几种有监督的模式识别方法及其程序包

马秀芳 王玉秀 王碧泉 编著

中国科学技术出版社

几种有监督的模式识别方法及其程序包

马秀芳　王玉秀　王碧泉　编著

中国科学技术出版社

内 容 提 要

本书主要介绍几种有监督的模式识别方法及其程序包，共分四章：第一章绪论，简介了几种有监督的模式识别方法的应用；第二章介绍了特征提取方法，CORA—3方法及其修改方法，Hamming方法及加权Hamming方法，以及控制试验等方法；第三章是一组配合上述模式识别方法的应用实例介绍；第四章是上述方法的程序包——PR程序包介绍。

本书所介绍的方法及其程序包已应用于地震危险区划、地震预测的研究中，实践表明也适用于其它科研领域。

本书可供地震学研究人员、其它科研领域中的有关人员与大专院校有关专业师生参考使用。

几种有监督的模式识别方法及其程序包

马秀芳　王玉秀　王碧泉　编著

责任编辑　周兆龙

中国科学技术出版社出版（北京海淀区魏公村白石桥路32号）

新华书店北京发行所发行　各地新华书店经售

北京三环印刷厂印刷

开本：787×1092毫米1/32　印张：2 7/8　字数：56千字

1986年10月第一版　1986年10月第一次印刷

印数：1—1500，册　定价：0.70元

统一书号：13252.1498　本社书号：1234

序

模式识别是近二十多年来发展迅速的一门学科，目前已广泛应用于文学识别、指纹辨认，细胞识别、疾病诊断、遥感技术、地震探测、天气预报、质量控制等方面。模式识别主要分为统计（或决策理论）方法和句法（或结构）方法，其中每一方面又包括了许多方法。例如在统计模式识别方面，各种无监督的集群(聚类)方法应用甚广；同时有监督的识别方法也日益发展。

目前，模式识别这一课题已受到人们极大的重视。无论在国外或国内，许多领域的科技工作者愈来愈多地将模式识别应用于本领域中，并将两者结合起来进行研究。其中模式识别在地震中的应用——“地震模式识别”即为一例。因此推广某些行之有效的方法及其程序包，使更多的非模式识别专业的科技人员能迅速掌握这些方法并应用于本学科的领域中，是有其积极意义的。

1984年12月，国家地震局地球物理研究所主持召开了地震界的第一届“地震模式识别交流讨论会”，到会者一致要求地球物理研究所举办学习班以推广有关方法。本书就是为这次学习班编写的。其内容主要是介绍几种有监督的模式识别方法及其程序包，现正式出版，希望能对各领域中有兴趣于模式识别研究和应用的科技人员在研究和应用这几种方法中有所帮助〔注〕。

限于我们的水平及时间仓促，不足之处，欢迎批评指正。

作者1985年6月

注：本书所介绍方法的相应程序包（PR程序包），原是在VAX机上运行的，其中多数程序已开发到PDP11/24机上运行，由于程序包的程序较长，不便在书中印出，欲购PR 程序包者，可与国家地震局地球物理研究所王玉秀、马秀芳联系。

目　录

第一章　绪论

第一节　模式识别简介

模式识别（Pattern Recognition）也称图象识别，粗略地说，它是用某些特征，对一组对象进行判别或分类。被分类的对象即为模式，分类的过程称为识别。

人们在生活实践中是经常进行模式识别的。例如要识别写在卡片上的数码字，判断它是0,1,2，…，9中的哪个数字，就是将数码字分成十类的问题。随着生产的发展和科学技术的进步，人们需要分类和识别的事物愈来愈多，被识别对象的内容也愈来愈复杂。这些对象（模式）可以是物理的实体：例如在鉴别癌细胞和正常细胞时它们是细胞的图象，在识别指纹时对象是某人的指纹；这些模式也可以是对物质对象或精神对象的描述：例如一张心电图或地震图可以按一定间隔采样为一组数值，对许多图形的诸组数值进行分类，通常用来诊断是否有心脏疾病或识别天然地震还是核爆炸。又如，当判断一个人是否有某种疾病时，还可以采用一组逻辑值（0或1）来描述疾病的各种症状：疼、晕、发烧等，有这种症状用“1”表示，没有这种症状用“0”表示。可见，虽然被识别的模式复杂了，但仍可归为“是”、“否”癌细胞，“是”“否”某一种疾病”，“是”“否”核爆炸等分为两类的问题。总之，许多事物都可以归为分成两类或多类的问题，从而进行模式识别。在地震学中，常需判定易发生强震的地点或预测强震发生的时间，这也可归为某一地区“是”或“否”为可能发生强震的地点，在某一时间段内

“是”或“否”发生强震的问题，即分为两类的模式识别问题。1980年出版的《国内外自动化发展动向报告集》中曾把这类“地震模式识别研究”作为广义模式识别的例子加以介绍。

人们是有模式识别能力的。例如，在马路上遇到几个熟人，一见面就知道那个人是谁。这是因为你脑子里有了每一个人的特点。在机器上最初是用模板匹配实现模式识别的。例如，要做一个能识别0—9共十个数字的机器，就要在机器中存放十个模板，要识别某一个数字，就需将该数字和机器中的模板进行比较，和哪个模板一样，就识别为那个数字。如果数字的大小或形状不一样机器就不能识别了，要想识别不同大小和形状的数字就要增加新的模板，这样模板会愈来愈多，机器就会愈来愈复杂。

解决这个问题的办法就是设法提取每一数字的主要特征，只要符合某些主要特征就识别为这个数字。正如人物漫画和照片并非完全一样，但一看就知道是谁，这就是抓住了人物的主要特征。提取模式的主要特征并用数字刻画出来，这就称为特征提取。每一模式可用一个维数与特征数相同的向量来描述。特征提取是一个很重要的问题，它对识别效果有直接影响。因此，必须研究所考察的模式，提取那些能描绘模式本质的特征。

一般说来，特征选得多一些较好，其优点是：（1）特征多则包含的信息量较大；（2）不必否定某些有争议的特征或某一专家的意见。但特征过多也会带来一些缺点：（1）计算量会大大增加；（2）各特征间的相关性可能增强；（3）当样本数一定时特征数增加到一定程度可能造成不稳定性。因此，有必要对特征作进一步的选择，使得最后用于识别的

特征总数减少，而又保留原来特征中的主要信息。这往往是利用原来的特征，通过一定方法淘汰掉一些特征，保留一些起主要作用的特征用以识别，例如Cora-3修改方法。或是用某些方法找出一些综合性的特征来（比原来的特征数目少），使这些综合性特征包含原有诸特征的主要信息，例如主成分分析法。这一过程称为特征选择。

进一步的问题，是要完成分类识别工作，这是模式识别工作要解决的关键问题。已述每一对象可用一个m维向量来描述，即对应于m维空间中的一个点。所谓将P个对象分成Q类就是把m维空间中的P个点分成Q个点集。"识别"就是要在m维特征空间中寻找合适的超曲面，使之能较好地分开这Q个点集。此超曲面称之为决策曲面。

现在可以用图1.1所示的流程图（取自〔1〕，略有修改）来说明模式识别的大致过程。图中上半部分是识别（或投票）部分，即对未知类别的模式进行分类；下半部分是分析（学习）部分，即由已知类别的模式，按某种识别方法得到判别函数及判别规则（有规则的学习），并用来对未知类别的模式进行分类识别。左边第二个框图是预处理过程。地震模式识别中，需要用地震目录、地质地貌特征和地震图等，这就需要进行地震选目、数据校正及地震图的数值化等，这些都属于预处理。此以外预处理中还包括数据的平滑、标准化等。图中右下角是自适应部分。即用训练对象经过学习得到判别函数，对训练对象进行检验，进一步改进判别准则直到满意为止。

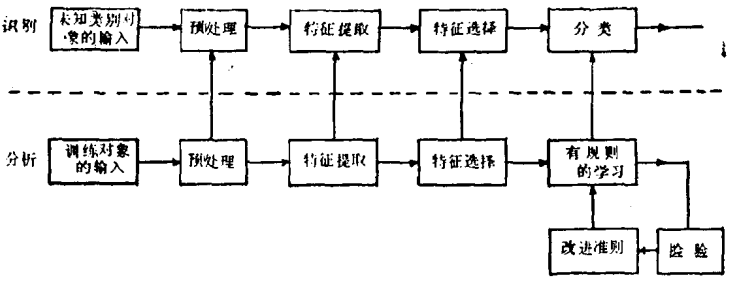

图 1.1 模式识别流程图

第二节 Cora—3和Hamming方法在地震学中的应用

60年代中期，苏联学者邦加德（(М.М.Бонгд)[2]就将模式识别中的Cora-3方法（俄文原文为“Кора-3”，为使与程序包中的名称一致，本书中采用英文译名Cora-3）应用于地学中，在勘探中分析地层。在地震学中的应用是始于1972年，由苏联应用数学研究所的格尔丰德(И.М.Гельфанд).和古别尔曼（Ш.А.Губерман)及地球物理研究所的著名地球物理学家凯利斯-博罗克(В.И.Кейлис-Борок)等人合作，将Cora-3方法应用于地震危险区的划分中，以识别未来可能发生强震的地点[3]。此后，他们发表了一系列论文〔4〕—〔12〕，其中文献〔4〕至〔11〕为一套论文，总标题是“可能发生强震地点的识别”，刊登在《计算地震学》(Вычислительная сйльмология)丛刊上。《模式识别应用于加州地震震中》一文[12]，可以认为是这一领域内经典的论文，该文为苏美地震预报合作研究的成果，苏方主要

参加者仍为上述的研究小组，美方参加者有著名地球物理学家诺波夫（L.Knopoff），前总统科学顾向普雷斯(F.Press)等人。文中报导了Cora-3修改方法，阐明了消除弱特征和等效特征的特征选择方法，从而形成了特征提取、特征选择、学习、投票的一整套模式识别算法——Cora-3及其修改方法，并完善了在地震区划中应用它识别潜在震源的方法。

我国在地震模式识别领域中的研究起步也较早，王碧泉等1979年起用此方法研究强震前的地震频度、相差半级的地震频度之比（b值）、以及它们随时间的变化等地震活动特征，用以预测强震发生的时间，并于1980年在自动化学会主持的全国第一届模式识别和机器智能会议上宣读了论文《大震前地震活动的图象识别》[13]。

目前，无论在国内或国外，Cora-3及其修改方法在地震学中已广泛应用，可大致概括为下列几个方面：

（1） 应用于地震区划，以识别潜在震源 这方面除前述苏美的工作外，我国马秀芳等[14]、张昇林〔注〕分别识别了北京附近及南北地震带中段易发生强震的危险区。国外也有阿卡亚(H、K、Acharya)[15]和卡普托(M.Caputo)等人[16]对中吕宋和意大利等地区的研究结果。

（2） 应用于强震发生时间的预测 前已述及，1979年开始王碧泉等[17]和马秀芳[18]等进一步研究了强震前的地震活动特征以及各种动力因子对强震的触发作用，金学申等〔注〕也研究了华北多个地震带的频度特征。1983年12月5日至16日在意大利举办的《模式识别及地震活动性分析学习班》上(Workshop on pattern recogntion and analysis of seismicity)，凯利斯一博罗克[19]和艾伦等(C、Allen)[20]也报告了他们用Cora-3修改方法预测强震发生

时间方面的一些结果。

（3） 应用于地震前兆的综合判断 王碧泉等[21]研究了16次强震的一组地震前兆—频度、b值、空区、地震的条带分布和地震的集中性等，定量地描述了这一组特征，并用Cora-3方法进行综合分析，以预测7级以上强震发生的时间及其震中。张郢珍等〔注〕则是综合分析形变、电阻率、地下水等几种前兆。

（4） 其他方面的应用中最突出的例子是普雷斯等[22]用此方法研究了钱德勒颤动、地震、地球自转和地磁变化之间的关系，并由此提出了构造运动激发钱德勒颤动的假设。

Hamming方法是继Cora-3方法后，由苏联学者格维希阿尼(А.Д.Гвишиани)等[23]引入地震学的，俄文原文为Хемминг，为与程序包中名称统一，本书中均采用英译名Hamming，在文献[23]中介绍Hamming算法，并论证了Cora-3、Hamming和Bayes方法在三个地区中识别强震可能发生地点之分类的统计意义。近两年来，我国吕宏伯、聂金宗、陈祖荫、马秀芳、王碧泉等人用Hamming和Bayes方法研究了北京及邻区易发震的地点[24]，并对原来使用二值化数据的Hamming方法进行了改进，使之适用于连续数据[25]，修改后的识别效果略优于原方法，而且避免了将原来连续的数据进行二值化而可能出现的信息损失。马秀芳等更进一步对比了Hamming、Cora-3和Cora-3修改方法的识别效果[26]，目前Hamming方法主要用于对危险地点的识别，但

注：见“地震模式识别交流讨论会论文汇编”，1984年12月4—8日，国家地震局地球物理研究所地震模式识别组印。

是毫无疑问它也可以用于其他问题的识别。

1983年以来，国家地震局地球物理所王碧泉等与北京工业大学陈祖荫等合作，研究并改进了各种集群方法，使之用于研究强震的孕震过程以预报强地震的发生时间[27]-[29]。可以预计，模式识别在地震学中的应用将日益广泛。囿于本书是应1984年12月举行的“地震模式识别交流讨论会”上与会者的要求，为推广Cora-3、Cora-3修改方法、Hamming方法及其程序包而编写的，因此，有关处理连续数据的Hamming修改方法，Bayes方法及各种集群方法均不在本书中介绍，各种方法在地震学中应用的结果也不在此叙及，读者可参考有关文献。

第二章　方　法

第一节　对象的确定和特征提取

一、对象（object）的确定

对象要根据需要解决的问题来确定。例如，在地震预测研究中这些对象可以是一系列时间段、一个个典型地震或一块块地区均可。记这些对象为E_1，E_2…E_n，若对象的类别未知，需对其中用于学习的对象预先进行分类。满足一定条件（或阈值）的称为D类（危险类），不满足该条件的称为N类（安全类）。下面举例说明：

例1：地震区划中，常需要研究某一地区是否列为易发生强震的危险地点。由图2.1中可见，强地震常常发生在断裂的交汇点附近。现取这些交汇点为研究对象，称为节点，如

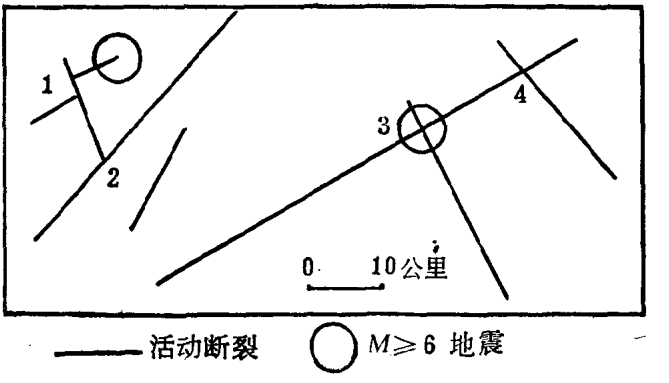

图 2.1　活动断裂、地震震中和节点示意图

图中编号所示。取下述条件事先将节点分类是合理的：节点附近（例如在半径为R＝10公里以内）曾发生M≥6地震，为危险类；否则为安全类节点。按此条件，图2.1中1号和3号为D类节点，2、4号为N类节点。实际例子可参看有关文献[3]、[12]、[14]。

例2：为预测强震发生的时间，可以ΔT为间隔将所研究的时间划分为若干个时间段，把这些时间段作为所考虑的对象。图2.2中ΔT＝3年，将1960年至1980年划分为7个时间段，并在相应的时间段中用圆圈标明了M≥7的地震。我们选取的条件是：某时间段内发生过一次以上M≥7的地震，则此时间段为D类，反之为N类。按此条件得到图2.2中第3、4、6、段为D类时间段，其余时间段为N类。具体实例参看文献〔13〕。

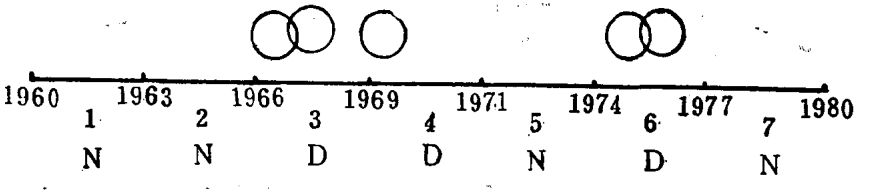

图 2.2 时间段的划分及其类别示意图

同样，也可能将某个M≥M₀的地震作为一个D类对象，并随机地选取一些“无震”作为N类对象，具体作法可参看文献[21]。

总之，首先需把所研究的对象按某种条件分成D类和N

类，如下表所示。

对　象	E_1	E_2	E_3	E_4	E_5 … Ek …… E_p
类　别	D	D	N	N	N… D N

为了便于研究，我们把D类和N类对象分别集中在一起，则上表可排成：

类　别	D类	N类
对　象	E_1 E_2… Ek	E_3 E_4 E_5…E_p

二、特征的提取

为了对研究的对象进行识别和分类，就要提取它们的特征（feature）。假定有m个特征，第j个特征用x_j表示（j=1，2…m），这样就形成一个描述某个对象的m维向量

$$X=(x_1, x_2, \cdots, x_m)^T \qquad (2.1)$$

对于每一个对象，一般说来，特征向量的元素xj可取任一数值，本书所介绍的方法中，x_j用“0”或“1”来描述，定义如下：

$$x_j=\begin{cases}1; & \text{当这个对象具备这个特征时}\\ 0; & \text{当这个对象不具备这个特征时}\end{cases}$$

这样就构成了表3.1所示的m×p矩阵，记为A矩阵，

通过以下步骤我们选取了要研究的对象，提取了代表这些对象特点的特征，并形成了m×p矩阵，用以进行识别，

表2.1　　A 矩　阵

对象 / 特征	D类					N类			
	E_1	E_2	…	Ek	……	E_3	E_4	……	Eq
x_1	1	1	…	0	……	0	1	……	0
x_2	1	0	…	1	……	1	0	……	0
⋮			⋮				⋮		
xm	0	1	…	1		0	0	……	1

第二节　Cor—3方法[2][3][12][13]

前已叙及，Cora-3方法是目前在地震学研究中广泛应用的算法之一，它是一种有监督的模式识别方法。即用一定准则，淘汰掉一些特征，并由保留的特征形成判别法规而进行识别的。此方法主要有两个步骤：学习和投票。

一、学习阶段

（1）构造新特征

为了刻画D类和N类的本质区别，先从特征向量

$$X=(x_1, x_2\cdots x_m)^T$$

出发，构造新的特征。对每个对象可定义三元（或多元）数组[13]：

$$\tau_0=x_q x_r x_l (q.\ r.\ l=1,2\cdots m) \qquad (2.2)$$

当q=r=l时，τ_0就与原特征相同；q.r.l不等时τ_0是原有特征的有序组合。这时，原特征是τ_0的子集：

$$\{x_j \mid j=1,2,\cdots m\} \quad \varepsilon\{\tau_0\}$$

由此可知τ_0的总数为

$$N\tau_0=C_m^1+C_m^2+C_m^3 \qquad (2.3)$$