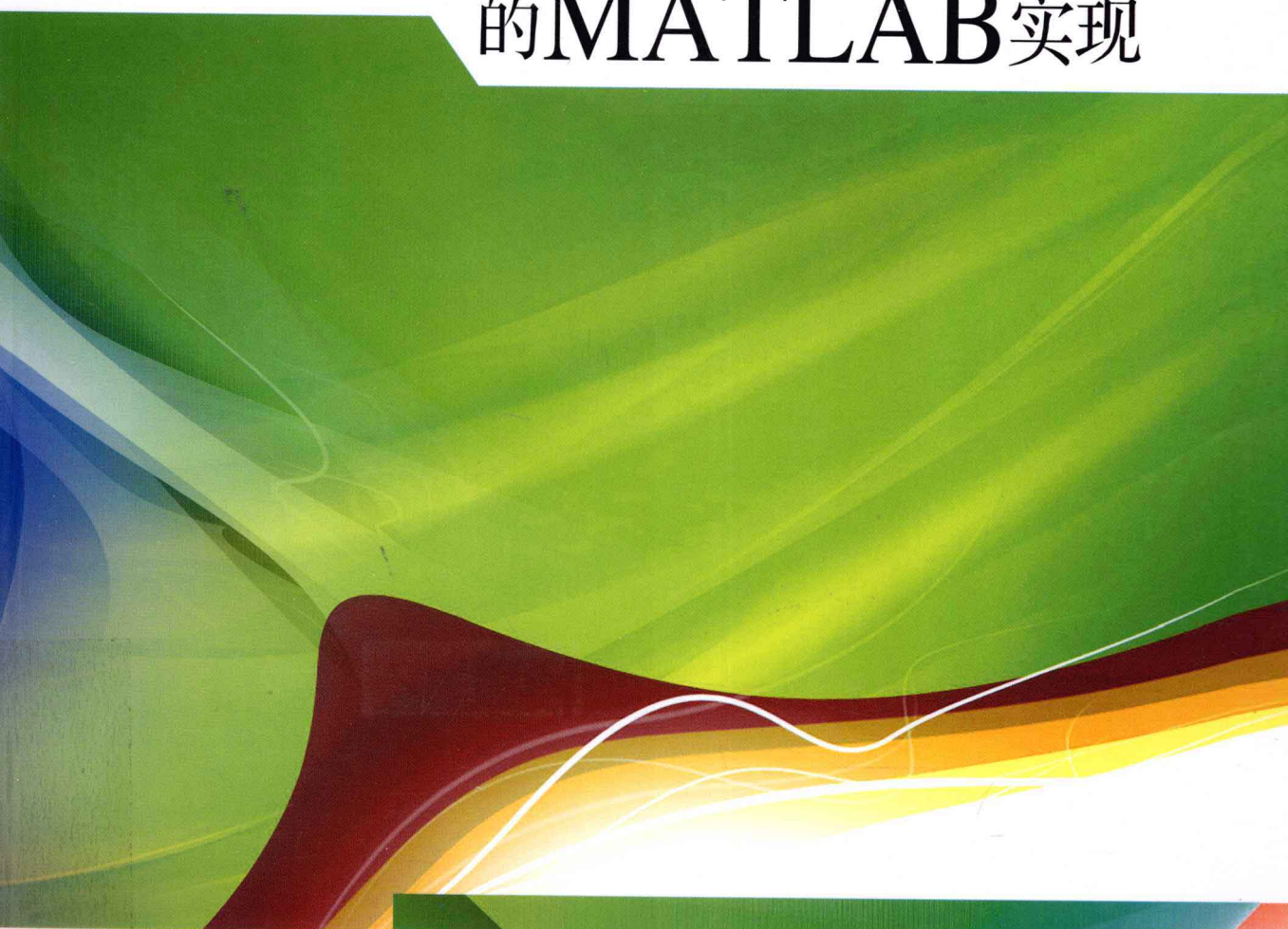
MATLAB
开发实例系列图书

在线交流 有问必答

模式识别与智能计算的MATLAB实现



许国根 贾 瑛 编著

北京航空航天大学出版社
BEIHANG UNIVERSITY PRESS

MATLAB开发实例系列图书

模式识别与智能计算的MATLAB实现

许国根　贾　瑛　编著

北京航空航天大学出版社

内 容 简 介

针对各学科数据信息的特点以及科学工作者对信息处理和数据挖掘技术的要求，本书既介绍了模式识别和智能计算的基础知识，又较为详细地介绍了现代模式识别和智能计算在科学研究中的应用方法和各算法的 MATLAB 源程序。本书可以帮助广大的科学工作者掌握模式识别和智能计算方法，并应用于实际的研究中，提高对海量数据信息的处理及挖掘能力，针对性和实用性强，具有较高的理论和实用价值。

本书可作为高等院校计算机工程、信息工程、生物医学工程、智能机器人、工业自动化、地质、水利、化学和环境等专业的研究生、本科生的教材或教学参考书，亦可供有关工程技术人员参考。

图书在版编目(CIP)数据

模式识别与智能计算的 MATLAB 实现/ 许国根，贾瑛编著. -- 北京：北京航空航天大学出版社，2012.7

ISBN 978-7-5124-0843-2

Ⅰ. ①模… Ⅱ. ①许…②贾… Ⅲ. ①模式识别-Matlab 软件②智能计算机-Matlab 软件 Ⅳ. ①O235②TP387③TP317

中国版本图书馆 CIP 数据核字(2012)第 125826 号

模式识别与智能计算的 MATLAB 实现

许国根 贾 瑛 编著

责任编辑 王慕冰 龚荣桂 王平豪

*

北京航空航天大学出版社出版发行

北京市海淀区学院路 37 号(邮编 100191) http://www.buaapress.com.cn

发行部电话：(010)82317024 传真：(010)82328026

读者信箱：bhpress@263.net 邮购电话：(010)82316936

涿州市新华印刷有限公司印装 各地书店经销

*

开本：787×1 092 1/16 印张：17.75 字数：454 千字

2012 年 7 月第 1 版 2012 年 7 月第 1 次印刷 印数：4 000 册

ISBN 978-7-5124-0843-2 定价：36.00 元

前　言

MATLAB是功能非常强大的计算机软件，在科学研究和工程实践中得到了广泛的应用。利用它来编制现代模式识别和智能计算等技术的程序，揭开这些在大多数人眼中极为深奥的数学方法的神秘面纱，使每个科学工作者都能非常容易地使用它们来解决实际问题，是作者学习MATLAB后，结合实际的科学研究经验产生的一个强烈的愿望。

本书理论联系实际，较为全面地介绍了现代模式识别和智能计算方法及其应用技巧。通过大量实例，讲解了模式识别和智能计算的理论、算法及编程步骤，并提供基于MATLAB的源代码。通过本书的学习，读者能够真正掌握模式识别和智能计算并应用于科学研究和工程实践中。

模式识别是当今高科技研究的重要领域之一，它创立于20世纪60年代，初期属于信息、控制和系统科学领域。模式识别是利用某些特征，对一组对象进行判别或分类，被分类的对象即为模式，分类的过程即为识别。模式识别所涉及的信息往往存在高维、影响因素多、关系复杂等特征，单靠人的思维往往难以有效地确定其规律，需要通过一定的数学方法借助计算机来完成。到了20世纪70年代后，随着大规模集成电路技术的发展，特别是计算机硬件的飞速发展，无论在理论上，还是在应用上，模式识别技术都有了长足的进步，同时，也推动了以计算机科学为基础的具有智能性质的自动化系统的实际应用，促进了人工智能、专家系统、景物动态分析、图像识别、语音识别等多学科的发展，广泛应用于人工智能、机器人、系统控制、遥感数据分析、生物医学工程、军事目标识别等领域，在国民经济、国防建设、社会发展等各个方面发挥着越来越重要的作用。

在众多学科的科学研究及工程应用中，人们往往通过对研究对象的观察和实验积累了海量的数据信息，并且由于对象的复杂性，使得这些数据具有高维、复杂非线性、强相关性和多噪声等特点。如何从这些数据信息中发现更多、更有价值的关系，找到其内在规律，建立的模型能良好地反映研究对象的实际特征和良好的可理解性，易与先验知识相结合，并能适应大规模数据处理的要求，正逐步成为科学工作者关注的焦点。常规数学手段已不能解决这个问题，现代模式识别和智能计算将起到十分重要的核心作用。

现代模式识别和智能计算从已知数据出发，参照相应的数学(或物理、化学)模型或经验规律得到一批特征量，然后进一步进行特征抽取以求得合适的特征量，张成模式空间或特征空间，最后通过模式识别算法进行训练和判别，以揭示已知数据信息中隐含的性质和规律，为研究者提供十分有用的决策信息和过程优化的重要信息。

现代模式识别和智能计算作为一种高效的信息处理技术，解决了众多学科研究和工程应用中许多重要的问题。例如化学研究中对化合物性质的分类、化合物的构效分析、基于化学方法的疾病诊断、药物的分类、环境质量评价等，都可以通过模式识别和智能计算对一组样本中某些化学组分的分析结果进行分析处理来实现；在机械故障诊断、军事目标识别、遥感数据分析、水利等学科的研究中，模式识别和智能计算也得到了广泛的应用。目前，模式识别和智能计算已经成为科学研究中的一种重要的数据处理手段，应用越来越广泛。

国内外论述模式识别和智能计算的参考书为数不少。由于这一领域涉及深奥的数学理

论，而且大部分书籍只是罗列模式识别和智能计算的原理及伪代码，看不到源代码以及算法的实际效果和各种算法的对比结果，往往使大多数科学研究者感到困惑，不知如何下手。对大多数科学工作者而言，目前确实还缺少一本具有较强的系统性、可比性及实用性的模式识别和智能计算的参考书。基于这点考虑，作者撰写了本书，目的是想通过系统的介绍和实例分析，让众多学科的科学工作者不仅具备模式识别和智能计算的理论，而且能掌握模式识别和智能计算的方法，可以在各自的学科实际研究中加以应用。

本书内容基本涵盖了目前模式识别和智能计算的重要理论和方法，包括了最近十几年来刚刚发展起来的并被实践证明有用的新技术、新理论，如支持向量机、神经网络、决策树、粗糙集理论、模糊集理论和遗传算法等。在介绍各种理论和方法的同时，将不同算法应用于科学研究的实际中。对于每一种模式识别方法，本书都通过实际问题，按照理论基础、实现步骤、编程代码三部分进行阐述，避免空洞的理论说教，着重介绍编程步骤及MATLAB源代码，具有较强的指导性和实用性，使读者不至于面对如此丰富的理论和方法无所适从，而是通过了解各种算法的实现思路和方法，体会算法的源代码的意义，这样即使所举的实例不属于读者从事的学科，通过举一反三，读者也能掌握模式识别和智能计算并应用于自己从事的科学研究中。

由于至今还没有统一的、有效的能够应用于所有的模式识别的理论，当前的一种普遍看法认为，不存在对所有的模式识别问题都适应的单一模型和解决识别问题的单一技术，即存在着所谓的无免费午餐定理(NFL定理)：若算法A在某些函数中的表现超过算法B，则在这类其他的适应度值函数中，算法B的表现就比A要好。我们所要做的是把模式识别方法与具体问题结合起来，把模式识别和智能计算与统计学、神经网络、数据挖掘、机器学习、人工智能等学科的先进思想和理论融合在一起，为读者提供一个多种理论和方法的测试平台，并在此基础上，深入掌握各种理论的效能和应用的可能性，互相取长补短，开创模式识别和智能计算在各学科中应用的新局面。

本书的出版得到了北京航空航天大学出版社的大力支持，编辑陈守平为本书内容编排等许多方面提出了宝贵的意见，在此表示衷心的感谢！同时对参考文献中列出的引用资料的作者表示衷心的感谢！

由于作者水平有限，特别是数学知识相对贫乏，书中难免存在缺点和错误，敬请读者批评斧正。

作　者

2011年于西安二炮工程大学

相信吗?这是一本会动的图书！无需怀疑，当您拿起此书，恭喜您，您已经找到了一条学习 MATLAB&Simulink 的捷径——图书+论坛。

目前国内最大的 MATLAB&Simulink 技术交流平台——MATLAB中文论坛（www.iLoveMatlab.cn）联合本书作者、编辑将为您提供所需要的问题答案和大量技术支持，让本书成为一个联系同行、链接相关知识点的活动载体，确保您增值无限！

请登录 www.iLoveMatlab.cn 提出您在图书阅读和代码调用过程中产生的疑问，本书作者将定期为您释疑，同时您还有机会和作者面对面交流；如果您对此书内容或代码有任何建议，也可以发帖反映，您的建议将是我们创造精品的动力和源泉。请您随我们一起“动”起来，让这条“读者—作者”交流渠道更畅通，让该书动得更炫！

“在线交流、有问必答”网络互动参与步骤：

① 在MATLAB中文论坛 www.iLoveMatlab.cn 注册一个会员账号并登录。

② 由本书配套的“在线交流卡”获得卡号和密码。

③ 在以下网址验证密码：http://www.iLoveMatlab.cn/book.php。

④ 验证完毕，进入本书版块，与作者在线交流。

目　录

第 1 章 绪 论

模式识别(pattern recognition)是当前科学发展中的一门前沿学科,也是一门典型的交叉学科,它的发展与人工智能、计算机科学、传感技术、信息论等学科的研究水平息息相关,相辅相成。

模式识别在数据挖掘、生物特征识别、自动检测、化学、遥感、文本分类、工业自动化、文档与图像分析等领域的应用越来越广泛,同时,模式识别技术的研究也越来越受到各方面的重视,与之相关的市场也越来越庞大,如智能交通管理系统、文字识别系统等都需要模式识别技术的支撑。

1.1 模式识别的基本概念

1.1.1 模式与模式识别的概念

模式识别是人类天生具备的能力。初生婴儿就能辨认自己的父母;大多数 5 岁的孩子便能辨认数字和字母,而且不论其大小、方向、手写体、印刷体,甚至这些数字或字母的一部分被遮住……。随着计算机技术和人工智能技术的迅速发展,人们迫切希望计算机能够完成模式的识别工作,诸如听懂人类所说的话,看懂人类所写的文字,而不论它是手写体还是印刷体……。这些愿望促进了模式识别学科的形成和发展。

模式识别研究的目的是利用计算机对物理对象即"模式"进行分类或描述,在错误率最小的条件下,使识别的结果与客观物体相符合。"模式"是一个内涵十分丰富的概念,可以把凡是人类能用其感官直接或间接接受的外界信息都称为模式,而把具有某些共同特性的模式的集合称为模式类。

1.1.2 模式的特征

在模式识别中,被观察的每个对象称为样本(或样品)。对于一个样本来说,必须确定一些与识别有关的因素,作为研究的根据,每一个因素称为一个特征。例如,医生给病人治病,病人就是样本;疾病特征或病人的一些生理指标即为模式识别的特征。

模式的特征集可用处于同一个特征空间(R^n)的特征向量表示。如果一个样本 $\boldsymbol{X}$ 有 n 个特征,则 $\boldsymbol{X}$ 可以看作是一个 n 维的列向量,其中的每个元素称为特征,该向量也因此称为特征向量:

$$\boldsymbol{X}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$$

若有 N 个样本,每个样本具有 n 个特征,则样本集的特征可以用一个 n 行 N 列的矩阵表

示。待识别的不同模式都在同一特征空间中考察，但由于性质上的差异，即各特征取值的不同，它们会在特征空间的不同区域(类)中出现。模式识别就是根据 $\boldsymbol{X}$ 的 n 个特征来判别模式 $\boldsymbol{X}$ 属于 $\omega_1,\omega_2,\cdots,\omega_M$ 中的哪一类。

1.1.3 模式识别系统

模式识别的整个过程大致要经历数据采集、数据预处理、特征提取与特征选择及模式识别4个阶段，实际上就是将预处理后的原始数据所在的数据空间经特征空间到类别空间的映射过程，如图1.1所示。

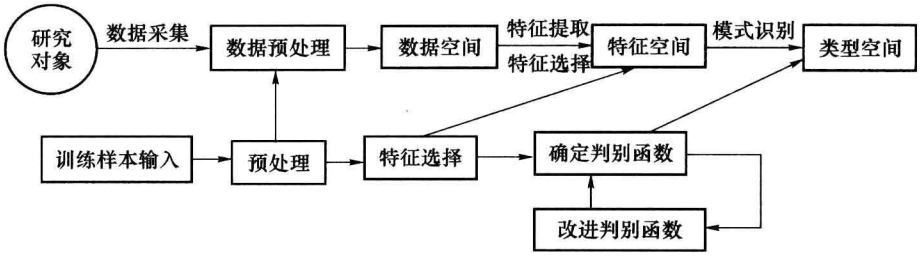


图1.1 模式识别过程

(1) 数据采集

一般获取的数据类型有以下几种：

- 二维图像，如文字、指纹、地图、照片等；
- 一维波形，如脑电图、心电图、时间序列等；
- 物理、化学等参数和逻辑值，如体温、水质化验数据、参量正常与否的描述等。

(2) 预处理

对测量数据进行去噪、复原、归一化和标准化等处理。

(3) 特征提取和选择

对原始数据进行分析，去掉对分类无效或易造成混淆的那些特征，得到最能反映模式本质的特征；有时还采用变换技术，将高维的特征空间转变为低维的特征空间。

(4) 模式识别

首先按设想的分类判决数学模型对训练样本模式进行训练，得到分类的判决规则；然后利用判决规则对待识别模式的特征进行识别，输出识别结果；最后将已识别的分类结果与已知类别的输入模式作对比，不断改进判决规则和特征选择与提取方法，制定出使错误率(风险率)最小的判决规则和特征选择与提取策略，即再学习过程。

1.2 模式识别的主要方法

目前，常用的模式识别方法包括三大类，即模板匹配方法、结构模式识别和统计模式识别。

模板匹配模式识别是应用最早且最简单的模式识别形式，它通过比较待识别模式和已有模式的相似程度来实现模式识别的目的。

在一些模式识别问题中，研究的模式或者非常复杂，或者类别很多，不能用简单的分类就

能解决。例如在对一幅画进行识别时，不仅要识别画中的简单物品（如桌子、杯子等），而且要对画有更完整的描述（如这些物体间的相互关系），并且产生出一个模式的结构的描述，这种描述一般采用形式语言的形式，这就是结构模式识别。

统计模式识别是研究得最多也最为深入的一种模式识别方法。在此模式中，每一个模式采用 n 维特征或测量值来表示，最终的目的是在由这些特征构成的空间中能将各模式有效地分开。

1.3 模式识别的主要研究内容

模式识别的主要研究内容包括三部分：模式分类、模式聚类、特征提取和选择。

模式分类是模式识别的主要内容，即将某个模式分到某个模式类中。在这个过程中首先需要建立样本库，然后根据样本库建立判别函数，这一过程由机器来实现，称为学习过程。然后对一个未知的新对象分析它的特征，并根据判别函数决定它属于哪一类。模式分类是一种监督学习的方法。可用于模式分类的方法有很多，经典的方法有贝叶斯分类、Fisher 判据和近邻法等，现代的方法有模糊模式识别、人工神经网络模式识别、支持向量机和基于核的分类方法等。

聚类分析是统计模式识别的另一重要工具。模式聚类时遵循“同一个聚合类的模式比不同聚合类中的模式更相近”的原则。它的基本原理就是在没有先验知识的情况下，基于“物以类聚”的观点，用数学方法分析各模式向量之间的距离及分散情况，按照样本的距离远近划分类别。聚类分析是一种无监督学习的方法。

如何确定合适的特征空间是设计模式识别系统一个十分重要的问题。如果所选用的特征空间能使同类物体分布具有紧密性，即各类样本能分布在该特征空间中彼此分割开的区域内，这就为模式识别成功地提供了良好的基础。当识别对象是波形或数学图像时，模式的特征是通过计算而得到的；当识别对象是实物或某种过程时，模式的特征则是由仪器设备测量而来的。这样产生的特征称为原始特征。一般需要对原始特征进行预处理，有时还需要转换。预处理是为了除去原始特征的噪声等影响模式识别的因素；转换是为了将高维的特征空间降为低维的特征空间而有利于后续的分类计算。

1.4 模式识别在科学研究中的应用

数十年来，模式识别研究取得了大量的成果，新理论及算法不断出现，并在许多领域得到了成功的应用，很难用一节篇幅作无遗漏的叙述，下面只作一些简单的介绍。

1.4.1 化合物的构效分析

据 1978 年估计，全世界用于找新药的费用每年达 20 亿美元左右，每发明一种重要的新药耗资为 4 000 万美元。为了更快更省地开发新药，迫切需要总结化合物分子结构和性能的关系，以提高探索的命中率。这种构效关系（Structure-Activity Relationship，SAR）研究可有演绎法、归纳法两种途径。演绎法是从量子生物学角度查明药物活性的机理，从而确定何种结构最有效。但目前的知识水平距这一目标尚十分遥远。归纳法则是利用模式识别等方法从大量实验结果中总结规律。这一方法虽然是纯经验性质或半经验性质，但切实可行。由于新药研究合成和药理试验工作量大，费用也相当高，即使是误报率相当大的模式识别方法，也能产生

一定的效益。

模式识别方法也是研究化合物结构与性能关系的有效工具。例如许多化合物具有致癌性或者抗癌活性，研究这些化合物的结构特点，对于人类预防及治疗癌症具有重要的意义。例如在对200个化合物(其中87个有抗癌活性)的抗癌活性与结构间的关系研究时，利用20个结构参数，用线性判别函数法和K近邻法判别各化合物的抗癌活性，分类率可达90左右，并发现下列结构特征(特征参数)与化合物抗癌活性关系较大：硫原子/总原子数；C—S键数/碳原子数；S—H键数；C=C键数/碳原子数；碳原子数/总原子数。

多环芳烃是含有多个芳烃环的化合物，多种多环芳烃是强致癌物质。人们通过大量实验和句法模式识别技术，发现多环芳烃的分子图形和致癌活性有很大关系。鉴于图像和图形信息在化学中的用途，特别是有机化学结构中图形信息尤其丰富，句法模式识别有可能在有机分子设计、药物设计等方面得到广泛的应用。

1.4.2 谱图解析

随着各种物理方法和物理化学方法在化学分析中的推广应用，获取质谱、光谱、色谱和电子能谱等谱图已成了专门的学问。各种谱图包含大量的化学信息，不但可以用来鉴定未知物的成分，测定某些成分的含量，而且可以用来探讨或确定分子或固体的结构、化学键的特征等。理想的做法应当是彻底弄清各种谱图产生的机理，从而从理论上完成从实测谱图到化学成分、分子结构、化学键特征等化学信息的交换。但实际上很难完全做到这一点。以最简单的光谱——原子光谱为例，重原子的原子光谱中迄今为止多数谱线不能从理论上解释。这就不能不用经验方法对谱图做鉴别和解析工作，以达到化学分析和结构分析的目的。由于化合物种类庞杂，谱图的数据亦急剧增加，单凭少数有经验的专家来做谱图解析已不能满足需要。随着计算机人工智能、模式识别和数据库技术的发展，用计算机做谱图解析的各种方法应运而生。其中有一类方法是数据库谱图显示方法，即将大量已知化合物的谱图保存在数据库中，通过检索的方法来识别谱图。另一类方法是模式识别方法，它利用已知谱图作训练点，对未知物的谱图作分类、鉴别以及结构测定等。由于化合物种类庞杂、数目很多且每年都大量增加，单纯依靠已知谱图的存储和检索不能完全解决谱图解析问题。由于模式识别方法有某种“举一反三”的功能，能从大量已知化合物的谱图做分类工作，所以模式识别方法在谱图解析、分析化学、结构确定等方面有重要的实际意义。迄今为止，质谱、原子光谱、红外光谱、拉曼光谱、核磁共振谱、γ-射线谱、色谱、极谱等的谱图识别都已用了模式识别方法，不同程度地收到效果，这方面的研究工作是现代分析化学的前沿课题。

1.4.3 材料研究

金属等各种材料具有不同的性质，人们往往根据其性能确定它的用途。但是寻找一种新材料的工作是十分艰苦的。一般要通过大量的“配方炒菜”式的实验工作，才能筛选出较好的材料。以高温合金为例，试制一种新的高温合金要初筛千百种配方，初选后还要做成千小时的高温长期性能测试。这一类先搞大批“配方炒菜”，再逐一测试性能的工作方法需要消耗大量的人力、物力和时间。如何利用计算机信息处理方法使寻找新材料的工作方式有所改进，以收到事半功倍的效果，是近数十年来许多科学家努力研究的课题。

瑞典钢铁公司试制了15种新钢种，在新钢种的钢材加工过程中，有9种钢材开裂，另6种不开裂。为了查明钢中微量元素对钢材开裂的影响，他们分析了这15种钢材中的17种微量