.4}) and (*TM*7=20.98, 33.74, 59.25, 72.00{*CF*≥0.4}))THEN id is C5 Certainty is very_close_to 0.95

.....

其中 *PCA*123 表示 *TM*1、*TM*2、*TM*3 的 KL 变换后的第 1 主成分。规则的前提部分采用模糊表达方式,如 *TM*=*a*,*b*,*c*,*d* 表示梯形的隶属关系。利用以上所获得的影像光谱特征,我们已经成功地应用于土地覆盖分类中[10]。

3 结　论

GIFEM 是在 GMDD 空间逐步搜索获得特征空间混合高斯密度分布的基础上,对多维遥感影像数据的特征空间进行挖掘和提取。GIFEM 方法具有很强的稳健性能,表现在:(1)预先影像特征类别可以未知或不限定;(2)具有一定的抗干扰的稳健力,不被周围的许多离散点干扰;(3)分布函数的参数估计与初始无关,是在搜索过程中自然获得;(4)密度分布具有可变性和多样性。除了高斯密度分布以外,其他分布的 MDD 模型也可根据同样的原理进行拓展,特别在分布模型中融合地学特征,使分布模型更接近实际地学分布规律,值得进一步地探索和实践。

参 考 文 献

1 Crowther P. Knowledge acquisition from satellite image for geographic expert systems —— show or tell? In: Proceedings of International Conference on Modeling Geographical and Environmental systems with Geographical Information Systems. 1997, 2:568～573.

2 Friedl M A, Brodley C E. Decision tree classification of land cover from remotely sensed data. Remote Sens Environ, 1997, 61:399～409.

3 Fischer M M, Leung Y. A genetic-algorithms based evolutionary computational neural network for modeling spatial interaction data. The Annals of Regional Science. 1998,32:437～458.

4 陈文伟编著. 智能决策技术. 电子工业出版社,1998.

5 Glymour C, Madigan D *et al*. Statistical Themes and Lessons for Data Mining. Data Mining and Knowledge Discovery. 1997,1:11～28.

6 Zhuang X H, Huang Y, Palaniappan K, Zhao Y. Gaussian Mixture Density Modeling, Decomposition, and Applications. IEEE Transactions on Image Processing. 1996,5(9):1293～1301.

7 Dave R N, Krishnapuram R. Robust clustering methods: A unified view. IEEE Transactions on Fuzzy Systems. 1997,5(2):270～293.

8 Holland J H. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, 1975.

9 Buckles B P, Petry F E. Genetic Algorithms. IEEE Computer Society Press,1994.

10 骆剑承,周成虎,梁　怡. 空间逐步寻优的数据挖掘法的多波段影像分类研究. 地球信息科学,1999,(1):52～59.

Robust Statistical Theory Based RS Image Feature Estimating Model

Luo Jiancheng, Zhou Chenghu

(*LREIS, Institute of Geography, Chinese Academy of Sciences, Beijing* 100101)

Ma Jianghong

(*Department of Mathematics, Xi'an Jiaotong University, Xi'an* 710049)

Abstract Gaussian Mixture Density Modelling and Decomposition (GMDD) is a hierarchical clustering method based on robust statistical theory. Firstly, GMDD is assumed with a mixture group of Gaussian distribution in feature space, then by optimization algorithm the feature which mostly accord with the assumed distribution is hierarchically extracted from space until all of the features in the space are decomposed to a group of featuring pattern. Compared with conventional statistical clustering methods, GMDD's main outstanding superorities are: (1) Initial number of features does not needed to be specified a priori; (2) The proportion of noisy data in the mixture can be large; (3) The parameters estimation of each feature is virtually initial independent; and (4) The variability in the shape and size of the feature densities in the mixture is taken into account. The article presents the model named the GMDD based remote sensing image feature estimation model (GIFEM), and the model of GA space searching optimization is also presented out.

Keywords Robust statistics, Gaussian Mixture Density, Image features, Genetic algorithm

(上接第 940 页)