MODERN DATA ANALYSIS

现代数据分析

◆ 吴今培 孙德山 编著

机械工业出版社
CHINA MACHINE PRESS

现代数据分析

吴今培　孙德山　编著

机械工业出版社

本书全面阐述了现代数据分析的基本理论和方法、实现技术和应用，内容反映了数据分析的最新发展。

本书共分8章，分别从数据的约简、分类、聚类、拟合、寻优、预测以及小概率数据挖掘等方面做了较详尽的论述，试图构建数据分析的一个新平台，以帮助读者掌握更高层次的数据分析技术，来解决更为复杂的实际问题。

本书可作为高等学校信息科学、计算机技术、控制工程及管理科学等相关专业的高年级大学生、研究生的参考教材，也适合于广大科技工作者学习与参考。

图书在版编目(CIP)数据

现代数据分析/吴今培，孙德山编著. —北京：机械工业出版社，2006.2

ISBN 7-111-18197-2

Ⅰ.现… Ⅱ.①吴…②孙… Ⅲ.统计数据-统计分析(数学) Ⅳ.0212.1

中国版本图书馆CIP数据核字(2005)第154436号

机械工业出版社(北京市百万庄大街22号 邮政编码100037)

策划编辑：吉 玲 责任编辑：吉 玲 版式设计：冉晓华

责任校对：申春香 封面设计：王伟光 责任印制：李 妍

保定市印刷厂印刷

2006年2月第1版第1次印刷

1000mm×1400mm B5·9.25印张·358千字

0 001—4 000册

定价：28.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

本社购书热线电话(010)68326294

封面无防伪标均为盗版

前　言

信息时代，数据如海！

然而，什么是数据？怎样科学地分析与有效地利用数据呢？

古代哲学思想认为：万物皆数。提出1是最神圣的数字。1生2，2生诸数，数生点，点生线，线生面，面生体，体生万物。因此，数产生万物，数的规律统治万物。

数学家认为：数是概念，不是物，是物的数量特征在人的头脑中反映为数，而不是数转化为物。数是用以表达世界上一切可以精确化、形式化的关系。

现代科学的观点认为：世界万物由三要素构成——物质、能量与信息。不是万物皆数，而是万物皆与数有关。

不是吗？一切实实在在的物质皆有形，形可以用数描述；运动与变化伴随着能量的交换与变化，能量可以用数表示；人的知识本质上是信息，信息可以用数记录。万物有质的不同，但质又可以用数刻划。

抽象的数只有大小没有量纲，当它们来自实验室中的测量、现场的采集记录时，就形成了有实际意义的数据。

随着数字网络的兴起与广泛应用，数据的来源越来越丰富，人们获得数据的代价越来越小，在众多领域都产生了海量数据，出现所谓数据过剩，而知识贫匮的局面。各行各业面对堆积如山的数据，人们有时会感到无所适从。为了把大量的数据作为信息，信息成为知识，知识付诸于应用，这就需要研究现代数据分析与推理的新途径和新方法。

在我们已经拥有收集、存储、查询数据的手段之后，理解这些数据，从数据中提取重要的特征；发现有用的知识，将巨大的数据资源转换为信息资源，以帮助我们进行科学的分析和决策，自然成为各行各业的迫切需要。然而，知识的发现和利用，将与数据的分类与聚类、归约与推理、模型化与最优化等问题相联系，我们称此为：从数据中学习。从数据中学习已成为当今知识工程领域的前沿学科，正吸引着

越来越多的科技工作者进行更加广泛而深入的研究。

本书较全面阐明现代数据分析与推理的基本理论和方法、实现技术和应用，内容的表达既简明扼要，立足于实用，又能涵盖本学科的要点。本书的部分章节，第一作者曾为北京交通大学、北京航空航天大学等相关专业的研究生讲授过，其中的大多数反馈意见也包含在书中。

全书共分8章，第1章为粗糙集与数据约简；第2章为支持向量机与数据分类；第3章为模糊划分与数据聚类；第4章为神经计算与数据拟合；第5章为遗传算法与数据寻优；第6章为时间序列分析与数据推测；第7章为现代数据分析的融合与协作技术；第8章为异常数据挖掘。读者通过目录可以了解全书内容的概貌和结构。

本书的问世，不只是作者的努力结果，而且还凝结了许多人的成果。作者在撰写过程中，参考和引用了大量的国内外有关文献和研究成果，在此对所涉及的专家学者表示衷心的感谢。

本书的出版得到许多人士的帮助，他们是常青教授、林湘桂副教授、马宁博士、郑晓星硕士以及机械工业出版社吉玲编辑等，对于他(她)们的辛勤劳动，作者表示深深的感谢。

感谢中国科学院院士戴汝为教授给予的关心与鼓励。

感谢北京交通大学副校长李学伟教授的热心帮助与大力支持。

最后特别指出，作为一本交叉学科的著作，需要不断面对新的知识，研究新的问题，但由于编著者学识水平有限，书中不当之处在所难免，我们真诚地希望读者不吝赐教，如果您能把宝贵意见与建议发往：jinpeiwu@sohu.com，我们将不胜感激。

吴今培

2006年1月

目　录

第 1 章　粗糙集与数据约简

1.1　不确定性理论

在自然界和人类的社会活动中所遇到的各种现象，大体上可分为两大类：确定性现象和不确定性现象。

所谓确定性现象是指，在一定条件下必然会出现的现象。例如，“同性电荷互相排斥”，“在标准大气压下，纯水加热到 100℃会沸腾”等等。

为了回答确定性问题，数学家提供了动力系统的方法，按照这个方法研究那些严格地遵从因果关系的系统，即所谓的确定性系统。

对确定性(或决定性)系统的数学描述，举一个大家熟知而又简单的例子，这就是迭代。如果某个系统服从确定性的因果关系，那么，它明天的状态 Y 与今天的状态 X 之间就有一个确定性的联系。在数学上，这叫做 Y 是 X 的函数，即

$$Y = F(X)$$

关系式既可表示今天和明天的状态之间的联系，也可以表示昨天和今天，明天和后天的状态之间的联系。如果 Z 是后天的系统状态，根据上面的关系式便有

$$Z = F(Y) = F(F(X))$$

而大后天的状态将是

$$W = F(F(F(X)))$$

一般来说，n 天之后的状态可以用函数 F 的 n 次迭代表示为

$$X_n = F^n(X)$$

而

$$\begin{cases} F^N(X) = F(F^{n-1}(X)) \\ F^O(X) = X \end{cases} \qquad (n = 1,2,3\cdots)$$

迭代运算是完全确定的。在计算机上作迭代特别适宜：一个固定了的计算机程序；给一个初始值；计算出的结果又当成初始值。反复多少次，完全不用人操心。因此，自从有了计算机，确定性系统的迭代模型引起了数学家的极大兴趣，同时迭代模型在自然科学、管理科学和工程技术科学领域获得广泛的应用。

为了研究不确定性现象，数学家提供了概率论和数理统计的方法，按照这个方法研究那些偶然性占统治地位的系统——随机系统。

所谓随机性现象，是指即使条件完全相同，事物的出现所产生的结果一般也不尽相同，或不完全确定和不能确切预言。

以抛硬币为例。掷一枚硬币出现正面还是反面是偶然的。即使硬币是绝对的均匀，我们也找不出任何理由断言它该出现正面还是反面。

当涉及人类系统的行为或处理与人类系统行为可相比拟的复杂系统时，确定性、随机性系统不再是十分有效的。例如，清早起来某人说一声“今天天气很好”，别人都能明白，可是计算机就不明白。因为“很好”是一个模糊概念，它不是简单的对事物的绝对肯定或否定，而是介于这两者之间，对于以二值逻辑为基础的计算机是无法理解的。由于人类思维的显著特点是具有模糊性，它是人类自然语言的本质属性，因此，这就给计算机和人脑之间造成了一条无形的鸿沟，使人机之间的接口受到阻碍。为了既能准确地表达大脑思维中的模糊性，又让计算机能够理解，把人们常常用的模糊语言设计变成计算机能接收的指令和程序(机器语言)，以便计算机能像人脑那样简捷灵活地作出相应的推理和判断，这就需要研究另一种不确定性——模糊性。美国控制论专家 L.A.Zadlh 教授于 1965 年建立了模糊集合和模糊逻辑的方法，按照这种方法研究那些模糊性占统治地位的系统——模糊系统。

在现实世界中存在着大量不完备、不精确的知识(或信息)，在知识不完备的情况下，人脑可以相应做出比计算机快速、精确得多的判断。但基于传统的知识处理方法，一旦遇到所给信息缺损或不完全时，其认知能力会急剧降低。为了处理具有知识不确定、不精确、不完善的系统，波兰数学家Z.Pawlak 于 1982 年首先提出了一种新的数学理论——粗糙集(Rough Set,RS)。按照对粗糙集的理解，我们对事物的认知程度取决于所拥有知识的多少，知识越多，则事物间的区分越精细；知识越少，则区分得越模糊。

综上所述，三类不确定性，即随机性、模糊性和粗糙性可以这样来描述：

(1) 随机性是因为事物的因果关系不确定，从而导致事件发生的结果不确定性，用概率来度量。概率表示事件发生可能性的大小。概率的客观意义可以由随机试验中所呈现的频率稳定性来承担。概率论的运用是从随机性中去把握广义的因果律——概率规律。

(2) 模糊性是因为事件在质上没有明确的含义，在量上没有明确的界限，导致事件呈现“亦此亦彼”的性态，是事物类属的不确定性，用隶属度来度量。隶属度表示事物多大程度属于某个分类。隶属度的具体确定包含着人脑的加工，其中有着某种心理过程。模糊集合论的运用则是从模糊性中去确立广义的排中律——隶属规律。

随机性是因果律的一种破缺，模糊性是排中律的一种破缺。

(3) 粗糙性是因为描述事件的知识(或信息)不充分、不完全，导致事件间的

不可分辨性。粗糙集把那些不可分辨的事件都归属一个边界域。因此，粗糙集中的不确定性是基于一种边界的概念，当边界域为一空集时，则问题就变为确定性的。粗糙集认为知识的粒度性是造成使用已有知识不能精确地表示某些概念(事件、对象等)的原因。在没有掌握所有关于对象域的知识的情况下，我们只能用一对近似集来刻划不精确性或含糊性。

粗糙集理论基于知识的不可分辨性，模糊集理论则侧重知识的模糊性。不可分辨性和模糊性实际上是不完全知识的两个不同侧面。因此粗糙集和模糊集并不是互相对立的理论，而是互补的。

从集合论的观点来看，粗糙集既是经典集合的延拓，又是模糊集合的补充。

经典集合认为一个集合完全由其元素所决定，一个元素要么属于这个集合，要么不属于这个集合，只有“真”与“假”之分。经典集合的隶属函数$\mu_X(x)\in\{0,1\}$是二值逻辑，只能处理确定性问题。

模糊集合认为事物具有中介过渡性质，而非突然改变，集合中每一个元素的隶属函数$\mu_{\tilde{X}}\in[0,1]$，即在闭区间[0,1]可以任意取值，隶属函数是连续光滑的。因此，模糊集合对不确定信息的刻划是精细而充分的，但隶属函数不可计算，凭人的主观经验给定。

粗糙集合把用于分类的知识引入集合。一个元素x是否属于集合X，需要根据现有的知识来判断，可分为三个情况：1)x肯定属于X；2)x肯定不属于X；3)x可能属于也可能不属于X。到底属于哪种情况依赖于我们所掌握的关于论域的知识。上述三种情况分别对应于粗糙的正区域、负区域和边界域。因此，大体上可认为粗糙集是一种三值逻辑(正区域、负区域、边界域)。粗糙集的隶属函数为阶梯状，对不确定性信息的描述是粗糙的，但粗糙隶属函数是可以计算的。粗糙集主要用于对信息系统进行约简和分类。

模糊集研究的是属于同一类的不同对象的隶属关系，重在隶属的程度。粗糙集研究的是不同类中的对象组成的集合之间的关系，重在分类。两种方法在处理不完善数据方面可以互为补充。

三种集合的隶属函数如图 1-1 所示。

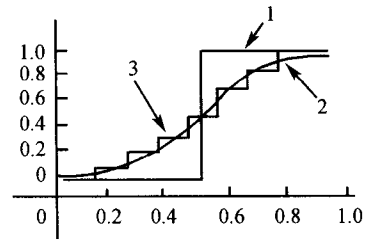

图 1-1　隶属函数

1—经典集合

2—模糊集合　3—粗糙集合

1.2　粗糙集的基本理论与方法

1.2.1　RS 的基本概念

1. 知识与分类

在粗糙集理论中，知识被认为是一种分类能力。人们的行为基本是分辨现实的或抽象的对象的能力。如在远古时代，人们为了生存必须能分辨出什么可以食用，什么不可食用；医生给病人诊断，必须辨别出患者得的是哪一种病。这些根据事物的特征差别，将其分门别类的能力均可以看作是某种知识。更抽象层次上的分类是推理、学习、决策的关键，是一种基础知识。

假定我们起初对论域内的对象(或称元素、样本、个体)已具有必要的信息或知识，通过这些知识能够将其划分到不同的类别。若我们对两个对象具有相同的信息，则它们是不可区分的，即根据已有的信息不能够将其划分开，显然这是一种等价关系，粗糙集理论的核心是等价关系，通常用等价关系代替分类，根据这个等价关系划分样本集合为等价类。从知识库的观点来看，每个等价类被称为一个概念，即一条知识(规则)。换句话说，每个等价类惟一地表示了一个概念，属于一个等价类的不同对象对该概念是不可区分的。

2. 知识表达系统

一般地，一个知识表达系统或信息系统 S 可以表示为有序四元组

$$S=\{U,R,V,f\}$$

式中，$U=\{x_1,x_2,\cdots,x_n\}$ 为论域，它是全体样本的集合；$R=C\cup D$ 为属性集合，其中子集 C 是条件属性集，反映对象的特征，D 为决策属性集，反映对象的类别；$V=\bigcup_{r\in R}V_r$ 为属性值的集合，V_r 表示属性 r 的取值范围；$f:U\times R\rightarrow V$ 为一个信息函数，用于确定 U 中每一个对象 x 的属性值，即任一 $x_i\in U$，$r\in R$，则 $f(x_i,r)=V_r$。如表 1-1 为某医疗信息表。

表 1-1　医疗信息表

属性 / 对象	条件属性 C			决策属性 D
	头疼 r_1	肌肉疼 r_2	体温 r_3	流感
x_1	是	是	正常	否
x_2	是	是	高	是
x_3	是	是	很高	是
x_4	否	是	正常	否
x_5	否	否	高	否
x_6	否	是	很高	是

3. 不可分辨关系

在粗糙集中，论域 U 中的对象可用多种信息(知识)来描述。如在表 1-1 中，U 中的对象是病人，病人可由他们发病的症状(条件属性)来描述。当两个不同的对象由相同的属性来描述时，这两个对象在该系统中被归于同一类，它们的关

系称之为不可分辨关系(indiscernibility relation)，即对于任一属性子集 $B \subseteq R$，如果对象 x_i，$x_j \in U$，$\forall r \in B$，当且仅当 $f(x_i, r) = f(x_j, r)$时，x_i 和 x_j 是不可分辨的，简记为 Ind(B)。不可分辨关系称为等价关系(equivalence relation)。

我们对不可分辨关系(或等价关系)还可做进一步解释。假设只用黑白两种颜色把空间中的一些物体划分成两类：{黑色物体}、{白色物体}，那么同为黑色的物体就是不可分辨的，因为描述它们特征属性的信息相同，都是黑色。如果再引入方、圆的属性，又可将物体进一步划分为四类：{黑色方物体}、{黑色圆物体}、{白色方物体}、{白色圆物体}。这时，如有两个同为黑色方物体，则它们还是不可分辨的，而两个白色圆物体间的不可分辨关系可以理解为它们在白、圆两种属性下存在等价关系。

不可分辨关系这一概念在 RS 中十分重要，它反映了我们对世界观察的不精确性。若两个对象分别处于 Ind(B)的不同划分中，那么它们就可以为现有的知识所分辨；若两个对象处于同一个划分中，它们是不能为现有的知识所分辨的。另一方面，不可分辨关系反映了论域知识的颗粒性，即通过现有的知识我们往往不能精确地认识每一个对象，属性相同的对象聚合在一起以颗粒的状态呈现在我们面前，彼此无法分辨开来。

一般地说，知识库中的知识越多，知识的粒度就越小，随着新知识不断加入到知识库中，粒度会不断减小，直至将每个对象区分开来。但知识库中的知识粒度并不是越小越好，粒度小必然导致信息量增大，储存知识库的费用增高。

4. 基本集合

由论域中相互不可分辨的对象组成的集合称之为基本集合，它是组成论域知识的颗粒。

在表 1-1 中，考虑条件属性：头疼和肌肉疼。对于 x_1，x_2，x_3 这三个对象，其条件属性头疼的值都是“是”，条件属性肌肉疼的值也都是“是”，因此，从条件属性头疼和肌肉疼的角度来看，这三个对象是不可分辨的。同样，x_4，x_6 在这两个属性上也是不可分辨的。由此构成的不可分明集$\{x_1, x_2, x_3\}$，$\{x_4, x_6\}$和$\{x_5\}$被称为基本集合。

设论域 U 是有限集，R 是 U 的等价关系簇，则 $K = (U, R)$称为知识库，知识库的知识粒度由不可分辨关系Ind(R)的等价类反映。若有 $B \subseteq R$，且 $B \neq \phi$ 则$\bigcap B$(B 中全部等价关系的交集)也是一种等价关系，称其为 B 上的不可分辨关系，记为 Ind(B)。若$\{x_i, x_j\} \in$ Ind(B)，则称对象 x_i 与 x_j 是 B 不可分辨的，即 x_i，x_j 存在于不可分辨关系 Ind(B)的同一个等价类中。依据等价关系簇 B 形成的分类知识，x_i 与 x_j 无法区分。U/Ind(B)中的各等价类称为 B 基本集。基本集是粗糙集中构成知识的基本模块。若集合 X 可以表示成某些基本集的并时，则称 X 是 B 可定义集，否则称为不可定义集，也可称 X 为 B 粗糙集。

5. 下近似集和上近似集

下近似集(Lower approximation)定义为：根据现有知识 R，判断 U 中所有肯定属于 X 的对象所组成的集合，即

$$R_-(X) = \{x \in U, [x]_R \subseteq X\} \tag{1-1}$$

式中，$[x]_R$ 表示等价关系 R 下包含元素 x 的等价类。

上近似集(Upper approximation)定义为：根据现有知识 R，判断 U 中一定属于和可能属于 X 的对象所组成的集合，即

$$R^-(X) = \{x \in U, [x]_R \cap X \neq \phi\} \tag{1-2}$$

综上所述，我们可以给出便于计算下近似集和上近似集的定义：给定知识表达系统 $S = \{U, R, V, f\}$，对于每个样本子集 $X \subseteq U$ 和等价关系 R，所有包含于 X 的基本集的并(逻辑和)为 $R_-(X)$；所有与 X 的交(逻辑积)不为空集(ϕ)的基本集的并为 $R^-(X)$。

6. 正域、负域和边界域

正域(Positive region)：$\mathrm{Pos}(X) = R_-(X)$，即根据知识 R，U 中能完全确定地归入集合 X 的元素的集合。

负域(Negative region)：$\mathrm{Neg}(X) = U - R_-(X)$，即根据知识 R，U 中不能确定一定属于集合 X 的元素的集，它们是属于 X 的补集。

边界域(Boundary region)：$\mathrm{Bnd}(X) = R^-(X) - R_-(X)$，边界域是某种意义上论域的不确定域，根据知识 R，U 中既不是肯定归入集合 X，又不能肯定归入集合 $\bar{X}$(即 $U - X$)的元素构成的集合。

边界域为集合 X 的上近似与下近似之差，如果 $\mathrm{Bnd}(X)$是空集，则称集合 X 关于 R 是清晰的；反之，如果 $\mathrm{Bnd}(X)$不是空集，则称集合 X 为关于 R 的粗糙集。因此，粗糙集中的“粗糙”(不确定性)主要体现在边界域的存在。集合 X 的边界域越大，其确定性程度就越小。

7. 粗糙度(近似精确度)

对于知识 R(即属性子集)，样本子集 X 的不确定性程度可以用粗糙度 $\alpha_R(X)$来表示为

$$\alpha_R(X) = \frac{\mathrm{Card}(R_-(X))}{\mathrm{Card}(R^-(X))} \tag{1-3}$$

$\alpha_R(X)$亦称近似精确度，式中 Card 表示集合的基数(集合中的元素个数)。

显然，$0 \leqslant \alpha_R(X) \leqslant 1$，如果 $\alpha_R(X) = 1$，则称集合 X 相对于 R 是确定的，如果 $\alpha_R(X) < 1$ 则称集合 X 相对于 R 是粗糙的，$\alpha_R(X)$可认为是在等价关系 R 下逼近集合 X 的精度。

8. RS 的计算举例