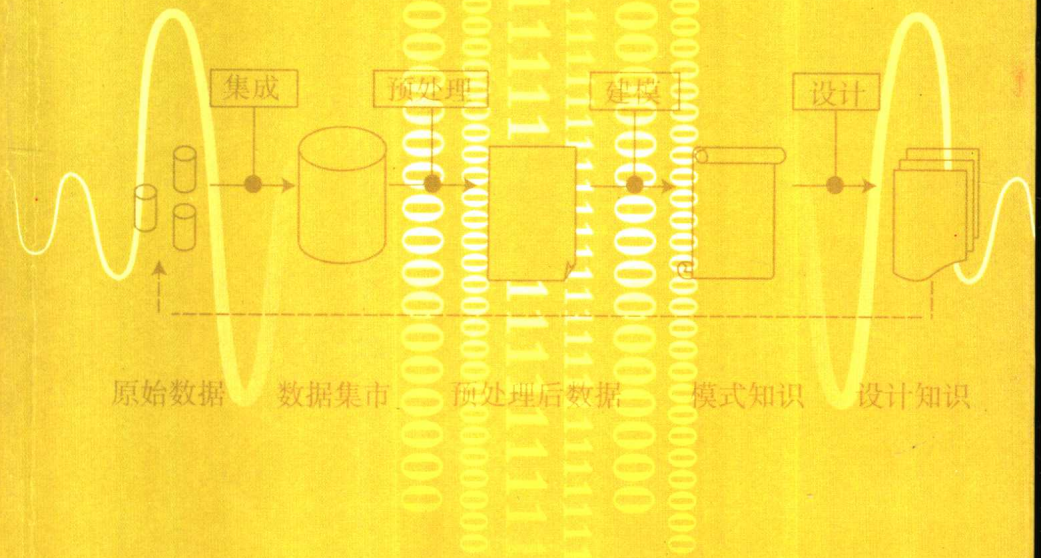
数据挖掘在治金产品质量控制中的应用

邢进生 著



国防工业出版社
http://www.ndip.cn

数据挖掘在冶金产品质量控制中的应用

邢进生 著

国防工业出版社
·北京·

图书在版编目(CIP)数据

数据挖掘在冶金产品质量控制中的应用/邢进生著．
北京:国防工业出版社,2004.6
ISBN 7-118-03507-6

Ⅰ.数… Ⅱ.邢… Ⅲ.数据采集－计算机应用－冶金工业－工业产品－质量控制 Ⅳ.F407.41-39

中国版本图书馆 CIP 数据核字(2004)第 046416 号

国防工业出版社出版发行
(北京市海淀区紫竹院南路 23 号)
(邮政编码 100044)
腾飞胶印厂印刷
新华书店经售
*
开本 850×1168 1/32 印张 7½ 205 千字
2004 年 6 月第 1 版 2004 年 6 月北京第 1 次印刷
印数:1－1000 册 定价:22.00 元

前　言

一个大型冶金企业，在其长期的生产过程中必然已积累了丰富详实的生产实绩数据，能否充分利用这些数据来提高产品质量，是工程技术人员、产品质量管理人员一直想解决的问题。数据挖掘技术是20世纪90年代迅速发展起来的计算机技术，其核心思想就是从企业堆积如山的数据中找出有效的、新颖的、具有潜在效用的、最终可理解的规律或模式，来指导企业制定管理产品质量的技术决策，为企业创造巨大的经济效益。因此，利用数据挖掘技术可帮助工程技术人员、产品质量管理人员解决目前面临的问题。

本书是作者在博士和博士后期间科研工作的总结，是以作者的博士论文和15余篇相关学术论文为基础，结合具体实践研究资料编著完成的。主要内容是以钢铁企业的大型多辊热连轧机生产线板材产品质量控制为目的，以提高产品质量、降低产品成本、适用辅助新产品新工艺设计为宗旨，提出了基于神经网络产品质量模型、逆质量模型和质量控制模型的新产品新工艺设计的数据挖掘(DM，Data Mining)方案，并给出其实现的方法和应用软件。主要研究工作有以下几方面。

(1) 针对热轧产品质量的控制问题，将通常的DM方案具体化；利用数据仓库等概念，设计了来自不同计算机系统关于热轧产品数据的数据集市；提出了数据集市中一些数据预处理的方法；通过分析数据和与厂方技术人员讨论得出影响热轧产品质量的关键输入变量（KIV，Key Input Variable）有32个，代表产品质量指标的关键输出变量（KOV，Key Output Variable）

有 3 个。

(2) 分析了神经网络，特别是 BP、RBF 神经网络的建模算法和结构，提出了适合现场生产数据的高维 BP 神经网络建模方法；研究了具体生产数据的最佳 BP 神经网络结构参数。

① 训练样本逐渐扩大的 BP 神经网络建模方法：把每次校验不合格的数据作为新的训练样本加入原训练样本中形成新的样本继续训练 BP 神经网络，直到 BP 神经网络质量模型的校验精度符合工程要求。用所建质量模型校验三个质量指标，校验结果命中率分别为 84.7%、98.5%、84.6%，符合工程实际要求。

② 两阶段混合算法的 BP 神经网络建模方法：把 BP 神经网络算法和随机搜索算法相结合，第一阶段的权值作为第二阶段的初始值。将这种神经网络算法用于大型多辊热连轧产品质量数据的建模中，经过 9000 多个实测数据建模及校验，85%样本的校验值与实测值的误差满足工程实际要求。应用实例及校验结果表明，这种算法适宜于解决高维输入数据的建模问题，对提高建模速度与精度具有良好效果。

③ 基于产品加工工序的建模算法：根据产品加工工序构造多输入层神经网络。

④ 用具体数据研究了 BP 神经网络质量模型的最佳结构，给出了一个数据集（其中有 915 条观测和 34 个变量）在什么结构参数下，所建模型校验的 RMSE（均方差开方）最小，即模型最好。

(3) 分析了模糊神经网络的建模问题，建立了适合热轧产品质量控制的基于矩形函数系的高维模糊神经网络质量模型。新模糊神经网络的建模不必求导，因此运算简便，收敛速度快，适宜于高维模型的控制，可用于热轧产品质量控制。经 9000 多个实测数据建模及校验表明，85%样本的校验值与实测值的误差满足工程实际要求，较之传统的基于公式模糊神经网络模型有较大提高。同时还研究了模糊神经网络的记忆问题。

(4) 提出了两种基于产品质量模型的新产品新工艺设计方法。

① 基于神经网络产品质量模型的新产品新工艺设计方法：它是用建模数据相关分析的结果作指导，对神经网络产品质量模型进行有目的的校验，然后利用模型值在某一确定区间内找出关于用户希望的新产品生产方案。对新产品质量的多个目标值，可以取每个目标值范围的交集。实例研究表明，只要所建的神经网络产品质量模型具有足够高的精度，该方法是可行的。

② 基于BP神经网络产品质量模型输出的两阶段优化的新产品新工艺设计方法：它是用梯度法优化目标值，然后再用随机搜索法进一步优化目标值。本方法证明了这种两阶段优化算法具有收敛性。对于由一组多辊热连轧机生产线热轧产品质量数据建立的神经网络产品质量模型，利用这种算法经 10000 步迭代，求出了能使神经网络输出优化的输入向量，其相应的神经网络输出值优于热轧产品质量数据中的质量指标最优值，确实起到了神经网络输出优化的作用。

(5) 提出了两种基于产品质量控制模型的新产品新工艺设计方法：

① 基于 BP 神经网络产品质量控制模型的新产品新工艺设计方法：用控制量作为模型的输出，产品质量指标作为模型的输入建立 BP 神经网络质量控制模型，然后给一个质量指标值，通过模型就可得到一个控制量的值，把这些控制量的值和其他有关变量的值合到一起就是要找的新产品新工艺的生产方案。

② 基于矩形模糊神经网络产品质量控制模型的新产品新工艺设计方法，是利用这种模糊神经网络模型，由实际的输出值可以方便地求出输入向量的某一区域，使得对此区域中的每一向量，实际的输出与模糊神经网络的输出值之差可以达到任意要求的精

度，所求出的输入向量的某一区域就是要找的新产品新工艺的生产方案。

本书的完成从构思、确定写作内容、收集资料到书的定稿，都是在导师汪应洛院士指导下完成的。导师严谨的治学态度、渊博的学识、勇于创新的求学精神、脚踏实地的工作作风及宽宏谦逊的人品使我受益终生。同时，还要特别感谢师母张教授在我求学期间对我生活上无微不至的关怀。师恩难忘，我惟有在今后的工作中倍加努力，不辜负老师对我的一片厚望，以此来表达我对导师最真诚的谢意。

另外，特别感谢我的博士导师万百五教授，正是因为我在攻读博士学位期间，得到他的精心指导和培养，才使我的科研能力有较大的提高，他对我的指导和影响在今后的科研工作中必将起到很大的作用。在本书内容研究中，我还有幸得到冯祖仁教授和冯建生教授的指导。他们敏锐的科学洞察力和孜孜不倦的教书育人的精神给我留下了深刻的印象，感谢他们对我研究工作的帮助。

在此，我还要感谢我的妻子李晋玲女士，是她的大力支持才使我能够完成科研工作和本书的写作。

在求学期间和与宝钢自动化所的项目合作中，得到了许多老前辈和同学的鼎力帮助，在此特别感谢宝钢的王洪水高工、华文成高工、吴少敏博士、王迎军博士、苏东平硕士、陈文明硕士、陈贻龙先生，以及西安交通大学的刘人境博士、安凯博士、郑亚林博士、杨森博士、钱富才博士、贾磊博士、阮小娥博士、李换琴博士。

还要感谢国防工业出版社和西安交通大学、山西师范大学为本书出版给予的帮助。

本书内容的研究和出版，得到了国家 863 计划（编号：863-51-945-011）、国家自然科学基金项目（编号：79800003）、国家博士后基金项目（编号：2001[5]）、上海宝山钢铁公司重大科

研项目(编号：9812010)和陕西省自然科学基金（编号：98-SL08）的部分资助，在此一并表示诚挚的谢意。

由于作者水平有限，时间仓促，再加上部分内容还是项目的阶段性成果，不妥、错误之处在所难免，希望专家和同行批评指正。

邢进生

目　录

第一章　引　论 1
1.1　数据挖掘技术及其研究现状 2
1.1.1　知识获取与数据挖掘技术 2
1.1.2　数据挖掘研究与应用的现状 3
1.2　数据挖掘的对象 6
1.2.1　数据库 7
1.2.2　数据仓库 8
1.2.3　文本 11
1.2.4　Web 信息 12
1.2.5　空间数据 12
1.3　数据挖掘的主要技术 13
1.4　数据挖掘过程及结果解释 17
1.5　数据挖掘建模设计方案 19
1.5.1　通用的数据挖掘框架 19
1.5.2　建模设计方案的基本框架 20
1.5.3　方案实施的系统环境 22
1.6　冶金产品质量控制问题分析 23
1.7　本书的主要工作 23
第二章　冶金产品质量数据集市的构建 25
2.1　数据仓库与数据集市 26
2.1.1　数据仓库概述 27
2.1.2　数据集市 38
2.2　热轧产品质量数据集市的建立 39

2.2.1 热轧数据的现状 39
2.2.2 热轧数据集市的实现 41
2.3 数据预处理 43
2.4 确定建模数据的输入输出变量 45
2.5 建模数据的筛选与归一化 46
2.5.1 建模数据的筛选 46
2.5.2 建模数据的归一化 46
2.6 小结 47
第三章 人工神经网络特征分析 49
3.1 人工神经网络概述 50
3.1.1 神经网络的结构及设计方法 51
3.1.2 神经网络的学习方法 55
3.1.3 基本人工神经元模型 57
3.2 感知器模型及算法研究 58
3.3 多层前向神经网络的误差反向传播(BP)算法 78
3.3.1 BP 神经网络学习方法分析 79
3.3.2 BP 神经网络学习方法的几种改进 84
3.3.3 影响 BP 神经网络建模的其他因素 94
3.4 RBF 神经网络算法 95
3.4.1 RBF 神经网络结构 95
3.4.2 RBF 网络的算法分析 100
第四章 基于神经网络的产品质量模型 103
4.1 逐渐扩大训练样本的 BP 神经网络质量模型 104
4.1.1 基于数据集 F 的 BP 神经网络模型 104
4.1.2 对三类钢的模型测试 109
4.1.3 输出变量为 ys_re1，ys_rm1 的质量模型 111
4.2 二阶段混合算法的 BP 神经网络模型 114
4.2.1 二阶段混合算法 115
4.2.2 实例 116
4.3 高维多输入层神经网络质量模型 117

4.3.1 引言 117
4.3.2 高维多输入层神经网络的结构 118
4.3.3 高维多输入层神经网络的学习算法 120
4.3.4 实例 124
4.3.5 结论 126
4.4 RBF 神经网络产品质量模型 127
4.4.1 引言 127
4.4.2 高维 RBF 神经网络质量模型的建立 128
4.5 两种改进结构的 RBF 神经网络产品质量模型 132
4.5.1 分布式 RBF 网络在 1580 热联轧机控制中的应用 132
4.5.2 重叠式 RBF 网络在 1580 热联轧机控制中的应用 135
4.6 基于具体数据集的 BP 神经网络结构研究 138
4.7 小 结 147
第五章 基于模糊神经网络的产品质量模型 149
5.1 基于矩形函数系的模糊神经网络质量模型 150
5.1.1 基于矩形函数系的模糊神经网络模型 150
5.1.2 在热轧数据质量控制中的应用 154
5.1.3 应用实例 155
5.2 基于公式的模糊神经网络产品质量模型 157
5.3 模糊神经网络的记忆研究 158
5.3.1 一次性记忆 160
5.3.2 逐个记忆 165
5.3.3 结束语 167
5.4 小 结 168
第六章 基于多种模型的新产品新工艺设计 170
6.1 基于产品质量模型的新产品新工艺设计 171
6.1.1 用相关分析对产品质量模型进行校验分析 172
6.1.2 气瓶钢钢种的产品质量设计 177

6.1.3 基于 BP 神经网络质量模型的气瓶钢新产品设计 179
6.2 基于产品质量模型输出优化的新产品新工艺设计 181
6.2.1 神经网络输出的优化 181
6.2.2 用梯度法对输出的优化 182
6.2.3 用随机搜索方法进一步优化 186
6.2.4 在热轧新产品设计中的应用 190
6.3 基于 BP 神经网络产品逆质量模型和质量控制模型的新产品新工艺设计 191
6.3.1 基于 BP 神经网络逆质量模型的新产品新工艺设计 191
6.3.2 基于 BP 神经网络质量控制模型的新产品新工艺设计 193
6.4 基于模糊神经网络质量控制模型的新产品新工艺设计 195
6.4.1 质量控制模型与求解 195
6.4.2 求解理论分析 196
6.4.3 在热轧新产品新工艺中的应用 198
6.5 小 结 199
第七章 基于模糊神经网络的 SAS 应用软件 201
7.1 数据挖掘工具简介 202
7.2 建模设计系统的设计 202
7.3 建模设计系统的开发 207
7.4 小结 213
第八章 结束语 214
参考文献 218

第一章

引　论

- 数据挖掘技术及其研究现状
- 数据挖掘的对象
- 数据挖掘的主要技术
- 数据挖掘过程及结果解释
- 数据挖掘建模设计方案
- 冶金产品质量控制问题分析
- 本书的主要工作

1.1 数据挖掘技术及其研究现状

随着计算机技术的快速发展和它在生产过程中的广泛应用，企业生成、收集、存储和处理数据的能力大大提高，数据量与日俱增。众所周知，数据就是财富，但这种价值是隐含的[1~3]。为了从堆积如山的数据中找出真正有价值的东西——知识，从20世纪90年代人们就开始了对数据挖掘(Data Mining)的研究[4~7]。值得注意的是针对不同的应用领域，应该设计特定的数据挖掘方案，以求达到知识获取的高效性。针对冶金企业生产特性和质量控制问题，利用数据挖掘技术可以从其领域数据中获取定性和定量的知识，以帮助质量工程师改善产品质量和进行新产品的辅助设计。

1.1.1 知识获取与数据挖掘技术

在智能控制中，知识获取一直是研究的难点问题之一[8~11]。由于知识在智能系统中扮演着重要的角色，而知识获取又往往不能自动进行，故其一直被公认为是构造智能系统时的“瓶颈”。

领域知识获取的传统方法是通过知识工程师与领域专家交流，由知识工程师整理、总结专家的经验，把它们形式化，再输入到计算机中[12]。这个过程较多考虑了人的因素，而没有利用领域数据。

人工智能中专门研究知识获取的分支之一是机器学习[13]，它能够从数据中提取知识，但机器学习使用的数据是专门为其特别准备的，与现实世界中的数据有所不同。在20世纪60年代，统计学家们在基于计算机的数据分析中率先使用了数据挖掘这个术语[14]。进入20世纪80年代后，随着计算机技术的飞速发展，各行各业都开始广泛采用计算机及相应的信息技术进行运营和管理，企业的数据量与日俱增。人们迫切希望能从堆积如山的数据

中找出真正有价值的东西，为决策支持服务。在这样的背景下，数据挖掘在 20 世纪 90 年代成了国际上的热门话题。在此有必要提及的是数据挖掘与 KDD 的联系，数据挖掘有狭义和广义两层涵义，狭义的数据挖掘是指 KDD 过程的一个重要组成部分，称为 DM；而广义的数据挖掘等同于 KDD，另外就 KDD 本身而言，有的学者认为是 Knowledge Discovery in Data 的缩写，而比较常见的说法是指 Knowledge Discovery in Databases 的缩写。本书中取数据挖掘的广义的涵义。

目前在国际上对数据挖掘还没有一个统一的定义，其中一种比较有代表性的观点认为[15]：数据挖掘是从大量数据中提取出可信的、新颖的、有效的、具有潜在价值的并能被人理解的模式的处理过程，这种处理过程是非平常的过程。从技术角度看，它是从大量的、不完整的、有噪声的、模糊的、随机的实际数据中，提取隐含在其中的、人们不知道的、但又是潜在有用的信息和知识的过程。

数据挖掘技术实际上为知识获取指出了一条新路。可以对领域数据进行分析，找出蕴藏在数据背后有规律性的东西，在与专家的意见达成一致后，所发现的模式经解释后就可上升为知识，可用于构造智能系统中的知识库，或提供问题求解的对策。

1.1.2 数据挖掘研究与应用的现状

从数据挖掘发展的历史来看，它是一门从应用中发展起来的边缘学科，数据挖掘实际上是一个统计学、机器学习、数据库技术、人工智能等许多相关领域技术的结合体，这些相关领域现在都比较成熟，一旦把它们有效组织起来，就会收到前所未有的效果[16~20]；但是目前在数据挖掘的研究中最欠缺的还是在整体上对系统的组织策略的研究，也即从系统的观点来看，还缺乏一种全局优化的机制，系统整合的效率并不高[21]。

国际上对数据挖掘的研究如火如荼，从 1995 年起，每年都举行 KDD 大会，以供这一领域的研究人员交流探讨。另外在许多

国际会议中，也把 KDD 作为征文和讨论的一个领域。在美国国家科学基金会(NSF)的数据库研究项目中，KDD 被列为 20 世纪 90 年代最有价值的研究项目。ACM 于 1999 年专门开办了 SIGKDD 的刊物,目前这方面的专业期刊还有 Datamation 和“Knowledge Discovery and Data Mining”杂志。

从 20 世纪 90 年代以来，数据挖掘首先与数据仓库结合。数据仓库将异构的数据集成起来，经过数据清洗等过程变成一个可直接使用的数据资源。数据仓库带有自己的 OLAP(OnLine Analysis Process,联机分析处理)工具，可为用户或知识工人提供数据分析和决策服务，进行初步的数据分析。数据挖掘和 OLAP 最本质的区别在于，数据挖掘是一种挖掘性的分析工具，它主要是利用各种分析方法主动地去挖掘大量数据中蕴含的规律，而 OLAP 则是一种求证性的分析工具，即已有一个假设，通过 OLAP 来得到验证。OLAP 所采用的验证方法多是基于数据立方体法，即通过对数据立方体的切片、切块、旋转、钻取等操作来实现对数据立方体快速地多维存取。数据挖掘和 OLAP 这两种分析工具本身是相辅相成的，因为 OLAP 可以帮助人们提出假设，可以验证数据挖掘预测出的结果，而数据挖掘能够挖掘出一个结论。传统的数据环境基本上是数据操作型的，传统的信息系统只负责数据的增、删及修改操作，而在数据库的基础上可实现的工作就是 OLTP(OnLine Transaction Process,联机事务处理)。由于 OLTP 和 OLAP 在用户和系统的面向性、数据内容、数据库设计、访问模式方面有很大的不同，所以现在随着数据积累的不断增多，人们需要分析型的数据环境，于是就出现了数据仓库，以此为基础则可以实现 OLAP 和数据挖掘。我国一些大型企业的数据仓库正在建设和应用之中，目的就是利用数据挖掘工具提高其产品的竞争力，从而提高企业的竞争优势。

其次是数据挖掘与面向 Internet 和 Web 的数据。Internet 上的数据的最大特点是半结构化的。半结构化是相对于结构化和非结构化而言的。传统数据库中的数据结构性很强，称之为完全结构

化的数据，一本书、一张图片等无结构的可称之为完全无结构的数据。但是 Internet 上存在的页面具有一定的描述层次，存在一定的结构，所以将它称为半结构化的数据。Web 上网站的信息也可以看作是一个数据库，一个更大的、复杂性更高的数据库。Web 上的每一个站点就是一个数据源，每一个数据源都是异构的。每个站点的信息和组织形式都不完全一样，这就构成了一个巨大的、异构的数据库环境。如果想要利用这些数据进行数据挖掘，必须要研究站点之间异构数据的集成问题。只有将这些站点上的数据都集成起来，提供给用户一个统一的视图或视角，才有可能从巨大的数据资源中获取所需的东西。要解决上述问题需要寻找一个半结构化的数据模型和一项技术能够自动地从现有数据中将这个模型提取出来。因为半结构化数据模型和半结构化数据模型提取是面向 Internet 的数据挖掘技术实施的前提，所以他们成为数据挖掘研究领域的最大热点。

近年来，国内外已推出了一些数据挖掘的产品和应用系统，并且获得了一定的成功应用，得到了业界的广泛关注。国外有 SAS 公司的 Enterprise Miner、ISL 公司的 Clementine、Angoss 公司的 KnoeledgeSEEKER、RightPoint Software 公司的 DataCruncher 和 IBM 公司的 IBM Intelligent Mine 等等。

数据挖掘技术应用前景相当广泛，在政府管理决策、商业经营、科学研究和企业决策支持等各个领域都有其用武之地。世界上许多大公司都在自己的经营管理中使用了数据挖掘技术，例如：

美国钢铁公司和神户钢铁公司利用基于数据挖掘技术的 ISPA 系统，研究分析产品性能规律和进行质量控制，取得了显著效果。

通用电器公司(GE)与法国飞机发动机制造公司(SNECMA)，利用数据挖掘技术研制了 CASSIOPEE 质量控制系统，被三家欧洲航空公司用于诊断和测试波音 737 的故障，带来了可观的经济效益。该系统于 1996 年获欧洲一等创造性应用奖。