数字信息模式识别理论与应用

史玉峰　靳奉祥　著

科学出版社
www.sciencep.com

数字信息模式识别理论与应用

史玉峰　靳奉祥　著

科学出版社
北京

内 容 简 介

测绘信息的分析处理一直是测绘科学技术领域研究的重点之一。本书从理论到应用全面地阐述了数据信息特性和数据信息特征提取方面的新颖见解，系统分析研究了数据压缩和特征提取的相关理论与方法，对信息模式识别的若干问题进行了深入的研究，突出强调了特征有效性和模式信息量。主要内容包括：模式识别和信息论的基础理论、信息模式相似性测度、基于统计理论的数据压缩与特征提取、基于投影寻踪的特征提取与模式识别、基于独立分量分析的特征提取和模式分离、数字模式识别理论的扩展与应用。

本书可供从事地学信息模式识别、测绘数据分析处理等信息科学领域的科研人员、高校教师和研究生参考。

图书在版编目(CIP)数据

数字信息模式识别理论与应用/史玉峰，靳奉祥著．—北京：科学出版社，2007

ISBN 978-7-03-020050-1

Ⅰ．数… Ⅱ．①史…②靳… Ⅲ．数字信号-模式识别 Ⅳ．TN911.72

中国版本图书馆 CIP 数据核字（2007）第 147339 号

责任编辑：牛宇锋 于宏丽/责任校对：陈玉凤

责任印制：刘士平/封面设计：耕者设计工作室

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂印刷

科学出版社发行 各地新华书店经销

*

2007 年 9 月第 一 版 开本：B5（720×1000）

2007 年 9 月第一次印刷 印张：13 1/4

印数：1—3 000 字数：254 000

定价：30.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

前　言

目前，以遥感、摄影测量和全球定位技术为主要手段的测绘数据信息获取技术已形成了覆盖全球的监测运行系统，成为快速获取和更新地球数据的主要技术手段。我们每天都可以获得大量甚至海量的数据。我们在享受科学技术进步的同时，也面临如何快速、自动、有效地分析处理这些已获取的大量的、高维的数据信息的严峻挑战。在实际工作中，面对研究对象和实际的需要我们可能要回答更多的问题，例如，数据中拥有多少类信息？各类信息之间有怎样的关系？每一类信息对对象的影响程度有多大？哪一种为主要的信息？以及数据中有没有所需要的信息等具有重要意义的问题？所有这些问题都需要进行一些定量的数据分析研究，需要从信息的角度去研究数据和分析数据，总结数据信息特征的性质，以实现信息的挖掘、度量、描述和利用。基于上述问题，本书作者以统计理论、信息论和模式识别理论为理论基础，从特征有效性和模式的信息量角度研究了测绘数据的信息特征。

本书是作者近年来在国家自然科学基金，山东理工大学科技基金资助下的部分研究成果的总结，同时也参考了国内外一些学者在该领域的研究成果。全书共分 7 章，第 1 章对模式识别理论、信息理论的应用做了概述，提出了本书主要研究内容；第 2 章简要阐述了模式识别和信息论的基础理论；第 3 章基于信息理论，对信息模式相似性测度进行了系统的研究，提出了交叉距离测度、信息距离测度等信息测度；第 4 章研究了基于统计理论的数据压缩与特征提取方法，分析了主成分分析的信息特性；第 5 章讨论投影寻踪理论及其应用，研究了基于投影寻踪的高光谱遥感数据特征提取与识别方法；第 6 章介绍了独立分量分析原理、准则函数、优化判据等独立分量分析基本理论，研究了独立分量分析在测绘数据特征提取和模式分离与识别中的应用；第 7 章研究了基于模式识别理论的母体均值变化判识方法，应用信息熵来表征测量结果的不确定度和识别观测结果中粗差的方法，最小描述长度原理及其应用，最大熵原理及其在空间数据分布模型建立中的应用，基于对称交叉熵的信息特征提取和基于互信息的特征提取等。

本书得到了徐州师范大学张莲蓬教授、中国矿业大学丁世飞教授、山东科技大学独知行教授等的大力支持，他们给予了许多指导与帮助，在此表示诚挚的谢意。

由于作者水平所限，书中难免有疏漏之处，敬请读者不吝指正。

史玉峰

目　录

第1章 绪 论

1.1 问题的提出

人类和自然是和谐统一的。在漫漫的历史长河中，人类学会了认识和利用自然；人类将认识事物的手段作了细致明确的分工，形成了众多的学科，建立了相应的理论体系和研究方法。经过长期的研究与实践，人们发现自然界的变化有着惊人的规律和秩序以及高度的组织性和系统性，它像一个有机生命体，内部的器官间有丰富、有序的信息传递，同时它与外部还有着信息交换和对外部信息的反映。所以这些为人类认识世界和考察事物提供了可能的信息来源。在如此宏大和复杂的信息集合中辨识事物现象与其本质间的关系、现象与现象之间的关系是一项十分有意义的研究。现代社会里，物质、能源、信息是构成现代社会大厦的三大支柱。物质是构筑社会的基础，能源是构筑社会的动力，而信息是构筑社会的神经系统。信息的重要性已经被人们所认识，信息理论也已经被广泛地应用到军事、医学、社会学、经济学、工业和农业等各个领域。信息科学的最新发展表明，建立在概率论基础上的 Shannon 信息论，只着重表达了信息的传递，但难以表达数据信息本身的涵义。而信息科学不仅要研究数据信息“量”的问题，更重要的还在于数据的信息的特征及信息的定性问题。这就涉及数据信息的提取、描述、推理、判断和决策等富有挑战性的处理工作。

目前，以遥感、摄影测量和全球定位技术为主要手段的测绘数据信息获取技术已形成了覆盖全球的监测运行系统，成为快速获取和更新地球数据的主要技术手段。我们每天都可以获得大量（甚至海量）数据。我们在享受科学技术进步的同时，也将面临如何快速、自动、有效地分析处理这些已获取的大量的、高维的数据信息的严峻挑战。戈尔在题为“数字地球”的演讲中清楚地描述了这种挑战，“充分利用这些浩瀚的数据的困难在于把这些数据变得有意义——把原始数据变成可理解的信息。今天，我们经常发现我们拥有很多数据，却不知如何处置。”

当我们面对大量采集到的测绘数据信息，在没有应用有关数据信息的分析方法（如各种统计方法）去处理它们时，这些数据信息对于我们来说，其所包含的信息量等于零；当在对所获得的数据进行预处理后，这些数据就转化为可被人们利用的各种数据格式的信息，但此时我们对它所蕴含的信息量的挖掘与提取是有限的，这些信息需要被继续分析（如进入地理信息系统）。在对信息的再处理过

程中，我们提取的信息量逐渐增高，信息的使用价值随之增大；特别地，当这些数据信息经过一些特殊处理信息系统的处理后，可以大幅度地提取其蕴含的信息，信息的使用价值增加更快；再经过辅助决策支持系统，这些信息就可以进一步地转化为更高级的信息形态——知识。把信息提炼成知识并把知识激活成智能，是信息学的核心和灵魂（钟义信，2001）。因此，信息的提取至关重要，也十分必要。

针对信息的提取问题，国内外许多学者（Kullback，1959；Guiau，1977；Sydenham，1982；Rissanen，1978；Ferraro et al.，2002；Torkkola et al.，2000；Friedman et al.，1974，1981；Fano，1961；Principle et al.，2000；李德仁等，2001；2002；靳奉祥等，1999；靳奉祥，2003；王晋年等，1999；何国金等，1999；李世中，1999；吴小培等，2001a、2001b；邸凯昌，2001；张学工，2000，刘桂雄等，1999；冯国章等，1998）作了许多研究工作。概括起来，他们的研究基础可以分为：统计分析理论、模式识别理论、信息理论。他们的研究集中在对数据集特征的分析、数据特征的提取、数据描述模型的建立等方面，以实现数据压缩、特征提取和模式分类的一般性目的。

在实际工作中，面对研究对象和实际的需要我们要回答更多的问题，例如，数据中拥有多少类信息？各类信息之间有怎样的关系？每一类信息对对象的影响程度（即信息量）有多大？哪一种为主要的信息？以及数据中有没有所需要的信息等具有重要意义的问题？正是这些问题的解决才使得数据产生知识，并指导着我们的决策。针对具体的大地变形分析问题，我们不能仅仅确定大地是否发生了变形，我们还要回答变形的方式（或模式），以及哪一种模式占主导地位，各种模式对变形的影响程度（或贡献）。所有这些问题都需要进行一些定量的数据分析研究，需要从信息的角度去研究数据和分析数据，总结数据信息特征的性质，以实现信息的挖掘、度量、描述和利用。

1.2 数据信息特征的涵义与研究内容

数据和信息是当今社会中广泛使用的术语，在不同的领域，站在不同的角度，人们对这些术语的内涵和外延有着不同的认识。

根据测绘数据的特点，我们从信息科学的角度给出数据的定义为：以时空数据为载体，以客观事物的属性数据为信息的数据集合体。它既可以是某一事物本身历史发展的积累，也可以是多个事物在某一时空的状态表达。具体地讲，这个事物可以是一个角度、一条边长、一点的高程、一像元的灰度，也可以是一幅遥感图像或是整个地球等。

信息是事物属性和变化规律的表征，通过它我们可以了解事物的存在和变

化；它的最基本形式是信号，它是可以被考察和测量的量，其测量的结果构成了一个完备的数据集合，这个集合的数据蕴藏着丰富的反映客观事物的信息；反过来讲，数据是信息的载体，有意义、有价值的数据才是信息。因此，我们需要研究数据的信息特征。

数据特征指数据本身所带有的、表明数据特性的、可以有效识别数据的一些信息，可以将其分为：物理特征、几何特征、数学特征和信息特征。数据的物理特征和几何特征是数据最基本、最直观的特征，通常我们直接利用这些特征来描述、判识数据，因为这样的特征易被视觉、触觉以及其他器官发现；数据的数学特征是指描述数据的参数，如概率密度函数、统计平均值、相关系数、方差-协方差矩阵、偏斜度、峰度、协方差阵的本征值和本征向量等，在利用计算机系统来识别数据时，我们通常用计算机来抽取数据的数学特征并进行模式识别；本书中所讨论的数据信息特征，指从信息论的角度，以信息量来描述数据特征的参数，如熵、互信息等，它是数据的数学特征的拓展和延伸。

数据特征的有效性是本书研究的一个重要问题。数据特征的有效性有两方面的意义，第一是数据压缩意义，即在不损失或少损失原始数据有用信息的条件下，来选择部分有效特征，而抛弃多余特征；第二是类别可分性意义，即所选择的特征相对于其他特征能够更有效地对数据集进行分类。

1.3 数据特征分析理论方法

人们对数据特征的分析由来已久，数据特征的分析方法归纳起来大致可以分为：统计分析方法、探索性数据分析方法、信息理论方法、模式识别理论方法、子空间方法和一些软计算方法。下面，简要讨论上述方法及其在数据特征分析中的应用。

1.3.1 统计分析理论方法

统计分析理论是常用的数据特征分析方法。统计方法有较强的理论基础，拥有大量的算法，可以有效地分析处理数据。多元统计分析中的回归分析、方差分析、主成分分析、因子分析、聚类分析等方法经常用于从数据中提取数据特征，而线性模型是我们在数据分析中用的最多的模型。

回归分析是从考察自变量和因变量间的关系入手，认为它们之间存在一种函数关系，应用最小二乘估计原理来建立它们之间的函数关系（陈希儒，1997；方开泰，1989；张金槐，1999；任若恩等，1996；Johnson et al.，1998）。回归分析中，在对自变量与因变量的关系不是十分清楚，或者对自变量间的关系不是十分清楚的情况下，常常采用逐步回归的方法，以选择跟因变量相关程度大的自变

量。回归分析中，自变量的筛选是一项十分重要的工作。

线性模型是我们经常使用的模型，众多学者研究了 Gauss-Markov 模型的性质、参数的估计和模型的描述等（张金槐，1999；陶本藻，1992；王松桂，1987）。在测绘领域，Gauss-Markov 模型也是我们应用最多、研究最深入的模型。靳奉祥等（1995）研究了 Gauss-Markov 的最大似然估计并将其做了必要的推广，得出如下结论：Gauss-Markov 模型未知参数的最大似然估计与最小二乘估计相同，均为无偏估计；方差的最大似然估计为有偏估计。靳奉祥等（1999）还以空间向量几何理论为基础，详细地分析了 Gauss-Markov 线性模型的几何性质和统计性质，以及它们之间的关系，在实现了模型图形化描述的基础上，利用各解向量的几何及统计性质构造出了不同情形下的统计检验方法和相应的检验统计量，为全面、深入地研究和灵活运用该模型提出了一种简洁直观的数学方法。归庆明（1997，2001）在分析根方估计、Stein 均匀压缩估计、Sclove 部分压缩估计各自特点的基础上，提出了一种新的有偏估计部分根方估计，并讨论了它的优良性质。周成虎等（2001）分析了传统遥感影像统计分析方法模型存在的缺陷，分别从传统的地学分析方法、神经计算模型、进化计算模型、稳健统计理论等几个不同的角度，提出了遥感数理统计分析模型扩展的方向。

主成分分析和因子分析是常用的数据特征分析方法，它们的主要作用是数据压缩和特征提取。主成分分析是从特征有效性的角度，将描述数据集的多个特征指标化为少数特征指标的一种统计方法。设有 n 个样本，每个样本测得 p 个指标，这样共有 np 个数据，然而指标间往往互有影响，主成分分析就是研究如何从 p 个指标中寻找出很少几个综合性的指标（主成分），利用对主成分的分析来达到我们的目的（陈希儒，1997；方开泰，1989；张金槐，1999；王松桂，1987；任若恩等，1996；Johnson et al.，1998）。这里要求主成分要尽可能多地反映原来资料的信息，而且彼此之间还应相互独立。因子分析的目的是，只要可能，就用几个潜在的但不能观测的随机变量去描述许多变量间的协方差关系，这些随机变量就叫做因子（任若恩，1996；Johnson et al.，1998）。因子分析模型是基于下述命题：可用变量间的相关性把它们分组。这就是说，在一个特定组内的所有变量，它们之间是高度相关的，而与不同组中的变量却有相对较小的相关性。实际上，可以把因子分析看成主成分分析的扩充，两者都可看作在力图逼近样本协方差矩阵。因子分析的主要问题是，数据是否与一个规定的结构一致。它们都是将描述数据的多个相关特征转化为少数几个不相关特征，其目的是使在高维空间中研究数据分布规律的问题，通过降维得到简化，并尽量保留原特征的信息量。主成分分析是通过变量变换把注意力集中在具有较大方差的那些主成分上，而将方差较小的主成分舍弃；因子分析是通过因子模型把注意力集中在少数几个不可观测的公共因子上，而舍弃掉特殊因子；它们都是基于数据样本方差-

协方差（相关系数）矩阵的数据特征分析方法。基于主成分分析方法，朱小鸽（2000）基于多重主成分分析，讨论了从遥感影像中提取有效信息的方法。廖明生（2000）分析了传统变化检测方法存在的不足，引进典型相关分析的基础理论，将不同时相的多通道遥感数据视为分组的多元随机变量，利用典型变换进行遥感数据的多元变化检测，提出的M变换方法用于多时相、多通道遥感影像的变化检测具有明显的优势和应用前景。

判别分析和聚类分析是直接利用数据信息来对数据分类的数据分析方法。判别分析问题用数学语言来说就是，有 k 个总体 $G_1,\cdots,G_k$，它们的分布函数分别为 $F_1(y),\cdots,F_k(y)$，每个 $F_i(y)$ 均是 m 维分布函数。对给定的一个子样 y，需要判断它来自哪个母体，这就是判别分析所要解决的问题。在判别分析中，通常按照一定的准则，将子样判给与其“距离”最近的母体，这是一种“有监督”的数据判识方法。

聚类分析是按一定的距离或相似性测度将数据分成一系列相互区分的组，以期从中发现数据集的整个空间分布规律和典型模式。它与判别分析的不同之处在于不需要背景知识而直接发现一些有意义的结构和特征，属于“无监督”数据判识方法。国内外许多学者对聚类算法进行了研究，提高了聚类分析的计算效率和聚类结果的可靠性。Murray 和 Estivill-Castro（1998）回顾了探索性空间数据分析的聚类发现技术。为了改善聚类分析中的分割算法，在聚类分析算法的基础上，NG 和 Han（1994）提出了随机搜索的改进 K _ medoid 算法，Ester（1995）用基于 R 树的数据聚焦法进一步提高了聚类的效率。层次算法将数据集分解成树状图子集，直到每个子集只包含一个目标，可用分裂或合并的方法构建。概念聚类是分割算法的一种延伸，它用描述对象的一组概念取值将数据划分为不同的类，而不是基于几何距离来实现数据对象之间的相似性度量（Pitte et al.，1998）。概念聚类能够输出不同类以确定其属性特征的覆盖，并对聚类结果给予解释。Murray 和 Shyy（2000）在分布显示和空间数据挖掘中集成了属性和空间特征，提出了一种交互的探索性空间数据聚类分析技术。

上述统计分析方法在对数据特征进行分析时，都是从统计意义上分析数据特征或对数据子集进行分类的，没有从信息的角度审视数据特征、分析数据子集的信息含量、评价模型的可靠性与有效性；而且，统计分析方法常常需要对数据集做满足统计不相关假设，而这种假设很多情况下在高维测绘数据中难以满足。因此，为提高数据分析的可靠性和有效性，提高数据特征的应用效率，一方面要改进数据分析方法，应用探索性数据分析（exploratory data analysis）方法分析处理数据，另一方面需要对数据的信息特征进行分析，这样就要将传统的统计分析方法与现代的信息理论相结合，应用信息论的思想来分析数据特征的信息特性，提高数据的信息含有率。

1.3.2 探索性数据分析方法

传统的统计分析方法是建立在总体服从某种分布，比如正态分布这个假设的基础上的，采用的是所谓的证实性数据分析方法（confirmatory data analysis），即“假定-模拟-检验”的方法。但实际问题中有许多数据并不满足正态分布，需要用稳健的或非参数的方法去解决。不过，当数据维数很高时，这些方法都将面临一些困难：随着维数的增加，计算量迅速增大；对于高维数据，存在着高维空间中点稀疏的“维数灾难（curse of dimensionality）”，非参数方法也很难使用；低维时稳健性能好的统计方法用到高维时稳健性变差。因此，传统的证实性数据分析方法对于高维非正态、非线性数据分析很难收到很好的效果。其原因在于它过于形式化、数学化，难以适应千变万化的客观世界，无法找到数据的内在规律和特征，远不能满足高维非正态分布数据分析的需要。为了克服上述困难，需要对客观数据不作假定或只作极少假定，而采用“直观审视数据-通过计算机模拟数据结构-检验”这样一种探索性数据分析方法。探索性数据分析是相对传统的统计分析而言的，它不预先假设数据具有某种分布或具有某种规律，而是一步一步地、试探性地分析数据，逐步地认识和理解数据的特征。它采用的是动态统计图形和动态链接窗口技术将数据及其统计特征显示出来，直到发现数据中非直观的数据特征或异常数据（Friedman et al.，1974；1981；Huber，1985）。

投影寻踪（projection pursuit）是探索性数据分析方法中的一种，它是处理高维数据，尤其是高维非正态数据的一类新兴的统计方法。它包括两方面的内容：一是利用计算机图像系统，在终端上显示出数据在任何 1～3 维子空间上的投影，使用者可以通过观察图像找出有意义的即能揭示数据的结构或特征的投影；二是按照实际问题的需要，事先确定一种衡量投影是否有意义的指标（projection index，投影指标），然后把数据投影到低维（主要是一维）子空间上，在计算机上自动找出能使该指标达到极大（或极小）的投影（成平等，1986）。Kruskal（1972）最先将高维数据投影到低维空间，发现数据的聚集结构和解决化石分类问题。随后 Friedman 和 Tukey（1974）提出了一种用整体上的散布程度和局部凝聚程度结合起来的新指标进行聚类分析，正式提出了投影寻踪概念。David Landgrebe 和 Luis Jimenez 等（1995，1999）将投影寻踪技术用于高光谱数据分析，将高光谱数据投影到低维子空间，成功地提取了高光谱数据的特征。李祚泳等（1997，1998）将投影寻踪技术应用于回归分析方法中，在水文、气象、环境等方面建立了合理的回归分析模型。

独立分量分析（independent component analysis，ICA）是一种新的模式识别理论，也有些学者把它归属于探索性数据分析方法。它是由法国学者 Jeanny Herault 和 Christian Jutten 于 1983 年提出的，其最初的应用是来解决“盲源分

离（blind source separation)”问题（Comon，1994)。近几年，ICA理论得到快速发展。它的目的是为非高斯分布数据找到一种线性变换，这样成分与成分之间是统计独立的或者尽可能独立的。在统计理论中，两个变量相互独立的涵义是指它们的联合概率分布密度等于它们各自概率分布密度的乘积，即 $p(x_1,x_2)=p_1(x_1)p_2(x_2)$。为了方便，在ICA中通常不直接用上述等式作为变量的独立性判据，而是利用观测量之间的熵信息作为独立性判据。Giannakis提出基于三阶量的独立分量分析方法，Gaeta和Lacoume（1990）提出最大似然估计法的独立分量分析方法，Cardoso（1989）提出运用四阶统计量实现独立分量的分析方法。近年来，芬兰学者Aapo Hyvärinen和Erkki Oja（2002）又提出了快速独立分量分析算法等。

国内外关于ICA的理论与应用研究如火如荼，吴小培等（2001b，2001c）基于独立分量分析理论，研究了混合声音信号的分离，针对独立分量分析输出结果排序的不确定性，提出了一种结合小波变换的独立分量分析解决方案；他还分析研究了独立分量分析在序列图像处理方面的应用，提出了基于独立分量分析的运动目标检测新方法。杨福生等（2006）系统地讨论了独立分量分析在生物医学工程中的应用，采用独立分量分析技术去除信号中干扰和伪迹等。

1.3.3 信息理论方法

Shannon于1948年在贝尔实验室杂志上发表长篇论文《通信中数学理论》中首次提出信息理论是一种基于统计意义的、狭义的信息理论，该理论从概率论出发，建立了信息熵（information entropy)、互信息（mutual information）等概念，比较科学地解决了概率信息的测度问题，并通过定量的研究，提出了一系列关于信源和信道的编码定理，为通信技术提供了数学基础，对通信技术的发展产生了持久和深刻的影响。但是，狭义信息理论对除通信技术以外的其他领域理论指导意义不显著。所以，自从Shannon以来，人们对更广泛意义下信息理论的研究一直没有停止过。Jaynes（1957）提出了最大熵（maximum entropy）原理；Kullback和Leibler在1951年首次提出，后又为Shore等在1980年发展了的鉴别信息（又称Kullback-Leibler相对熵或Kullback-Leibler信息散度）及最小鉴别信息原理的理论，这些信息原理在信息的重建、估计和识别等领域中得到了非常成功的应用。Kolmogorov提出了关于信息度定义的3种方法，即概率法、组合法和计算法，Chaitin系统地发展了关于算法信息的理论，使信息的统计定义得以进一步推广并对非统计意义的信息给出了一种量度。信息的科学量度已经系统地发展成为信息处理的一种准则，这一准则在信息处理领域正逐渐取代代表功率的最小均方误差准则，而成为信息处理的一般理论基础。

信息理论作为应用统计方法的延伸，是研究通信系统中信息传递和信息处理

问题的科学，它规定信息是减少可能事件出现的不确定性的量度，信息量等于消除的不确定性的数量（姜丹，2000；孟庆生，1986；周炯磐，1983）。对于认识主体（人、生物或机器系统）来说，如果他（它）在接受信息后，一点确定性都消除不了，那么信息量最小（等于零），若所有的不确定性都消除了，则信息量为最大。

Shannon 所定义的信息熵是信息量的测度，由于其物理意义清晰，使用方便，得到广泛的应用。其应用主要集中在两个方向：一是利用信息熵的基本定义和扩展定义计算信息熵，或基于信息熵基本性质将其作为度量有关数据特征的指标；二是在自由或受限极大熵的假设条件下，分析推导出数据分布的结构特征或某一要素的分布特征并进行实际检验。

半个多世纪来，信息科学获得了巨大的发展，从狭义信息理论发展成广义的信息科学，已成为一门具有普遍指导意义的基础学科，并已经在生命科学、分析化学、机械学、物理学、医学、经济学和地学等应用学科中取得了丰硕的研究成果。人们开始采用崭新的信息科学方法论来研究高级事物的复杂行为，以物质和能量为中心的传统科学将逐渐让位于以物质、能量和信息为中心的现代科学（钟义信，1996）。

基于信息理论，众多学者在其所从事的领域进行了研究。基于信息理论，Wallace 和 Boulton（1968）首先应用最小消息长度准则（minimum message length）解决聚类分析问题；Rissanen（1978）发展了最小消息长度准则，提出更一般的最小描述长度（minimize describe length，MDL）准则，发展了统计推断理论。基于最小描述长度准则，Rissanen（1978）、Hansen（2001）、Davies（2001）和 Grünwald（2000）等研究了模型的选择以及应用等问题。

熵是信息理论中的重要概念，已在许多领域得到应用。熵作为水文系统复杂性的统计测度，能够从系统可达状态的宏观概率层次和状态内部微观层次上的变化共同描述水文系统复杂性（冯国章等，1998）；熵还可以解决水文系统概率分布推导和参数估计问题、水文水质站网布设评估问题等（王栋等，2001）。陈必红（1998）研究了观测过程中信息熵的变化，定义了非标准数，借助广义均匀分布的概念和非标准数，用概率密度函数统一描述所有概率分布，给出了统一的信息熵的定义，并利用信息理论知识，详细描述了在观测过程中观测主体的知识函数熵减现象。刘桂雄等（1999）应用信息理论，对机器人位姿精度进行了研究；把机器人的位姿传递过程视为信息传输过程，把机器人各运动参数、各关节和手臂以及手末端位姿分别看作信息理论信息传输模型中的信源、信道和信宿，而机器人各误差源则看做是信息传输模型中信道处的干扰，从而将机器人位姿信息传递过程同信息理论的信息传输模型完全等效起来，建立机器人位姿信息传输模型。张九龙等（2000）应用最大熵原理，最大限度地利用已有信息对未知分布来

作出最合理的推测，进而对待识模式进行判识。王海涛等（1998）采用概率化预处理方法，把多种前兆观测值时间序列转化为概率值时间序列，应用信息理论基本原理，计算由多种前兆观测项目构成的信源系统的合成信息熵。赵鸿（1996）研究了由信息理论定量计算两序列 S、Q 之间的互信息的方法；根据周期系统、混沌系统、随机系统的一维时间序列及其时间延迟序列间的互信息估算结果，分析了数据的可预测性，得出序列预测准确性的最大极限。根据以上三类系统的二维时间序列间的互信息估算结果，估价了数据所提供的信息的新颖性，从而可以避免一些重复性的工作所造成的浪费。白雪梅（2001）剖析了信息概念，指出信息与消息、信号、知识、情报及数据的区别与联系；揭示了熵与信息的关系，并给出熵的计算公式；还介绍了信息与熵在决策和相关分析中的应用。

靳奉祥（2003）从信息的角度讨论了 K-L 的信息表示特性，提出了用表示熵进行信息度量的计算思想。郭大志和范爱民（2001）提出了整体 GIS 数据质量指标的概念，用条件信息熵来描述数据集的整体质量，研究了基于误差熵不确定带的数据质量评价模型。周成虎和张健挺（1999）从信息熵的概念出发，认为地学空间子集划分产生的互信息源于子集划分，使得各子集的不确定性降低，并且子集间的差异性增大。因此具有最大互信息的子集划分方案代表一定的地学模式和地学规律，以此为基础，讨论了地学数据属性要素的子集划分产生多维属性关联规则，以及通过空间和时间的子集分割进行聚类的方法，并将其用于地学空间数据挖掘模型的建立。王新洲（1999）基于信息扩散理论，研究了适合小样本情况下估计概率密度函数的信息扩散理论，导出了参数估计的信息扩散估计法；该方法不仅具有良好的抗差性，而且具有方法简便、便于应用等优点，并将其应用于水准网平差（王新洲，2000）。寻找合适的窗宽，是扩散估计的最大难题之一，也是影响扩散估计精度的最关键因素之一。在深入讨论信息扩散估计的基本大样本性质的基础上，提出了最优窗宽的理论和算法，较圆满地解决了窗宽问题（王新洲，2000）；在此基础上，王新洲和游扬声（2001）将信息扩散原理的估计方法从一维推广到多维，提出了适用于 Gauss-Markov 模型的信息扩散估计方法，并对其性能做了大量模拟试验研究。何国金等（1999）阐述了卫星遥感数据的信息涵义，认为遥感信息包含物质信息和场信息两部分，通过对它们的深入理解，促进了遥感信息提取与识别方法的改进与提高。何宗宜等（1998）讨论了信息理论在地图制图中的应用。范爱民（2001）研究了 GIS 中误差熵不确定带模型。李大军等（2002）分析了 GIS 中点位不确定性的信息熵指标的确定方法。孙海燕（1994）系统讨论了熵与不确定度区间的关系，得到几种常用分布的密度函数与熵之间的关系。

1.3.4 模式识别理论方法

模式识别是运用已知信息，按一定的准则确定未知类别样本的类别属性，即把某一样本归属于多个类型中的某一个（傅京孙，1984）。数据特征，特别是数据信息特征分析是模式识别理论的重要组成部分。

模式识别技术起源于20世纪20年代，随着大规模集成技术的发展和人工智能的兴起，在20世纪60年代初迅速发展为一门学科（李金宗，1994）。由于识别活动是人类的基本活动，人们希望机器能代替人类进行识别工作，因此模式识别的理论和方法引起了人们极大的兴趣并进行了长期的研究，现已发展成一门多学科的交叉科学。这门学科涉及的理论与技术相当广泛，涉及多种数学理论、神经心理学、计算机科学、信号处理等（边肇祺等，2000）。从本质上讲，这门学科实际上是数据处理及信息分析；而从功能上讲，可以认为它是人工智能的一个分支。特别是近二十年来，随着计算机技术的发展，它所研究的理论和方法在很多科学和技术领域得到广泛的重视。

模式识别按其理论基础可分为统计模式识别（statistical pattern recognition）、句法模式识别（syntactic pattern recognition）、模糊模式识别（fuzzy pattern recognition）、人工神经网络模式识别（artificial neural network pattern recognition）和人工智能模式识别（artificial intelligence pattern recognition）等（孙即祥等，2002）。

统计模式识别 统计模式识别是发展较早、理论较完善、应用较多的模式识别方法。它是通过对大量样本的统计分析，选择抽取最有代表性的统计特征作为分类决策的依据。图1.1给出了统计模式识别的不同方法，统计模式识别现已形成了一个比较完整的理论体系。上述1.3.1节的统计分析理论方法和1.3.2节的探索性数据分析方法中的许多理论和方法都是统计模式识别的理论基础。

句法模式识别 统计模式识别方法是先抽取和选择模式特征，用一个特征向量来代表模式，然后进行识别，得到模式的分类信息。它已成功地解决了许多模式识别问题，但这类识别方法也遇到不少困难，在许多模式识别问题中，研究的模式或是非常复杂，或是类别很多，不是用简单的分类所能解决的。有时人们对于事物的完整识别不仅限于简单的模式分类，而且还应对模式的结构做较全面的描述。分析一个模式，并且产生它的结构描述问题是相当困难的，一些方法仍在探索中，目前比较成功的结构模式识别方法是句法结构模式识别方法。句法模式识别方法（傅京孙，1984）将对象分解为若干基本单元，这些基本单元称作基元；用这些基元以及它们的结构关系来描述对象，基元以及这些基元的结构关系可以用字符串或图来表示；然后运用形式语言理论进行句法分析，根据其是否符合某一类的文法而决定其类别；模式的识别过程是由句法分析来完成。

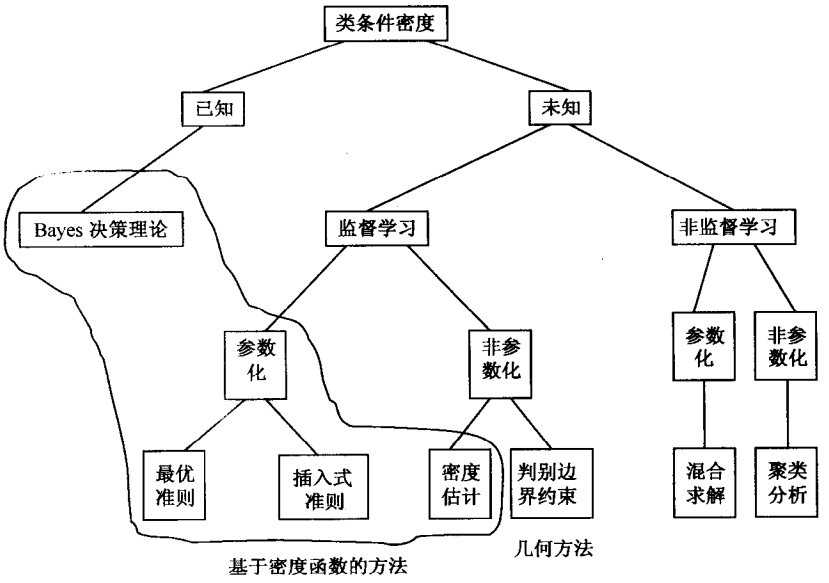

图 1.1 统计模式识别的不同方法

模糊模式识别 Zadeh 提出了著名的模糊集（fuzzy set）理论；该理论引起了数学界和科技工程界的极大兴趣并对其进行了深入广泛的研究，理论成果和应用成果不断出现，从而创建了一门新的学科——模糊数学（fuzzy mathematics）。将模糊技术应用于各个不同的领域，就产生了一些新的学科分支，如应用到人工神经网络中，就产生了所谓的模糊神经网络；应用到自动控制中，就产生了模糊控制技术和系统；应用到模式识别领域来，自然就是模糊模式识别、模糊聚类等。它与普通的模式聚类方法（硬划分）有许多“平行”之处，但也存在本质的差别，主要是概念的不同，将待识别类、对象作为模糊集及其元素，因此可以称为“软划分”，Zadeh 的模糊集理论为这种软划分提供了有利的分析工具，人们开始用模糊的方法来处理模式识别问题（Ruspini，1969）。利用模糊划分这一概念，人们开始用模糊的方法来处理模式识别问题，同时还有基于数据集的凸分解、动态规划方法等。总之，人们针对一些模式识别问题设计了相应的模糊模式识别系统；另一方面，对传统模式识别中的一些方法，人们用模糊数学对其进行了改进，这些研究逐渐形成了模糊模式识别这一新的学科分支。

人工神经网络模式识别 人工神经网络是模拟生物神经系统之间的复杂机理过程，由大量的称为神经元的简单信息处理单元构成的网络，整个神经网络的信息处理是通过这些神经元的相互作用来完成的。ANN 是一种动态信息处理系统，具有联想记忆、自组织、自学习和容错性等特征。从系统观点看，人工神经