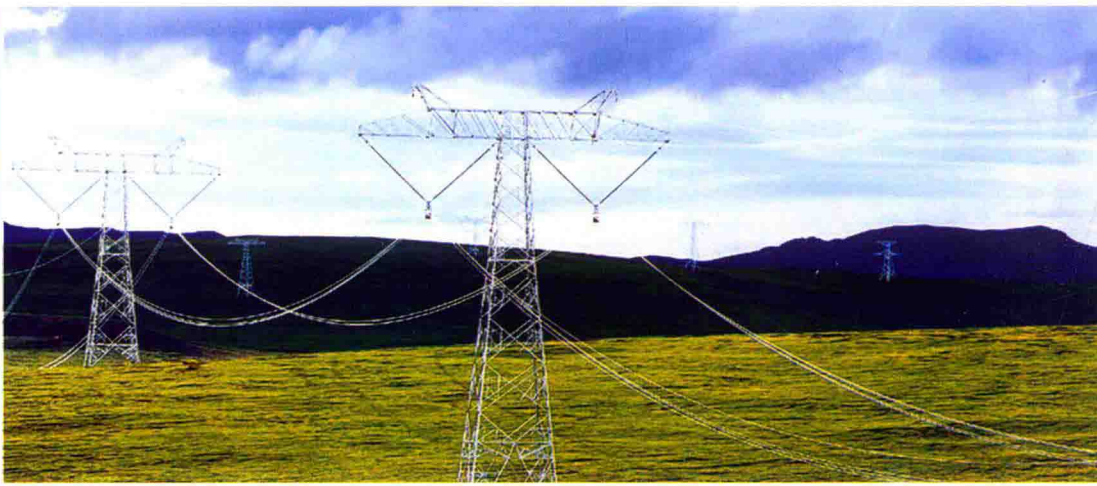
研究生教材

电能计量大数据及应用

李宁　袁铁江　孙谊媊　栗遇春　杨金成 等　著



中国电力出版社
CHINA ELECTRIC POWER PRESS

研究生教材

电能计量大数据及应用

李　宁　袁铁江　孙谊媚　栗遇春　杨金成　王　璐
张　科　王新刚　张建文　刘卫新　王　刚　刘国亮
张　龙　徐一晨　史旭东　段志尚　山宪武　著

中国电力出版社
CHINA ELECTRIC POWER PRESS

内 容 提 要

本书基于国网新疆电力公司在电能计量技术领域的理论研究和工程实践经验，对智能电能表及其数据采集系统的基本工作原理、结构、常见模块的故障及其产生的原因、基于大数据原理的故障判断和可靠性评估的基本原理和方法等进行了系统的介绍。

本书可为电力系统工程师、电气工程设计人员和电气工程专业的师生在相关领域的理论学习和工程实践提供一定的参考。

图书在版编目（CIP）数据

电能计量大数据及应用/李宁等著. —北京：中国电力出版社，2017.10

研究生教材

ISBN 978-7-5198-1245-4

Ⅰ.①电… Ⅱ.①李… Ⅲ.①电能计量－数据处理－研究生－教材 Ⅳ.①TB971

中国版本图书馆 CIP 数据核字（2017）第 250674 号

出版发行：中国电力出版社

地　　址：北京市东城区北京站西街 19 号（邮政编码 100005）

网　　址：http://www.cepp.sgcc.com.cn

责任编辑：王　娟　（010-63412522）

责任校对：常燕昆

装帧设计：张俊霞

责任印制：吴　迪

印　　刷：北京雁林吉兆印刷有限公司

版　　次：2017 年 10 月第一版

印　　次：2017 年 10 月北京第一次印刷

开　　本：710 毫米×1000 毫米　B5 开本

印　　张：9.25

字　　数：153 千字

定　　价：35.00 元

前　言

Preface

随着国家电网公司“三集五大”体系建设的深入推进，“大营销”推广的用电信息采集系统建设已具规模，特别是智能电能表已获得大范围的推广应用。在电力系统采集系统中，每年产生大量的智能电能表采集数据，但对这些数据的处理还停留在简单的数据备份、查询及简单统计阶段，并没有对这些电力数据进行整理并进行深入的分析。为满足广大电力工作者需要，国网新疆电力公司电力科学研究院计量中心组织撰写了本书。本书基于国网新疆电力公司在电能计量技术领域的理论研究和工程实践经验，结合实际案例系统地介绍电能计量大数据相关的理论和方法。

本书第 0 章对数据挖掘的基本概念进行了介绍，对电能计量数据的应用背景、国内外研究现状及其发展趋势等做了简要介绍。第 1 章对智能电能表及其数据采集系统的基本工作原理和结构等进行了介绍。第 2 章介绍了智能电能表常见模块的故障及其产生原因。第 3 章介绍了基于智能电能表计量数据的计量故障判断方法及其数学模型。第 4 章介绍了基于计量大数据的智能电能表可靠性预计的基本理论和方法。

通过阅读本书，读者能够系统地了解基于大数据原理的智能电能表的故障预测和可靠性评估的基本原理和方法。希望本书能够对电力系统工程师、电气工程设计人员和电气工程专业的师生在相关领域的理论学习和工程实践提供一定的参考。

在写作过程中，作者得到包括大连理工大学、清华大学等单位的大力支持，也得到了业界专家、学者们的无私帮助，在此一并表示感谢。

本书的理论与应用研究有待于进一步探索和完善，由于作者的知识水平和经验，书中的观点和结论若有不妥之处，望读者原谅和批评指正。

目　录

Contents

电能计量大数据及应用

0 绪 论

0.1 引言

全球性的能源危机，使基于化石能源架构的传统电网可持续发展面临重大挑战。人们在实现现有电网的精细化管理、新能源的开发、促进传统电网建设转型等方面做了不懈的努力，继而催生了大数据技术在电力系统的广泛应用，并取得初步成果。这直接推动了全球能源互联网、智能电网的建设与发展。建设具备着自愈、清洁、经济等优点的智能电网成为电网发展的一个重要趋势，智能电网的建设与推行已成为我国电力事业发展目标。

国网智能电网研究院的数据显示，截至 2016 年底，国家电网公司管理结构化数据为 49.75TB，非结构化数据为 213TB，营销基础数据为 130TB，用电信息采集数据达 43TB，且信息化数据平均每天以 10TB 的速度增长。因此，研究和应用大数据已成为提质增效和推动电网发展方式、公司发展方式转变的迫切要求。国家电网公司“三集五大”体系和坚强智能电网建设，积累了体量大、类型多、价值高、速度快等典型大数据特征的运营数据，具备了推广大数据应用的基础条件。

与此同时，随着经济与社会的发展，新能源的接入不断增长，电力行业的工作方式以及人民的生活方式都已经发生了深刻的变化，这些变化对电能计量提出了新的要求。需量电价和分时电价的实施，电能质量监控和无功计量的应用，预付费、网上缴费、远程停送电等电子商务模式在电力生活中的发展，传统的技术手段已不能经济地满足电网业务型管理需求，在优化电网管理、创新电力企业业务模式、提升电力企业价值方面难以发挥应有的作用。

继互联网、物联网、云计算的不断发展及智能终端的普及，海量复杂多样的数据呈现出爆炸式的增长，促使着大数据时代的到来。作为重要的生产因素，大数据已成为蕴含巨大潜在价值的战略资产，推动着产业升级和崛起，影响着科学思维与研究方法的变革。然而，大数据在依托其丰富的资源储备和借助强大的训—算技术发挥优势的同时，也带来了挑战。海量、动态及不确定的数据使得传统数据处理系统而面临着存储和计算瓶颈，同时，就如何从复杂的大数据中实时、快速地挖掘出有价值的信息和知识，传统的数据挖掘技术自身受限的功能已无法满足用户的需求。因此，大数据环境下需要一种适用技术，即大数据挖掘，来应对当下存在的挑战。鉴于大数据环境下，参照传统数据挖掘的构建思想开发的挖掘系统无法提供令用户满意的服务，因此为满足建设与

应用需求大数据挖掘的架构则必不可少。

0.2 数据挖掘概述

人们把原始数据看作是形成知识的源泉，就像从矿石中采矿一样，原始数据可以是结构化的，如关系数据库中的数据，也可以是半结构化的，如文本、图形、图像数据，甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，可以是非数学的，也可以是演绎的或是归纳的。发现了的知识可被用于信息管理、查询优化、决策支持、过程控制等，还可用于数据自身的维护。可以说，数据挖掘是一门广义的交叉学科，它汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。数据挖掘技术从一开始就是面向应用领域，它不仅是面向特定数据库的简单检索查询调用，而且，要对数据进行微观乃至宏观的统计、分析、综合和推理，以指定实际问题的求解，企图发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测。

大数据挖掘是从体量巨大、类型多样、动态快速流转及价值密度低的大数据中挖掘有巨大潜在价值的信息和知识，并以服务的形式提供给用户。与传统数据挖掘相比，其同样是以挖掘有价值的信息和知识为目的。然而，就技术发展背景、所面临的数据环境及挖掘的广度深度而言，两者却存在差异。

0.2.1 数据挖掘的定义

Usama M-Fayyad 等 1989 年在美国底特律举行的第十一届国际人工智能学术会议给出了 KDD 最初的描述性定义，即 KDD 就是数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解模式的非平凡过程。可是这次会议并没有给出数据挖掘的定义，实际上这导致许多学者在这两个术语上的混用。Meta Group 1996 年给出了数据挖掘（Data Mining，DM）的定义，即“Data Mining is the application of artificial intelligence（AI）techniques（Neural Network，Fuzzy Logic，Genetic Algorithms）to large quantities of data，to discovery hidden trends，pattern，and relationship”。现在数据挖掘界普遍认为数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的，但又是潜在有用的信息和知识的过程。

这个定义包括多层含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识应该可接受、可理解、可运用；并不要求发现放之四海皆准的知识，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明。实际上，所有发现的知识都是相对的，是有特定前提和约束条件、面向特定领域的，同时还要易于被用户理解，最好能用自然语言表达所发现的结果。数据挖掘所得到的知识应具有先前未知这个特征。先前未知的知识是指预先未曾预料到的，甚至是违背直觉的信息或知识，挖掘出的信息越是出乎意料，就可能越有价值。

那么何为知识？从广义上理解，数据、信息也是知识的表现形式，数据是指有关事实的集合，记录和事物有关的原始信息。信息是根据表示数据所用的约定，赋予数据的意义。但是人们更把概念、规则、模式、规律和约束等看作知识，因为这些东西是对数据包含的信息更抽象的描述。人们把数据看作是形成知识的源泉，好像从矿石中采矿或淘金一样。原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形和图像数据；甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现的知识可以被用于信息管理、查询优化、决策支持和过程控制等，还可以用于数据自身的维护。因此，数据挖掘是一门交叉学科，它把人们对数据的应用从低层次的简单查询，提升到从数据中挖掘知识，提供决策支持。在这种需求牵引下，汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员，投身到数据挖掘这一新兴的研究领域，形成新的技术热点。而从商业角度上来看，数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。

因此，数据挖掘可以说是一类深层次的数据分析方法。数据分析本身并不是新鲜的东西，它已经有很多年的历史，只不过以前数据收集和分析的目的是用于科学研究，而且限于当时计算的能力，对大数据量进行分析的复杂数据分析方法发展缓慢。现在，由于各行业业务自动化的实现，商业领域产生了大量的业务数据，分析这些数据主要是为商业决策提供真正有价值的信息，进而获得利润。但所有企业面临的一个共同的问题：企业数据量非常大，而其中真正有价值的信息却很少。因此从大量的数据中经过深层分析，获得有利于商业运作、提高竞争力的信息，就像从矿石中淘金一样，数据挖掘也因此而得名。因

此，数据挖掘也可以描述为：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的先进、有效的方法。

0.2.2 数据挖掘发展背景

在技术的先进程度及数据的体量及复杂程度和处理分析能力方面，传统数据挖掘没有大数据时代的充实环境技术条件，在数据库、数据仓库及互联网发展等背景下，实现了从独立、横向到纵向数据挖掘的发展。而大数据挖掘则在大数据背景下得益于云计算、物联网、移动智能终端等技术产生与发展。它针对大数据的特征及现存挖掘系统面临的问题，借助先进技术加以系统地整合与改进。相比传统数据挖掘已相当成熟的应用、算法研究及系统工具开发的研究与应用还处于不断发展中，对于海量数据的挖掘主要由基于云计算进行相关技术的集成来实现。

0.2.3 数据挖掘方法分类

在数据挖掘中，数据分为训练数据、测试数据和应用数据三类。数据挖掘的关键是在训练数据中发现事实，以测试数据作为检验和修正理论的依据，把知识应用到数据中。数据挖掘利用了分类、关联规则、序列分析、群体分析、机器学习、知识发现及其他统计方法，能够通过数据的分析，预测未来。数据挖掘有以下几种常用方法：

(1) 关联规则挖掘。1993 年，R-Agrawal 等人首先提出了关联规则挖掘问题，描述的是数据库中一组数据项之间某种潜在关联关系的规则。一个典型的例子是：在超市中，90%的顾客在购买面包和黄油的同时，也会购买牛奶。直观的意义是：顾客在购买某种商品时有多大的倾向会购买另外一些商品。找出所有类似的关联规则，对于企业确定生产销售、产品分类设计、市场分析等多方面是有价值的。关联规则是数据挖掘研究的主要模式之一，侧重于确定数据中不同领域之间的关系，找出满足给定条件下的多个领域间的依赖关系。数据项之间的关联，即根据一个事务中某些数据项的出现可以导出另一些数据项在同一事务中的出现。在关联规则挖掘法的研究中，算法的效率是核心问题，如何提高算法的效率是要解决的关键问题。目前最有影响的是 Apriori 算法，实现探查逐级挖掘。Apriori 的性质是频繁项集的所有非空子集都必须是频繁的。

（2）决策树方法。决策树（Decision Tree）根据不同的特征，以树形结构表示分类或决策集合，产生规则和发现规律。利用信息论中的互信息（信息增益）寻找数据库中具有最大信息量的字段，建立决策树的一个结点，再根据字段的不同取值建立树的分枝。在每个分枝子集中，重复建立树的下层结点和分枝的过程，即可建立决策树。

决策树起源于概念学习系统 CLS（Concept Learning System），其思路是找出最有分辨能力的属性，把数据库划分为多个子集（对应树的一个分枝），构成一个分枝过程，然后对每一个子集递归调用分枝过程，直到所有子集包含同一类型的数据。最后得到的决策树能对新的例子进行分类。CLS 的不足是它处理的学习问题不能太大。为此，Quinlan 提出了著名的 ID3 学习算法，通过选择窗口来形成决策树。从示例学习最优化的角度分析，理想的决策树分为 3 种：①叶子数最少；②叶子结点深度最小；③叶子数最少且叶子结点深度最小。寻优最优决策树已被证明是 NP 困难问题。ID3 算法借用信息论中的互信息（信息增益），从单一属性分辨能力的度量，试图减少树的平均深度，却忽略了叶子数目的研究。其启发式函数并不是最优的，存在的主要问题有：

1）互信息的计算依赖于属性取值的数目多少，而属性取值较多的属性并不一定最优。

2）ID3 是非递增学习算法。

3）ID3 决策树是单变量决策树（在分枝节点上只考虑单个属性），许多复杂概念表达困难，属性间的相互关系强调不够，容易导致决策树中子树的重复或有些属性在决策树的某一路径上被检验多次。

4）抗噪声性差，训练例子中，正例和反例的比例较难控制。

针对上述问题，出现许多较好的改进算法，刘晓虎等在选择一个新属性时，并不仅仅计算该属性引起的信息增益，而是同时考虑树的两层结点，即选择该属性后继续选择属性带来的信息增益。Schlimmer 和 Fisher 设计了 ID4 递增式算法，通过修改 ID3 算法，在每个可能的决策树结点创建一系列表，每个表由未检测属性值及其示例组成，当处理新例时，每个属性值的正例和反例递增计量。在 ID4 的基础上，Utgoff 提出了 ID5 算法，它抛弃了旧的检测属性下面的子树，从下面选择属性构造树。此外，还有许多算法使用了多变量决策树的形式，著名的 C4.5 系统也是基于决策树的。

（3）神经网络方法。模拟人脑神经元方法，以 MP 模型和 HEBB 学习规则为基础，建立了 3 大类多种神经网络模型，即前馈式网络、反馈式网络、自

组织网络。神经网络是一种通过训练来学习的非线性预测模型，可以完成分类、聚类等多种数据挖掘任务。

神经网络（Neural Network）是由大量的简单神经元，通过极其丰富和完善的连接而构成的自适应非线性动态系统，并具有分布存储、联想记忆、大规模并行处理、自组织、自学习、自适应等功能。网络能够模拟人类大脑的结构和功能，采用某种学习算法从训练样本中学习，并将获取的知识存储于网络各单元之间的连接权中，神经网络和基于符号的传统 AI 技术相比，具有直观性、并行性和抗噪声性。目前，已出现了许多网络模型和学习算法，主要用于分类、优化、模式识别、预测和控制等领域。在数据挖掘领域，主要采用前向神经网络提取分类规则。

神经网络模拟人的形象直觉思维，其中，最大的缺点是“黑箱”性，人们难以理解网络的学习和决策过程。因此，有必要建立“白化”机制，用规则解释网络的权值矩阵，为决策支持和数据挖掘提供说明，使从网络中提取知识成为自动获取的手段。通常有两种解决方案：①建立一个基于规则的系统辅助。神经网络运行的同时，将其输入和输出模式给基于规则的系统，然后用反向关联规则完成网络的推理过程。这种方法把网络的运行过程和解释过程用两套系统实现，开销大，不够灵活；②直接从训练好的网络中提取（分类）规则，这是当前数据挖掘使用得比较多的方法。从网络中采掘规则，主要有以下倾向：

1）网络结构分解的规则提取。它以神经网络的隐层结点和输出层结点为研究对象，把整个网络分解为许多单层子网的组合。这样研究较简单的子网，便于从中挖掘知识。Fu 的 KT 算法和 Towell 的 MofM 算法是有代表性的方法。KT 算法的缺点是通用性差，且当网络比较复杂时，要对网络进行结构的剪枝和删除冗余结点等预处理工作。

2）神经网络的非线性映射关系提取规则。这种方法直接从网络输入和输出层数据入手，不考虑网络的隐层结构，避免了基于结构分解的规则提取算法的不足。Sestito 等人的相似权值法以及 CSW 算法（将网络输入扩展到连续取值），是其中的两种典型算法。当然，在数据挖掘领域，神经网络的规则提取还存在许多问题，即如何进一步降低算法的复杂度，提高所提取规则的可理解性及算法的适用性，研究提取规则集的评估标准和在训练中从神经网络动态提取规则，以及及时修正神经网络并提高神经网络性能等，都是进一步研究的方向。

（4）粗集方法。粗集（Rough Set）理论的特点是不需要预先给定某些特

征或属性的数量描述，如统计学中的概率分布，模糊集理论中的隶属度或隶属函数等，而是直接从给定问题出发，通过不可分辨关系和不可分辨类确定问题的近似域，从而找出该问题内在规律。粗集理论同模糊集、神经网络、证据理论等其他理论均成为不确定性计算的一个重要分支。

在粗集理论中，分别用 3 个近似集合来表示正域、负域和边界。在数据挖掘中，从实际系统采集到的数据可能包含各种噪声，存在许多不确定的因素和不完全信息有待处理。传统的不确定信息处理方法，如模糊集理论、证据理论和概率统计理论等，因需要数据的附加信息或先验知识（难以得到），有时在处理大量数据的数据库方面无能为力。粗集作为一种软计算方法，可以克服传统不确定处理方法的不足，并且和它们有机结合，可望进一步增强对不确定、不完全信息的处理能力。

粗集理论中，知识被定义为对事物的分类能力。这种能力由上近似集、下近似集、等价关系等概念体现。因为粗集处理的对象是类似二维关系表的信息表（决策表）。目前，成熟的关系数据库管理系统和新发展起来的数据仓库管理系统，为粗集的数据挖掘奠定了坚实的基础。粗集从决策表挖掘规则、辅助决策，其关键步骤是求值约简或数据浓缩，包括属性约简。Wong SK 和 Ziarko W 已经证明求最小约简是一个 NP - hard 问题。最小约简的求解需要属性约简和值约简两个过程，决策表约简涉及核和差别矩阵两个重要概念。一般来讲，决策表的相对约简有许多，最小约简（含有最小属性）是人们期望的。另一方面，决策表的核是唯一的，它定义为所有约简的交集，所以，核可以作为求解最小约简的起点。差别矩阵突出属性的分辨能力，从中可求出决策表的核以及约简规则。约简算法 MIBARK，但对最小约简都是不完备的。此外，上述方法还只局限于完全决策表。Marzena K 应用差别矩阵，推广了等价关系（相似关系）、集合近似等概念，研究了不完全决策表（属性的取值含有空值的情况）的规则的发展问题，从而为粗集的实用化迈出了可喜的一步。Marzena K 还比较了几种不完全系统的分析方法，得出如下结论：①一个规则是确定的，如果此规则在原不完全系统的每个完全拓展中是确定的；②删除从不完全决策表包含空值的对象后，采掘的知识可能成为伪规则。

粗集的数学基础是集合论，难以直接处理连续的属性。而现实决策表中连续属性是普遍存在的，因此，连续属性的离散化是制约粗集理论实用化的难点之一，这个问题一直是人工智能界关注的焦点。连续属性的离散化根本出发点，是在尽量减少决策表信息损失的前提下（保持决策表不同类对象的可分辨

关系），得到简化和浓缩的决策表，以便用粗集理论分析，获得决策所需要的知识。最优离散化问题（离散的切点数最少）已被证明是 NP - hard 问题，利用一些启发式算法可以得到满意的结果。总体上讲，现有离散化方法主要分为非监督离散化和监督离散化。前者包括等宽度（将连续值属性的值域等分）和等频率离散化（每个离散化区间所含的对象相同）。非监督离散化方法简单，它忽略了对象的类别信息，只能用在属性具有特殊分布的情况。针对上述问题，监督离散化方法考虑了分类信息，提高了离散效果。目前，比较有代表性的监督离散化方法有以下几种：①Holte 提出了一种贪婪的单规则离散器（one rule dis-cretizer）方法；②统计检验方法；③信息熵方法等。这些方法各有特点，但都存在一个不足，即每个属性的离散化过程是相互独立的，忽略了属性之间的关联，从而使得离散结果中含有冗余或不合理的分割点。针对这个问题，有人给出了一种连续属性的整体离散化方法，实验表明，它不仅能显著减少离散化划分点和归纳规则数，而且提高了分类精度。连续属性离散化目前还存在的问题是缺乏递增的离散化方法，即当新的对象加入决策表时，原有的分割点可能不是最优或最满意的。

粗集理论和其他软计算方法的结合，能够提高数据挖掘能力。Mohua Banerjee 等利用集理论获得初始规则集，然后，构造对应的模糊多层神经网络（规则的置信度对应网络的连接权），训练后可得到精化的知识。粗集与其他软计算方法的集成是数据挖掘的一种趋势。目前，基于粗集的数据挖掘在以下方面有待深化：

1）粗集和其他软计算方法的进一步结合问题。

2）粗集知识采掘的递增算法。

3）粗集基本运算的并行算法及硬件实现，将大幅度改善数据挖掘的效率。已有的粗集软件适用范围还很有限。决策表中的实例数量和属性数量受限制。面对大量的数据，有必要设计高效的启发式简化算法或研究实时性较好的并行算法。

4）扩大处理属性的类型范围，实际数据库的属性类型是多样的，既有离散属性，也有连续属性；既有字符属性，也有数值属性。粗集理论只能处理离散属性，因此，需要设计连续值的离散算法。

（5）遗传算法。遗传算法（GA，Genetic Algorithms）是模拟生物进化过程，利用复制（选择）、交叉（重组）和变异（突变）3 个基本算子优化求解的技术。遗传算法类似统计学，模型的形式必须预先确定，在算法实施的过程