

武汉理工大学

硕士学位论文

基于粗糙集理论的金矿矿化信息挖掘与分析

姓名：朱雅琼

申请学位级别：硕士

专业：环境科学

指导教师：袁艳斌;崔巍

20071101

Abstract

At present, the mineral data are characterized by time-space and multi sources and the means for acquiring mineral information are diversiform, which make the mineral resources exploration and evaluation increasingly complicated. GIS technology has provided the capability of integrated management, query and research of multi-source geological information. Computer science and mathematical geology is the technical support for optimal selection of target and comprehensive analysis. The exploration of mineral information and quantitative analysis of multivariate geological information are the trend in future development.

Geological phenomena and ore-forming process have inner complexity, which has transcended the confines of linear means. The sampling observation in deposit research is random, the information extract is deficient, and the linear prognosis model is limit, which make the research result in geology impossible to be accurately recurrence, so that most deduced prognosis have multi solutions. Therefore, it's necessary to study on the application of nonlinear theory and the extract means of faint information, which is a hotspot in qualitative and quantitative geological research.

In mineral exploration research, the field data collecting and the indoor information disposing are affected by human thinking mode. Rough Set can combine with qualitative and quantitative genes, independent of any prior knowledge, absolutely aim at data to evaluate the contribution ratio of mineral genes scientifically, extract the relation and rules between geological genes through numerous data, and provide scientific basis for prospecting evidences.

The work is supported by National Science Foundation of China. Based on geology and metallogenic regulation, make use of mathematic tools and GIS technique, and analyze geological variables and metallogenic probability quantitatively based on Rough Set to get best combination of variables. Apply characteristic analysis and Neural Network to establish logical prospecting model for complex geological problem with multi goals and multi genes, analyze deposit variable quantitatively, to meet the goal of target optimization.

Multi-source data are integrated on GIS platform. Considering metallogenic background, data disposal environment and data quality, essential mineral information of geological cells are extracted to establish decision table. Attributes are reduced through Rough Set to get the attribute core. Based on the reduced attribute table, approximate set is acquired through Variable Precision

Rough Set model. Certain rules are acquired from lower approximation and the attribute values are reduced by applying decision matrix. The reduced result show good agreement with practice.

The main idea of Rough Set is the expression and reduction of knowledge and the main function is to find out the mineral prospecting information of geological evidence. The combination of Rough Set and characteristic analysis indicates the error of the model establish by reduced variables is minor. The combination of Rough Set and Neural Network can reduce complexity of the network and simplify the model. Results of two models are in substantial agreement, which indicates the integrated prognosis method based on Rough Set has a certain reference value for metallogenic prognosis. The method is worth further research to inherit and develop the mathematic geology.

Key Words: Rough Set, mineral information, quantitative prognosis, characteristic analysis, neural network

独 创 性 声 明

本人声明，所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含任何其他人已经发表或撰写过的研究成果，也不包含未获得武汉理工大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：朱雅琼 日期：2007.11.19

关于论文使用授权的说明

本人完全了解武汉理工大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

研究生签名：朱雅琼 导师签名：袁锦斌 日期：2007.11.19

第 1 章 绪论

1.1 研究背景

矿产资源评价与物探、化探、遥感、地理信息系统(GIS)、钻探、坑探等工程勘察技术关系密切, 数据处理及间接信息提取方面与计算机技术也密不可分, 图 1-1 为矿产资源评价与各学科的关系。目前地矿信息获取手段多样化导致海量信息的多解性, 地矿数据的时空耦合特性以及多源化导致成矿专属性和特殊性, 造成了矿产资源勘查和评价的不确定性, 如何有效地利用这些数据以提高矿产勘查效果, 已成为国内外地学工作者共同关心的问题。因此, 两方面的工作一直在不间断地进行: 一是加强成矿地质理论和实验研究, 深入了解各类矿床形成的环境和条件以及矿床分布规律及产出特征; 二是加强找矿技术方法研究, 进一步查明指示矿床存在的各种标志和现象, 有针对性地开发识别、获取、加工、分析和解释海量找矿信息的手段^[1]。

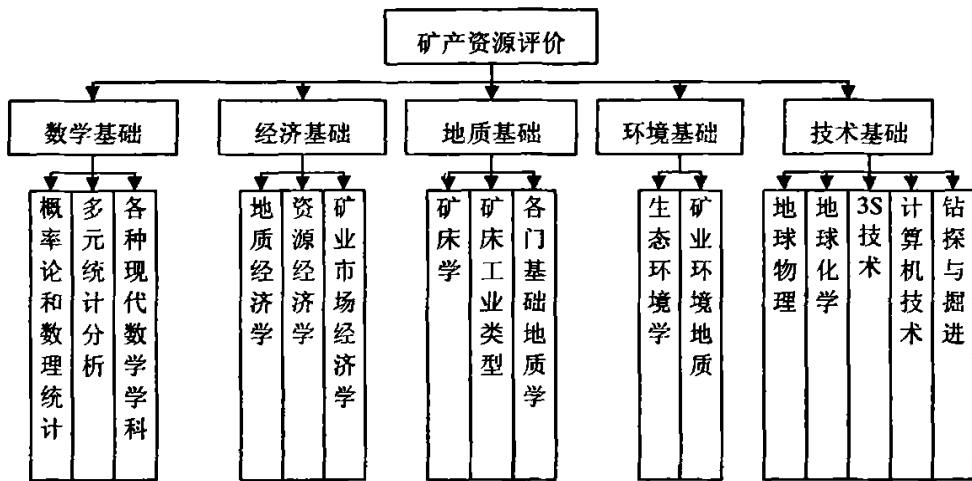


图 1-1 矿产资源评价与各学科的关系

长期以来, 地质学家们运用传统的观察、比较、历史分析等研究方法, 定性描述地质现象和地质过程, 已形成一套成熟的野外地质研究方法。然而, 随着社会经济和生产的发展, 以计量地学为主导的信息时代的到来, 要求对资源进行更为精确的定量分析和评价。如当代 GIS 技术提供了计算机辅助下的多源地学信息集成管理和查询检索能力, 而且可在经验与模型的指导下, 通过各种

空间分析方法建立相应的信息模型,对与成矿有关的各种空间信息各环节进行综合分析解释,确定成矿的有利地区或地段^[2~5]。计算机科学和数学地质建模方法是找矿靶区优选和综合评价的重要技术支持,成矿信息模型的全面探索和多元地学信息的定量化分析是当今矿产勘查和成矿预测评价的主流方向^[6~9],定性和定量研究只是成矿预测的不同环节的工作内容。

矿床统计预测是最早应用于矿产定量预测及评价的数学方法,依据的资料及数据可以是单一的地质变量或物、化探变量,也可以是地、物、化、遥等各种数据的综合信息。其成果形式体现为四定,即定成矿远景区空间位置,定矿产资源数量及质量,定成矿与找矿概率及定探矿因素或找矿标志的有利数值区间^[10]。然而,地质现象与成矿过程具有内在复杂性,超越了线性方法讨论的范畴^[11,12],目前单一确定型或随机型数学模型,都不足以表达复杂的地质历史过程。矿床观察研究的抽样性和随机性,现行信息提取的饱和性和不充分性,线性预测模型的局限性,造成地质学中大多数实验结果不可能准确地重复和再现,很多推断预测的成果具有多解性。为此,必须研究在强干扰下、深部、隐蔽及微弱信息的提取和非线性理论的应用^[6]。

勘查信息的多解性和不确定性,与地质专家对地质现象的认识层次以及专家思想并存,致使我们在矿产勘查的量化研究方面,从最初的野外数据采集到室内测试等的取样,都被动地受到人的干扰,在信息处理方面,同样受到人思维方式的影响。理想的方法应该是不受人为因素在认知和取样上的影响,不依赖假设,完全基于野外观测和室内测试等数据主动确定在不同分布和地质条件下矿床变量的信息量预测方法。粗糙集理论最大优点就是能够提供这一工作的技术支持。粗糙集理论可以同时考虑定性和定量因素,不依赖任何先验知识,完全基于数据本身科学地评估地矿因素的贡献率,从海量的数据中挖掘出各地质因素之间的联系和规则,找出控矿因素并为找矿标志的确立提供科学的依据。

1.2 研究意义

预测过程既是知识驱动又是数据驱动,如果所分析的数据是不完备的,直接影响预测效果。因此矿产预测的必要前提是多源地质数据的收集、存储、管理以及如何从这些数据中提取最有价值的信息。我们把地理信息系统平台的应用和数据分析模型的建立作为本研究的核心。目前,地学问题的研究中常用的经典预测方法包括多元统计,时间序列分析和马尔科夫法等,这些方法大多是建立在概率统计基础上,上述理论方法应用广泛,已成为成矿预测工作中必不

可少的应用工具^[13,14]。但是,地质变量的空间局域性、连续性、各向异性以及高度的非线性关系,造成了以上方法都存在局限性。在矿产预测中引入非线性科学,必将大大提高矿产预测的水平和可靠性。

本课题来源于国家自然科学基金资助项目“粗糙集支持下特征矿化信息挖掘的粒子群演化方法”,项目号 NO.40572166。该项目以矿产勘查的具体理论和方法流程为指导,考察矿化信息定量化研究的多种方法模型,基于粗糙集理论挖掘特征矿化信息,分析评价现有多种矿化信息量化处理以及特征提取的理论基础和解算方法的缺陷,利用多智能体技术改进和完善适应特征矿化信息提取的新型算法,研究新型算法下特征矿化信息挖掘的适用性,使其满足当前矿产预测信息组合有效数值区间的精度水平要求,并适于 GIS 系统来实现多源矿化信息预处理以及后处理要求。

矿产预测中不确定信息的定量化分析是矿产勘查中决策支持的重要过程,不确定性因素的度量能够反映预测的准确程度,因此引进的数学模型就无可厚非必须能够处理非线性、不确定性和不精确性方面的问题。为了有效地提取成矿预测信息,有必要客观地对诸多变量进行筛选,减少变量个数,突出成矿密切相关的变量,得到最佳变量组合。粗糙集理论不需要对知识或数据给出主观评价,不需要提供除问题所需数据集之外的任何先验信息,可以对非数值型数据进行编码赋值,对连续型数据离散化,将两种数据结合分析,计算各种属性的重要度和特征信息之间的依赖度,提取规则知识。粗糙集理论运用于地矿勘查与矿产预测领域是一项前沿性的科学研究,也是解决地学非线性复杂问题新的尝试。

1.3 研究内容

本研究针对复杂不确定系统的特性,以地质、成矿规律研究为基础,以数学为工具,以计算机为手段,基于粗糙集理论对有关的地质变量、矿化信息特征与矿床成矿可能性大小进行量化分析,获取控矿变量的最优组合和成矿区间,在特征分析法、神经网络模型的预测理论与方法的基础上,对具有多目标、多因素的复杂不确定地矿问题构建合理的预测模型,定量分析矿床变量值,找到成矿区间,从而达到靶区优选的目的。本论文研究内容主要包括以下几点:

(1) 学习和应用粗糙集理论对地矿信息进行分析。结合多源信息,考虑成矿背景、数据获取环境、数据质量等因素,提取相关信息,组成数据集,构建决

策表。利用粗糙集对属性进行约简,找出各个地矿信息对成矿的关联度和重要度,得到规则知识。

(2) 粗糙集与特征分析法、神经网络模型的结合应用。粗糙集理论的计算方法是知识的表达和约简,可以描述对象组成的集合之间的关系,从而提取规则知识。要实现矿产信息的定量分析,必须构造连续特征函数,即要查明各种控矿因素和找矿标志的找矿信息量,从而计算矿床变量值。

(3) 基于 GIS 的数据管理。矿产资源评价涉及的所有信息几乎都直接或间接地与空间位置有关,都属于地理信息的范畴。矿产资源评价的过程就是信息的搜集、整理、处理、成矿信息的提取、综合分析、成矿区带或找矿靶区的确定以及成果表示的过程。作为空间信息管理系统的 GIS 可贯穿于矿产资源评价的整个过程,为空间数据的认识和分析提供了一个方便的平台。

1.4 国内外研究现状

1.4.1 矿产资源评价现状分析

成矿预测的基本目的是能够预测矿床的位置,并大体知道这些矿床的类型、规模和品位。成矿定量预测结果可作为找矿勘探工作部署的依据,减少找矿勘探工作的盲目性和风险性,增加预见性,从而提高找矿工作的效率。定量预测必须以基本的地质认识和成矿规律为前提,对控矿地质因素和控矿类型的详细研究是成矿预测的关键问题^[15]。定量预测的发展过程就是计算机技术、数学模型或预测方法不断丰富过程。

在技术支持方面,七十年代末地质学家们就开始尝试在矿产资源评价中应用 GIS 技术,主要是实现多源地学信息的集成管理。经过二十年的努力,在空间数据库的建立、多种成矿信息综合分析方法的研究与应用、基于 GIS 的矿产资源评价专用软件的开发以及如何合理地组织人力资源适应新技术的应用要求等方面取得了长足的进步。

在定量预测模型方面, Gorelov(1999)强调用地球物理信息定量圈定地质异常的作用,并利用重磁场组合异常指数研究矿田地质结构异常,圈定找矿可行地段,从而开辟了应用“非相似类比”法开展矿产资源潜力评价的新途径。Harris 等(1993)根据相对例外原理提出了应用“一致性地质单元”代替传统的“网格单元”的矿产资源定量评价方法。王世称等(1990)提出了综合信息找矿与多方法、多测度矿产资源定量评价的思路。赵鹏大等(1999)经过多年的实践和探索,

特别是在定量预测方面开展大量工作的基础上,系统总结出矿床统计预测的基本理论、准则和方法,将矿床统计预测的基本理论概括为相似—类比理论、求异理论、定量组合控矿理论,以现代计算机技术和信息处理技术为手段,通过对地质、地球物理、地球化学和遥感地质异常信息的提取、转换和合成等一系列信息处理过程,最终应用多学科信息圈定综合致矿信息区,达到矿产资源定量评价的目的^[16]。

在定量预测方法方面,已经逐步探索出各种有效的数学方法。20世纪50~70年代,主要将概率统计及多元统计等定量方法用于矿产资源定量评价,用于处理物化探等定量数据;80年代初期,人们认识到地质数据中包含大量的定性数据,使得数量化理论得到发展;80年代中后期,模糊集方法流行,反映了人们对地质现象、过程模糊性的认识;90年代以来,灰色理论、人工神经网络、分形理论等方法流行,反映了人们对地质现象、过程的非线性认识^[17~19]。统计预测方法可以克服人为干预,提高系统分析的效率,但是样本需求量大,只能处理数值型数据,对于地质中的一些非线性问题的描述比较薄弱。模糊数学能够较好的处理成矿远景区与非成矿远景区、成矿有利与成矿不利等诸多模糊概念,但是模糊数学不能解决评价指标相关造成的信息重复问题,对隶属度与隶属函数的确定带有强烈的主观色彩,多目标模型中隶属度的确定也较为繁琐。分形理论被认为是非线性科学研究中取得的最重要成果之一,但是对地质因素之间的复杂关系考虑较少,且不能同时结合多类型数据进行分析。神经网络擅长于表达那些只有数据而无法用公式表达的系统,遗传算法具有全局寻优的特点,被认为是比较适合非线性模型预测处理的优化技术,称为现在地质学者们研究的热点,但这两种方法对样本要求较高,且易产生算法上的“早熟”,面对类型繁多、高信息维度的地质数据,往往使系统变得十分复杂,影响应用效果,仍然存在着一一定的不足和局限性^[11,17]。

地学领域非传统矿产资源定量预测方法正在不断的丰富和完善,其新理论方法都以数字化和定量化识别、揭示和提取新型的、隐式的和深层次的成矿地质信息为重点,把预测对象放到预测地区的地质成矿时空及成因演化系统中去考查,通过揭示区域成矿的本质规律实现矿床预测^[12]。

1.4.2 粗糙集应用现状分析

粗糙集理论最初是由波兰数学家 Z. Pawlak 于 1982 年提出来的。欧洲国家比较注重理论研究,北美学者比较注重应用,日本在粗糙集和概率论相结合方

面以及在医学的应用上比较突出,我国在知识约简、与信息论的结合、粗糙逻辑、粒计算、知识的不确定性研究方面取得了较大成功。特别是近十年来,由于粗糙集理论在机器学习与知识发现、数据挖掘、决策支持与分析等方面的广泛、成功的应用,成为当前计算机、人工智能、信息科学等领域的研究热点之一。

目前在理论上,主要研究粗糙集和其他软计算方法或人工智能方法的结合,例如模糊理论^[20]、神经网络^[21]、遗传算法^[22]等。针对经典粗糙集理论框架的局限性,拓宽粗糙集理论的框架,将建立在等价关系的经典粗糙集理论拓展到相似关系甚至一般关系上的粗糙集理论^[23~25]。根据属性的重要程度,提出了带隶属度及权重的粗糙集模型^[26],克服经典粗糙集分类过于严格、对噪音过于敏感、某些隐藏在边界中的规则丢失等缺陷。在应用上,粗糙集理论在许多领域得到了应用,如电力系统燃料的管理,味觉信号识别,区域水资源系统的评价,室内环境评价,城市岩土参数重要性评估,矿山不确定多属性问题的研究^[27~33]等。算法上,研究了粗糙集属性约简算法和规则提取启发式算法,如基于属性重要性和信息度量的启发式算法^[34]等。

粗糙集可以评价特定条件属性的重要性,进行属性约简,从决策表中去除冗余属性,从约简的决策表中产生决策规则,并利用规则对新对象进行决策。其传统建模过程包括对数据的预处理、连续属性的离散化、数据约简、发现依赖关系、规则生成和分类识别等多种方法。粗糙集在矿产预测领域的应用处于起步阶段,其在矿产预测领域中的应用研究本身就具有前沿性。鉴于粗糙集诸多特性和优势以及资源勘查领域自身的特点,其在矿产预测领域必定会发挥巨大的功效。基于粗糙集决策方法的矿山不确定多属性问题的研究^[33],及基于粗糙集理论的模糊综合评判权值确定^[35],为该方法在矿产资源预测中的应用提供了有利的参考。

1.5 技术路线

模型的建立基于数据,地矿数据的精确性和地矿因素的高度关联性是建立特征函数对矿产资源定量评价的基础,因此基于粗糙集的数据预处理是本研究的关键。基于粗糙集方法,从数量众多的变量中筛选特征变量,目的是要达到“变量结构最优化”,即要具有最优变量组合。这种筛选可以减少空间维数,简化系统,同时又不损失与研究对象有直接和间接联系的主要信息。粗糙集约简

后的数据参与定量建模，可以减小模型复杂度，提高模型的预测精度。本研究的基本流程如图 1-2 所示。

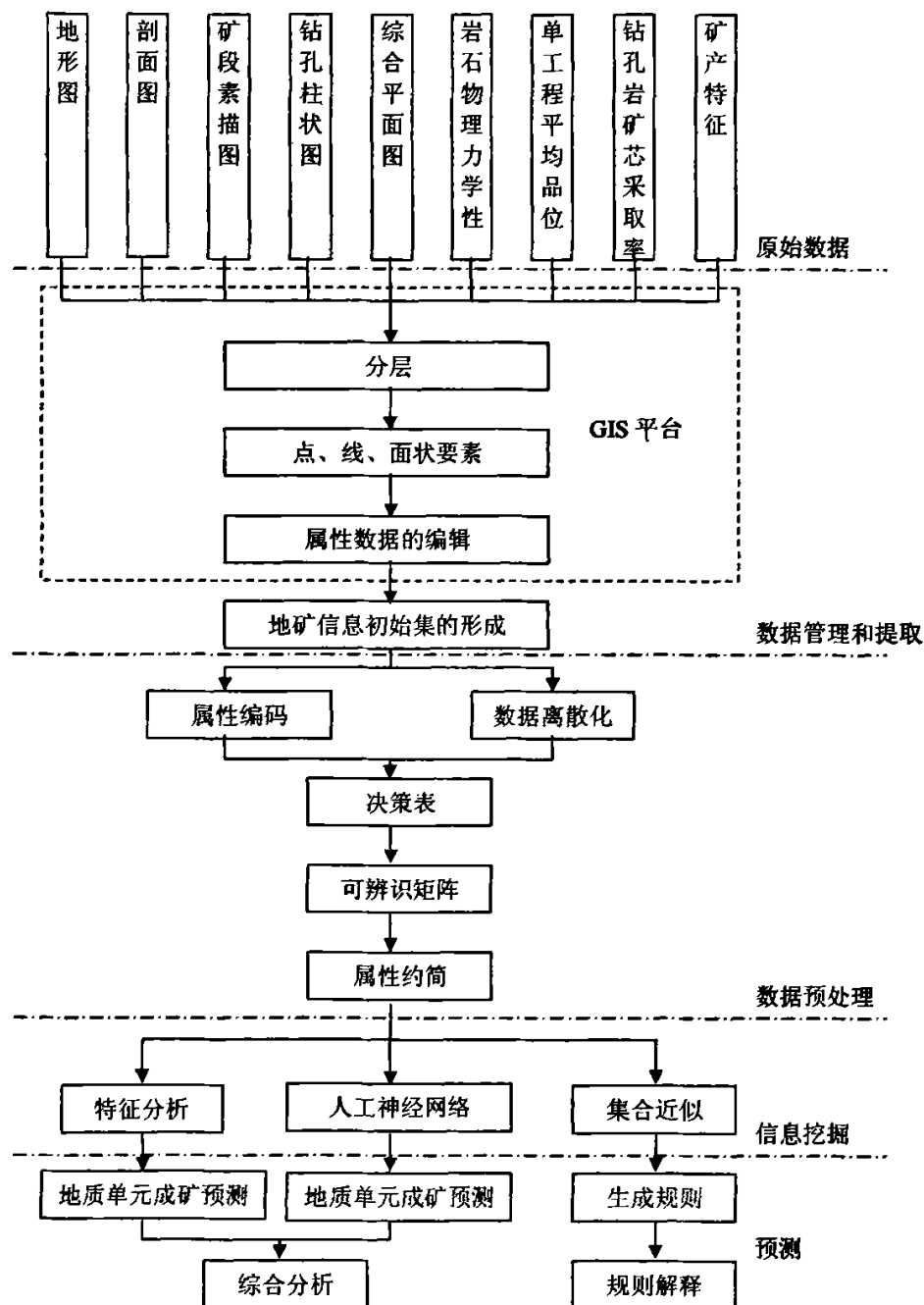


图 1-2 工作流程图

第2章 粗糙集理论

知识是实践经验的总结和提炼,来源于人类将对象进行分类的能力。粗糙集的主要思想就是在保证分类能力不变的前提下进行属性约简,提取决策或分类规则,将不精确或不确定的知识用已知的知识库来近似刻画^[36]。基于粗糙集模型,直接从原始决策表中求取近似集,并运用推理引擎,分别从下近似集中获取确定规则,从上近似集中获取可能规则,从海量的数据中挖掘更加概括、精炼的信息^[37]。矿产资源评价涉及到地质、物探、化探等众多因素的影响,人们在分析研究的过程中难以给出完整和确定的信息,除此之外,随着引入数据的增加,问题复杂性呈指数增加。一般一个地区成矿概率的大小与有利因素组合程度有关,也与关键因素是否存在相关。为了避免矿化信息的组合性爆炸,要求我们必须最大限度地查明“控矿变量组合”,提取、构置、优化各种成矿信息,并加以综合定量处理。矿产资源预测即是对研究区是否含矿产资源的判断,是一种分类知识的获取。根据已查明矿点提供的信息,基于粗糙集理论提取特征矿化信息,不可分辨关系和近似集思想从理论上回避了矿产资源评价中信息不完整和不确定性问题,此外完全针对数据的处理方式也使得处理结果更加客观真实。下面本章详细介绍粗糙集的相关概念。

2.1 信息系统

现实世界中的一个对象或个体通常使用属性-值的集合来表示,信息表这种数据表格是对客观对象的描述和罗列,表达的是说明性的知识。问题研究中将信息系统中的属性分为条件属性和决策属性两类。需要研究两类属性的关系,获取决策知识。本研究针对地质单元提取特征找矿信息作为科学研究和决策的依据。系统的形式为 (U, A, F, d) ,其中 U 为论域, U 中的元素 $x_i (i \leq n)$ 称为研究对象,一般记论域为 $U = \{x_1, x_2, \dots, x_n\}$,描述研究的全体对象;每个研究对象对应的属性集为 A ,记为 $A = \{a_1, a_2, \dots, a_m\}$, A 中的每个元素 $a_l (l \leq m)$ 描述一个属性; F 为 U 与 A 之间的关系集,即 $F = \{f_l: U \rightarrow V_l (l \leq m)\}$,其中 V_l 为 $a_l (l \leq m)$ 的值域; d 为决策信息, $U \rightarrow V_d$, V_d 取有限值。当信息表包含的数据足以反映论域的时候,通过属性所对应的等价关系就可以体现论域中的过程知识,即概念之间的逻辑关系或规则知识。

2.2 不可分辨关系

粗糙集理论中一个重要概念是不可分辨关系。信息系统记为 $S = (U, A, F, d)$, 属性集 A 的任意子集 B 定义论域 U 上的一个二元关系 R_B , $x_i, x_j \in R_B$ 。给定属性集 B , 如果对于任意 $a \in B$, $a(x_i) = a(x_j)$, 则 x_i 和 x_j 不可分辨, 其中 $a(x)$ 表示元素 x 对于属性 a 的属性值, 记为 $R_B = \{(x_i, x_j) | f_i(x_i) = f_i(x_j) (a_i \in B)\}$ 。显然 R_B 表示等价关系, R_B 所有等价类的集合记为 U/R_B 。不可分辨集被称为基本集, 如表 2-1 所示, 属性子集 $B=\{a_2\}$ 定义三个基本集 $\{x_1, x_2, x_5\}$, $\{x_3\}$ 和 $\{x_4, x_6, x_7\}$, 属性子集 $B=\{a_1, a_3\}$ 定义三个基本集 $\{x_1, x_2, x_3, x_4, x_6\}$, $\{x_5\}$ 和 $\{x_7\}$ 。如果 (x_i, x_j) 属于关系 R_B , 那么 x_i, x_j 记为 B -不可分辨, 关系 R_B 的等价关系定义为 B -基本集。令 $X \subseteq U$, 当 X 能用属性子集 B 确切描述, 即 X 是属性子集 B 所确定的 U 上的不分明集的并时, 称 X 是 B 可定义的, 否则称 X 是 B 不可定义的。 B 可定义集也称作 B 精确集, B 不可定义集称作 B 非精确集或 B 粗糙集。

表 2-1 实例数据集

U	a ₁	a ₂	a ₃	a ₄	d
x ₁	1	1	1	2	2
x ₂	1	1	1	1	2
x ₃	1	2	1	2	2
x ₄	1	3	1	1	1
x ₅	0	1	1	0	0
x ₆	1	3	1	2	1
x ₇	1	3	0	1	0
属性描述	a ₁ :地层	a ₂ :岩性	a ₃ :围岩蚀变	a ₄ :构造	d:金品位
	a ₁ (0): D ₁ ps ¹	a ₂ (1):石英砂岩	a ₃ (0):有	a ₄ (0):整合无断裂	d(0): 0
	a ₁ (1): O ₁ s	a ₂ (2):辉绿岩	a ₃ (1):没有	a ₄ (1):不整合无断裂	d(1):低
		a ₂ (3):其它		a ₄ (2):不整合断裂	d(2):高

2.3 近似集

近似集可以根据 B 的属性值来描述对象集 X 。令 $X \subseteq U$, $B \subseteq A$, 如果对象集 X 不能用 B -基本集的并确切地描述, 那么称为粗糙集, 否则称为精确集。例如, 给定属性子集 $B=\{a_3, a_4\}$, 对象集 $\{x_1, x_2, x_5, x_7\}$ 即为粗糙集, 因为根据条件属性子集 B , 对象 x_1 和 x_6 、 x_2 和 x_4 不可分辨, 不能根据条件属性 B 来对所有对

象是否属于集合 $\{x_1, x_2, x_5, x_7\}$ 作精确判定。对于所有 $X \subseteq U$ ，定义 X 的上近似集和下近似集来描述一个对象子集。下近似集是所有包含在对象集中的 B -基本集的并集，上近似集是所有与对象集交集非空的 B -基本集的并集，分别记为 $B_-(X)$ 和 $B^+(X)$ ，即

$$B_-(X) = \{x | (x \in U \wedge [x]_B \subseteq X)\} \quad (2-1)$$

$$B^+(X) = \{x | (x \in U \wedge [x]_B \cap X \neq \emptyset)\} \quad (2-2)$$

令 $B = \{a_3, a_4\}$ ， $X = \{x_1, x_2, x_5, x_7\}$ ，那么所有 U 上的 B -基本集为

$$U | IND(B) = \{\{x_1, x_3, x_6\}, \{x_2, x_4\}, \{x_5\}, \{x_7\}\}$$

令 $B_1 = \{x_1, x_3, x_6\}$ ， $B_2 = \{x_2, x_4\}$ ， $B_3 = \{x_5\}$ ， $B_4 = \{x_7\}$ ，那么 X 和 B -基本集的关系如下

$$X \cap B_1 = \{x_1\} \neq \emptyset$$

$$X \cap B_2 = \{x_2\} \neq \emptyset$$

$$X \cap B_3 = B_3 = \{x_5\} \neq \emptyset$$

$$X \cap B_4 = B_4 = \{x_7\} \neq \emptyset$$

得到下近似集和上近似集

$$B_-(X) = B_3 \cup B_4 = \{x_5, x_7\}$$

$$B^+(X) = B_1 \cup B_2 \cup B_3 \cup B_4 = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$$

2.4 正域、负域和边界域

一个集合的下近似和上近似，将论域划分为三个不相交的区域：正区域 $POS_R(X)$ ，负区域 $NEG_R(X)$ 和边界域 $BND_R(X)$ 。

X 的 B 正区域 $POS_R(X) = B_-(X)$ ，表示根据知识 B ， U 中所有一能归入集合 X 的元素构成的集合。

X 的 B 负区域 $NEG_R(X) = U - B^+(X)$ ，表示根据知识 B ， U 中所有不能确定一定归入集合 X 的元素的集合。

X 的 B 边界域 $BND_R(X) = B^+(X) - B_-(X)$ ，表示某种意义上论域的不确定域，边界域中的元素不能肯定的属于集合 X ，也不能肯定属于 \bar{X} 。

2.5 集合的精度和粗糙度

集合的不精确性是由于边界域的存在而引起的，集合的边界域越大，其精度越低。由于存在边界区域，即有些既不能在全域 U 的某个子集上被分类，也不能在它的补集被分类，而这些元素归于这种边界线区域，它的大小是衡量该

子集关于 U 上的等价关系 R 的近似精度, 为了更精确的表示这种近似精度的思想, 引入下面不精确性的数值量度。

设 $X \subseteq U$ 且 $X \neq \Phi$, 则称 $\alpha_R(X) = \text{card}(R(X)) / \text{card}(R^+(X))$ 为 $X \subseteq U$ 的近似精度, 其中 $\text{card}(S)$ 表示 S 的基数。 $\alpha_R(X)$ 表示我们获得关于集合 X 的知识是否完全的程度。

当 $\alpha_R(X)=1$ 时, $R(X) = R^+(X) \rightarrow \text{BND}(X)=\Phi$, 即全域 U 上的每个元素 x 都可以精确定义, X 是精确集;

当 $\alpha_R(X) < 1$ 时, 则 $R(X) \neq R^+(X) \rightarrow \text{BND}(X) \neq \Phi$, 因此 X 是 B 不可定义的;

同理, 也可以用其他量度来定义集合 X 的不精确程度, 引进粗糙度, 即 X 的 R 粗糙度为 $\rho_R(X)=1-\alpha_R(X)$ 。

X 的 R 粗糙度与精度恰恰相反, 它表示关于集合 X 知识 R 的不完备程度。

2.6 变精度粗糙集模型

在考虑的属性集范围内, 不同个体对象可能具有相同或者相似的描述, 在这些情况下, 由于一些属性值的缺失或者无法获得, 造成对象集合的描述不完全, 就导致了不完备信息系统的出现。在地质问题中, 由于数据具有多来源、多维数、多类别、多变量和多应用主题的“五多”特征^[38], 致使问题研究不可避免的具有不完整性 and 不确定性。为了提高对噪音数据的适应能力, Ziarko 提出了一种可变精度的粗糙集模型^[39]。

在可变精度粗糙集模型中, 定义了下面的条件概率:

$$P(D_j | [u_i]_{\text{IND}}) = \frac{P(D_j \cap [u_i]_{\text{IND}})}{P([u_i]_{\text{IND}})} = \frac{\text{card}(D_j \cap [u_i]_{\text{IND}})}{\text{card}([u_i]_{\text{IND}})} \quad (2-3)$$

给定一个决策表, 假定有条件属性集合 B 导出的等价类为 $\text{IND}(B) = \{B_1, B_2, \dots, B_k\}$, β 是依赖于数据中噪音程度的一个取值在 $[0, 0.5)$ 上的数, 则

$$(1) \beta \text{ 正域定义为 } \text{POS}_B(D_j) = \bigcup_{P(D_j | B_i) \geq 1-\beta} \{B_i \in B\} \quad (2-4)$$

$$(2) \beta \text{ 边界域定义为 } \text{BOS}_B(D_j) = \bigcup_{\beta < P(D_j | B_i) < 1-\beta} \{B_i \in B\} \quad (2-5)$$

$$(3) \beta \text{ 负域定义为 } \text{NEG}_B(D_j) = \bigcup_{P(D_j | B_i) \leq \beta} \{B_i \in B\} \quad (2-6)$$

条件属性集 B 和决策属性 D 之间的相关程度定义为

$$K_{\beta}(B,D) = \frac{\text{card}(\text{POS}_{\beta}(D) \cup \text{NEG}_{\beta}(D))}{\text{card}(U)} \quad (2-7)$$

$K_{\beta}(B,D)$ 是决策表中能够粗糙的或精确的划分到 β 正域和 β 负域的样本的百分比。

第3章 基于粗糙集的约简思想

基于粗糙集理论的知识获取,是在保持决策表决策属性和条件属性的依赖关系不发生变化的前提下对决策表进行约简,包括属性约简和属性值约简。在本研究中,属性约简主要基于条件属性子集的不可分辨关系,将多源化、高维数的地质信息进行浓缩,提取主要的特征矿化属性,从而得到约简后的条件属性对于决策属性的决策规则;属性值约简主要针对条件属性值对决策属性值所起的作用,删除特征矿化属性中的冗余值,从而得到最简的决策规则。基于粗糙集的属性约简可以从代数集合观点和信息论的信息熵观点进行系统分析,该方法已经得到了大量的研究,属性约简效果也得到很好的验证^[27-33]。属性约简的基本思想为一般约简算法、基于可辨识矩阵的约简算法和启发式算法;属性值约简的基本思想有归纳值约简算法,基于决策矩阵约简算法等。基于粗糙集属性约简的算法相对成熟,可以适用于本研究中特征矿化信息的提取。

3.1 属性约简思想

3.1.1 一般约简算法

在粗糙及约简算法中,最直观的就是删除法^[40-44]。这个方法依次从数据表中删除属性,将删除属性后的数据表和原数据表的决策类的等价关系进行比较,如果等价关系没有变化,那么就可以继续从新生成的数据表中删除属性,继续比较;如果等价关系发生了变化,那么就恢复为前一个数据表,删除另外一个属性。直到所有的属性都不能删除。

算法描述:

输入:信息系统 $S = (U, A, F, d)$, 其中 U 为论域, A 为属性集, $A = C \cup D$, C 为条件属性集合, D 为决策属性集合;

输出: 约简集 red

步骤:

(1)初始化: $red = C$;

(2)令 $temp = red$;

(3)如果 $temp \neq \Phi$, 循环: 取 $a \in C$, 判断: 如果 $POS_{red-\{a\}}(D) = POS_{red}(D)$, 则: $red = red - \{a\}$, $temp = red$; 否则 $temp = temp - \{a\}$

(4)输出 red

3.1.2 基于可辨识矩阵的约简算法

可辨识矩阵是由斯科龙(Skowron)教授提出的,是近年来在粗糙集约简上出现的一个有力工具。利用这个工具,可以将存在于复杂的信息系统中的全部不可区分关系表达出来。

令决策表系统为 $S = (U, A, F, d)$, $A = C \cup D$ 是属性集合,子集 $C = \{a_i | i = 1, \dots, m\}$ 和 $D = \{d\}$ 分别为条件属性集和决策属性集, $U = \{x_1, x_2, \dots, x_n\}$ 是论域, $a_i(x_j)$ 是样本 x_j 在属性 a_i 上的取值。 $C_D(i, j)$ 表示可辨识矩阵中第 i 行 j 列的元素,则可辨识矩阵 $C_D(i, j)$ 定义为

$$C_D(i, j) = \begin{cases} \{a_k | a_k \in A \wedge a_k(x_i) \neq a_k(x_j)\}, & d(x_i) \neq d(x_j) \\ 0, & d(x_i) = d(x_j) \end{cases} \quad (3-1)$$

其中 $i, j = 1, \dots, n$ 。

由上面区分矩阵的定义可知,当两个对象的决策属性取值相同时,它们所对应的区分矩阵元素的取值为 0;当两个对象的决策属性不同且可以通过某些条件属性的取值不同加以区分时,它们所对应的区分矩阵的元素的取值为这两个对象属性值不同的条件属性集合,即可以区分这两个对象的条件属性集合。当两个对象发生冲突时,即所有的条件属性取值相同而决策属性的取值不同时,则它们所对应的区分矩阵中的元素的取值为空集。

根据区分矩阵的定义,可以得到区分矩阵如下的性质:

(1)区分矩阵是一个对称矩阵;

(2)如果矩阵中存在一个元素,其取值只有一个属性,则表明该属性是区分这个矩阵元素所对应的两个样本所必需的属性,也是唯一能够区分这两个样本的属性。可辨识矩阵中的这些属性就是该决策表系统的属性核,即可辨识矩阵中凡是条件属性组合中包含有核属性的矩阵元素都可以仅用核属性就把决策不同的记录区分开来,也就是说属性组合中凡是包含有核属性的可辨识矩阵项的其它条件属性都是多余的;对于不包含核属性的属性组合必然每个组合都至少有一个元素成为约简后的一个条件属性,否则决策表中的某些记录将无法识别。

算法描述:

输入: 信息系统 $S = (U, A, F, d)$, 其中 U 为论域, A 为属性集, $A = C \cup D$, C 为条件属性集合, D 为决策属性集合;

输出: 约简集 red

步骤:

(1) 计算决策表的可辨识矩阵 $C_D(i, j)$;

(2) 将可辨识矩阵中包含核属性的元素值修改为 0;

(3) 对于可辨识矩阵中的所有取值为非空集合的元素 C_{ij} ($C_{ij} \neq 0, C_{ij} \neq \emptyset$), 建立相应的析取逻辑表达式 L_{ij} ,

$$L_{ij} = \bigvee_{a_i \in C_{ij}} a_i \quad (3-2)$$

(4) 将所有的析取逻辑表达式 L_{ij} 进行合取运算, 得一个合取范式 L , 即

$$L = \bigwedge_{C_{ij} \neq 0, C_{ij} \neq \emptyset} L_{ij} \quad (3-3)$$

(5) 将合取范式 L 转换为析取范式的形式, 得

$$L' = \bigvee_i L_i \quad (3-4)$$

(6) 输出约简结果。析取范式中的每个合取项就对应一个属性约简的结果, 然后将所有核属性加入析取范式中的每个合取项, 每个合取项中所包含的属性组成约简后的条件属性集合。

3.1.3 基于互信息的启发式算法

在求取决策表属性约简的时候, 可以利用决策表条件属性和决策属性之间的互信息。在决策表中增加某个属性所引起的互信息的变化的大小可以作为该属性重要性的度量。

信息系统 $S = (U, A, F, d)$, $A = C \cup D$ 。有条件属性 C 和决策属性 D 在论域 U 上的划分分别为 U/C 和 U/D :

$$U/C = \{C_1, C_2, \dots, C_n\}$$

$$U/D = \{D_1, D_2, \dots, D_m\}$$

U 上的任一划分都可看作是定义在 U 的幂集中的随机变量, 其概率分布为

$$[U/C, p] = \begin{bmatrix} C_1 & C_2 & \dots & C_n \\ p(C_1) & p(C_2) & \dots & p(C_n) \end{bmatrix} \quad (3-5)$$

$$[U/D, p] = \begin{bmatrix} D_1 & D_2 & \dots & D_m \\ p(D_1) & p(D_2) & \dots & p(D_m) \end{bmatrix} \quad (3-6)$$

其中 $p(C_i) = |C_i|/|U|, i = 1, 2, \dots, n$; $p(D_j) = |D_j|/|U|, j = 1, 2, \dots, m$ 。

知识 C 的熵 $H(C)$ 定义为

$$H(C) = - \sum p(C_i) \log_2 p(C_i) \quad (3-7)$$

知识 D 相对于知识 C 的条件熵 $H(D|C)$ 定义为

$$H(D|C) = - \sum p(C_i) \sum p(D_j|C_i) \log_2 p(D_j|C_i) \quad (3-8)$$

其中 $p(D_j|C_i) = |D_j \cap C_i| / |C_i|$ 。 $H(D|C)$ 反映了知识 D 相对于知识 C 的依赖程度，C 与 D 的互信息为

$$I(C;D) = H(D) - H(D|C)。 \quad (3-9)$$

设信息系统 $S = (U, A, F, d)$ ， $B \subseteq A$ ，那么在 B 中添加一个属性 $a \in A$ 后互信息的增量为：

$$SGF(a, B, D) = I(B \cup \{a\}; D) - I(B; D) = H(D|B) - H(D|B \cup \{a\}) \quad (3-10)$$

该信息越大，说明在已知属性 B 的条件下，属性 a 对决策 D 就越重要。基于互信息的属性约简算法就是将这个增量 $SGF(a, B, D)$ 作为属性重要性的启发信息，一次选择最重要的属性加入核中，直到满足终止条件，便得到信息系统或决策表的一个约简。

算法描述：

输入：信息系统 $S = (U, A, F, d)$ ，其中 U 为论域， A 为属性集， $A = C \cup D$ ， C 为条件属性集合， D 为决策属性集合，相对核 $core_D(C)$ ；

输出：约简集 red

步骤：

(1) $red_D(C) = core_D(C)$ ；

(2) $C' = C - red_D(C)$ ；

(3) 在 C' 中找到使得 $SGF(a, red_D(C), D)$ 取最大值的属性 a ，如果使 $SGF(a, red_D(C), D)$ 取最大值的属性多于一个，则从中选取一个与 $red_D(C)$ 的值的组合数量小的属性作为 a ；

(4) $red_D(C) = red_D(C) \cup \{a\}$ ， $C' = C' - \{a\}$ ；

(5) 若 $I(red_D(C), D) = I(core_D(C), D)$ ，则终止，否则转步骤(3)。

3.2 属性值约简

通过属性约简，可以将决策表中对决策分类不必要的属性省略，从而实现决策表的简化，这有利于从决策表中分析发现对决策分类起作用的属性。但是，属性约简只是在一定程度上去掉了决策表中的冗余属性，没有充分取掉决策表中的冗余信息。这就需要对决策表属性值进行约简，提取满意的决策规则。

决策矩阵的值约简算法是 Ziarko 等人针对可变精度粗糙集模型所提出的，用于获取具有最大适应度的决策规则，并成功应用于一个水资源调度系统的设计中^[45]。

该方法的主要思想是：以约简后的决策表为处理对象，针对决策属性 $d \in D$ 及其特定值 V_d ，关注满足 $d(x) = V_d$ 的对象 x 的集合 $\{V_d(x)\}$ 。用矩阵形式可以将区分所有属于集合 $\{V_d(x)\}$ 对象和属于 $\{U - \{V_d(x)\}\}$ 集合对象的属性值对表示出来。

用 x_i 代表任何一个属于集合 $\{V_d(x)\}$ 的对象， $i = 1, 2, \dots, m$ ， $\text{card}(\{V_d(x)\}) = m$ ； x_j 代表任何一个属于集合 $\{U - \{V_d(x)\}\}$ 的对象， $j = 1, 2, \dots, n$ ， $\text{card}(\{U - \{V_d(x)\}\}) = n$ 。决策矩阵为

$$M_{ef} = \{(a, a(x_i)); a(x_i) \neq a(x_j)\} \quad (3-11)$$

将 M_{ef} 的各个元素作为一个布尔表达式，决策规则集合可以表达为如下形式的布尔函数： $B_e = \bigwedge_f (\vee M_{ef})$ 。可以看出布尔函数 B_e 的基本蕴含实际上是当前决策类的所有最大泛化规则，再针对每一个最大泛化规则，以属性条件的并作为产生式规则条件，以决策属性的值作为结论构成产生式规则。

3.3 小结

一般约简算法采用搜索策略获得所有可能的属性约简，是一个组合爆炸问题，穷尽的搜索所需要的时间和空间代价都很高，而且每个约简都必须经过记录之间的反复比较运算，复杂度非常高。基于区分矩阵的约简将对属性组合情况的搜索演变成为逻辑公式的化简（合取式转换为析取式），从而简化问题，虽然利用区分矩阵的对称性质可以使计算减半，但是将区分函数转换成析取式，也是一个 NP 问题。互信息法从信息熵的角度考察属性约简，用添加某个属性引起的互信息变化的大小来反映该属性的重要程度，将重要性度量引入算法作为启发信息，可以减少搜索空间，提高算法效率。

本研究中样本数据较少，可以应用可辨识矩阵直接获得决策表属性核，进而计算属性约简集。基于可辨识矩阵和逻辑运算的属性约简算法可以得到决策表的所有可能的属性约简结果，这些结果不一定是最小约简集，而基于互信息法可以直接获得长度最短约简集。针对互信息法获得的最小约简集，进而应用决策矩阵进行属性值约简，获得最简规则。

第4章 基于粗糙集的特征矿化信息提取

目前已有的单一确定型或随机型数学模型,在表达复杂的地质历史过程中,由于条件的限制,都存在各自的局限,这也是数学地质目前尚不能很好解决地质学定量研究的根本症结。为了有效地提取成矿预测综合信息,有必要客观地对诸多原始观测信息进行筛选,突出成矿密切相关的致矿因子,得到最佳变量组合。本章基于 GIS 平台集成管理多源地学数据,考虑成矿背景、数据获取环境、数据质量等因素,提取地质单元地矿信息,构建决策表,基于粗糙集方法对属性进行约简,应用可变精度粗糙集模型从约简的属性表中求取近似集,从下近似集中获取确定规则,并运用决策矩阵法对属性值进行约简,剔除冗余信息,约简结果与研究区勘探资料完全相符。

4.1 研究区地质

研究区位于华南褶皱系滇东南褶皱带文山-富宁断褶束西畴拱凹北缘,文山-那洒弧形构造东段。地壳活动经历了由地槽(加里东期)到地台(华力西期),再到地槽(印支期)的复杂演化过程。矿区出露地层主要为上寒武统博菜田组(ϵ_{3b})中厚层一块状粉晶灰质白云岩;下奥陶统独树棵组(O_{1d})厚层、块状硅化石英砂岩、石英砂岩,闪片山组(O_{1s})中厚层状白云岩、白云质灰岩、生物碎屑灰岩,老寨组(O_{1l})薄—中厚层状石英砂岩,夹少量薄层泥岩;下泥盆统坡松冲组(D_{1ps})薄—中厚层状石英砂岩、硅化石英砂岩,上部夹少量薄层粉砂岩、粉砂质泥岩,坡脚组(D_{1p})薄层泥岩夹粉砂质泥岩,古木组(D_2)厚层一块状白云岩夹生物碎屑灰岩及灰岩,东岗岭组层孔虫灰岩及生物碎屑礁灰岩。

矿区处于那洒短轴背斜北翼,受区域性多期应力影响,测区范围内构造也表现出多期性特征。断层主要发育近南北向、近东西向和北西向三组;区内围岩蚀变主要有硅化、褐铁矿化、黄铁矿化、辉锑矿化、碳酸盐化、粘土化、绢云母化等。区域矿产丰富,现已发现的矿(床)点有九克、皂角树、革夺、韭菜坪、俄里、田坊等(锑)金矿床(点),本研究区为老寨湾金矿点。

4.2 研究区勘探情况

研究区勘探工作共施工了 40 个槽探,5 个剥土,工作量共 20083 立方米;7 个浅井,工作量共 60.8 米;4 个坑道;52 个钻孔,完成工作量 4079.08 米。研

究区内构造、岩性简单,地质观察点加上众多的槽、井、坑、钻等工程,基本查明了矿区内的地层层序,含矿层位,岩浆岩的种类,形态特征,主要构造特征,矿化带和蚀变范围。研究区勘探工作为本研究提供了丰富的地质数据,保证了实验和分析的顺利进行。

4.3 基于 GIS 的数据处理

4.3.1 GIS 平台及其应用

ArcGIS 是 ESRI 公司的代表产品,它提供了一体化的完整的地图绘制、显示、编辑和输出的集成环境,具有人性化设计和所见即所得的界面,支持各种复杂、动态的表达及强大的空间分析能力。这些都为地质数据的显示、表达、查询等提供了有力的平台。

GIS 技术的优势是提供集成管理多源地学数据、方便地建立模型及进行模拟的能力,将搜集到的各种图形信息(图件)、文字描述信息、数字数据通过合理、有效的空间数据库进行管理可大大提高地质数据分析的效率。合理、有效的空间数据库建立起来以后,可实现灵活的图形信息与属性的双向查询检索,既可从图形检索其各种属性,也可根据地质体的各种专题属性检索相应的图形,还可以根据多种属性进行组合条件检索。GIS 提供的空间分析(叠加、包含、相邻关系、缓冲区、地形分析)及空间信息计算(面积、周长、距离等)等功能,实现了传统方法难以解决的对各种地质体的多种空间关系的定量分析,这种空间分析对于研究地质现象之间的制约关系与相互作用进而提取与矿床或矿化有关的地质标志较为有效。

4.3.2 地质数据的管理

本研究中对地质数据的管理主要包括如下两个方面的工作:

(1) 依据地质实体的空间信息的分层管理。目前,空间数据的组织形式一般为图层。多个具有某些相同或相似特性的空间对象的集合,在数据库中以表的形式进行组织和表达,一个普通表图层一般由空间数据表、属性信息表和空间索引表组成。为了提高地图中各个要素的检索速度,便于数据的灵活调用、更新及管理,在空间数据库中,往往将不同类不同级的图元要素进行分层存放,每一层存放一种专题或一类信息^[46]。按照一定的需要或标准把某些相关图元要素组合在一起成为图层,它表示地理特征以及描述这些特征的属性的逻辑意义

上的集合。根据空间数据的现有数据源和实际使用情况，本研究区的要素分层结构如图 4-1 所示。

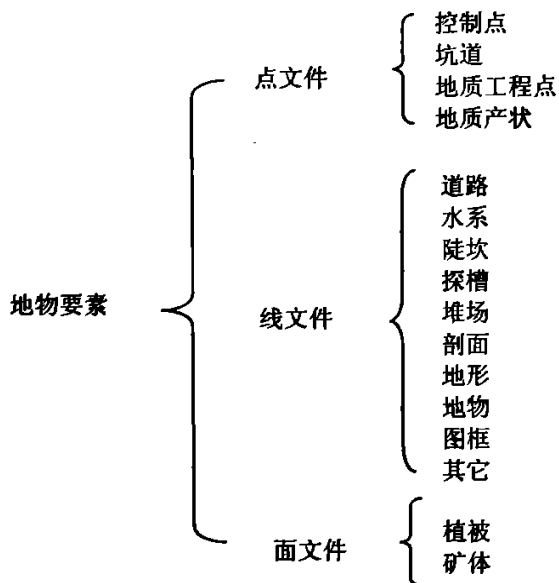


图 4-1 云南某地区图形要素分层结构图

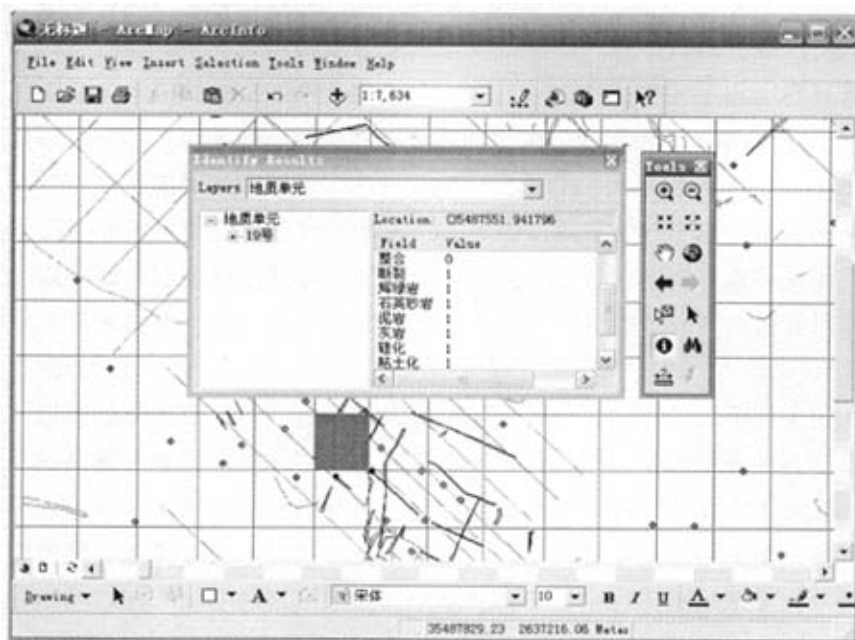


图 4-2 成矿信息的查询

(2)地质综合信息的分析。如 GIS 的叠加功能可形象地理解为计算机化的透图台,是资源评价用得最多的空间分析功能,可在地质图空间数据库中根据地层、岩性、构造等属性检索出相应的地质体;还可根据地质现象的点(矿床、钻孔、采样点等)、线(断层、线性构造、河流等)以及面(地层、岩体、异常区等)检索它们的属性。不同图形信息的叠置既可用于地质成矿信息的提取,也可用于多种成矿信息的综合分析^[47],如图 4-2 所示。计算机化的透图台还可以通过综合分析的方法反映信息之间的关系。

4.3.3 变量的选取

在地学研究中,广义的地质变量是根据表达形式和数学性质分成的定性变量、定量变量^[48]。定性变量一般都是离散型变量,主要通过“鉴定”区分不同的对象或个体,如岩性,不同类型的断层和褶皱等,可用 0、1 等符号表示取值,当描述单元中与成矿作用有关联的地质构造时,可用 1 表示存在,0 表示不存在。定量变量则主要为连续型变量,也可为离散型变量,这种数据彼此间不仅能比较其大小,而且可以定量地表示这种差异,如品位值等。

选择变量应以地质研究为基础,要注意到地质背景、控矿条件、成矿规律这一基本前提。变量的选择应该依据以下原则:

(1) 由于对矿床的成矿理论主观认识的不同,以及地质工作程度和研究程度的限制,不可能完全了解与矿有关的地质因素和标志,因此,在开始取地质变量时,应尽可能多取,以免漏掉有用的信息,然后用数学方法进行挑选;

(2) 变量选择应注意其纵向和横向的代表性,尽可能保证尺度的一致性;

(3) 对于数学方法选用而地质意义不明确的变量,应进一步分析其地质意义,以挖掘其隐蔽地质信息;对于地质意义明确且与研究对象有关的变量,用数学方法未被选上时,应对地质变量的取值和变换进行研究,使其尽量能被数学方法选上。

根据成矿地质背景、成矿模式和成矿规律的分析对比,可以看出本研究区金矿的形成受地层层位和断裂、褶皱构造以及岩性、围岩蚀变的控制。含矿地层层位主要是下泥盆统坡松冲组,岩石类型主要为石英砂岩、灰岩等,研究区大型褶皱和断裂都表现出复杂构造特征,同时岩浆活动对矿床的形成也都有不同程度的影响,蚀变也是主要的矿化地址特征之一。因此,本研究从地层、构造、岩浆岩三个方面选取地质变量,其描述见表 4-1。

表 4-1 地质特征属性表

地质特征	地质特征描述
地层	下泥盆统坡松冲组（不整合，一段 D_{1ps}^1 、二段 D_{1ps}^2 ）、下奥陶统闪片山组(O_{1s})
断层	近南北向断层，近东西向断层，北西向断层
褶皱	背斜
岩性	辉绿岩、石英砂岩、粉砂质泥岩、石英砂质粘土、灰岩
围岩蚀变	硅化、褐铁矿化、黄铁矿化、辉锑矿化、碳酸盐化、粘土化、绢云母化等

4.3.4 模型单元的选取

在矿产资源评价中，研究的对象是特殊的地质体，须用样本的观测结果来描述总体特征和确定远景区，因此，首要条件应当保证抽样的随机性和样品的代表性。为此，对研究区进行网格单元^[49]划分，网格化单元是独立的，单元中已知矿床是随机分布的，有利于随机抽样，形成简单样本。同时还可研究矿床的分布规律。



图 4-3 老寨湾矿区网格划分单元示意图

在单元中选取变量,如果单元面积过小,会产生变量不足或变量大量雷同的现象,影响预测结果的可靠性。为了选取足够的变量,本研究以经纬线网格为划分标准,将研究区划分为54个等面积正方形区域,在单元中分布有已知矿床的列为已知单元,剩余的单元,只有变量同已知单元有可比性,才能作为预测单元。

老寨湾成矿地段地质勘探程度较高,已经发现一些矿床和矿点,为了进一步挖掘找矿潜力,首先对该矿区进行规则网格划分单元,编号单元为已勘探单元,如图4-3所示。其中勘探程度高的3号、4号、6号、7号、8号、11号、12号、13号、18号、19号、20号、22号、23号、24号、28号、30号、31号和35号单元为模型单元,建模后预测剩余单元情况。

4.4 地质变量的取值和变换

所谓取值是指获取某个地质特征的具体数值。取值方法很多,如计数、分级可获得有序性变量;鉴别可标度名义型数据。通过这些途径所获得的数据统称为地质变量的原始观测值,然后根据所应用的数学模型以决定其是否需要变换。

对研究区勘查的数据进行整理,分析研究区观测数据和各大比例尺地质图和截面图,提取该研究区关于地层、断裂、岩性、围岩蚀变等地质特征。决策表的构建包括名义型数据的编码和多源勘查数据的离散化。在矿产信息中,名义型数据即一些描述性的属性值,如断层、岩性等,本研究中采用“二态数据”,即“0-1”数据进行基于粗糙集数学分析的探讨,描述地质单元某种属性的有或无;对于如金品位等类属性,需要进行离散化处理。

4.4.1 数据量纲的统一

不同的数学模型对地质变量的要求不同,如判别分析要求变量呈正态分布,回归分析要求因变量呈正态分布,聚类分析要求各变量量纲一致等,因此,地质变量的变换一定要根据数学模型要求,有的放矢地去进行。为了使数据量纲一致,可对原始数据进行标准化、极差化或均值化变换。本研究为了对比分析,应用极差变换将数据归一化处理,同时可以反映地质单元的成矿概率。

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, i = 1, 2, \dots, n \quad (4-1)$$

式(4-1)中 x_i 为原始数据, x_{\min} 为变量的最小值, x_{\max} 为变量的最大值,变换

后的数据有统一量纲，其最大值为 1，最小值为 0，所有数据变化在 0-1 之间。变换后变量间相关程度不变，其几何意义相当于把坐标原点移至变量最小值的位置。

4.4.2 基于粗糙集的数据离散化处理

运用粗糙集理论处理决策表时，要求决策表中的值用离散数据表达。如果某些条件属性或决策属性的值域为连续值，则在处理前必须进行离散化处理，而且，即使对于离散数据，有时也需要通过将离散值进行合并得到更高抽象层次的离散值。

表 4-2 决策表

样本号	单元号	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	d
1	3	1	1	0	1	1	0	1	1	0	1
2	4	1	1	0	1	1	0	1	1	0	1
3	6	0	1	0	1	0	1	1	1	0	2
4	7	0	0	0	1	1	0	0	1	0	0
5	8	1	0	0	1	0	0	1	1	0	2
6	11	0	0	0	1	1	1	1	0	0	0
7	12	1	1	0	1	0	1	0	0	1	2
8	13	1	1	0	1	1	0	1	0	1	2
9	18	1	0	0	1	0	0	0	0	0	0
10	19	1	1	1	1	1	1	1	1	1	2
11	20	1	0	0	1	1	0	1	0	1	2
12	22	1	0	0	1	1	1	0	0	0	1
13	23	1	0	0	1	1	1	0	0	1	2
14	24	1	0	1	1	0	0	1	0	0	2
15	28	1	0	0	1	1	0	1	1	1	2
16	30	1	0	0	1	1	1	0	0	0	0
17	31	1	0	0	1	1	1	0	0	0	1
18	35	1	0	0	1	0	1	0	0	0	0

离散化本质上可归结为利用选区的断点来对条件属性构成的空间进行划分，假设某个属性有 m 个属性值，则在此属性上就有 $m-1$ 个断点可取。选择断

点的过程也是合并属性值的过程，通过合并属性值，减少属性值的个数，减小问题的复杂度，有利于提高知识获取过程中所得到的规则知识的适应度。

本研究中将地质单元勘探的金品位离散化处理后，分为高、低和零三类，分别用 2, 1 和 0 表示。根据研究区经济技术条件和金矿成因类型，同时参照行业标准，断点选取为 0, 0.5 和 1.0。如果金品位值在区间 $[0, 0.5)$ 内，决策值为 0；如果金品位值在区间 $[0.5, 1.0]$ 内，决策值为 1；在采样点中，金品位最大值为 1.84g/t，如果金品位值在区间 $(1.0, 1.84]$ 内，决策值为 2。根据上述分析，构建表 4-2 的决策表，其中， a_1 ：不整合； a_2 ：断裂； a_3 ：辉绿岩； a_4 ：石英砂岩； a_5 ：泥岩； a_6 ：灰岩； a_7 ：硅化； a_8 ：粘土化； a_9 ：褐铁矿化。

4.5 基于粗糙集的规则获取模式

目前，基于粗糙集的规则获取主要有两种模式^[50]。一种是通过寻找属性核并去掉多余的属性求出约简的决策表，并从最简的决策表中获取相应的确定性规则；另一种的主要思想是直接从原始决策表中求取近似集，并运用推理引擎，分别从下近似集中获取确定规则，从上近似集中获取可能规则，如图 4-4 所示。

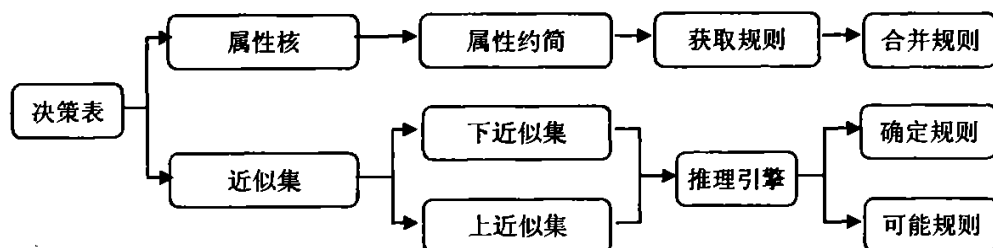


图 4-4 基于粗糙集规则推理模式

应用粗糙集原理所获得的规则越简洁，规则集的规模越小，就越具有较好的可理解性和较强的泛化能力。在信息决策中，我们大量面对的是不协调决策信息系统，即数据挖掘是在非协调数据上进行的。例如，两个地质工程点所获取的属性信息完全相同，但由于不同的概念分成不同的类，即决策属性值不同，这种情况就称为非协调，它表明在数据库中一些属性丢失^[51]。因此针对这种信息系统研究决策问题，更具有实践意义。粗糙集即处理这类非协调决策信息系统的有力工具，它不会将这些不协调的数据从数据库中移除，而是通过上近似和下近似的集合来重新描述概念，使问题描述更加精确。

4.6 特征矿化信息的分析

4.6.1 特征矿化信息的提取

构建好决策表后，利用粗糙集对属性值进行约简，找出各矿化信息与成矿的关联度。基于粗糙集方法，从数量众多的初始变量中筛选最重要的变量，目的是要达到“变量结构最优化”，即要具有最优变量组合。这种筛选可以减少空间维数，简化系统，同时又不损失与研究对象有直接和间接联系的主要信息。

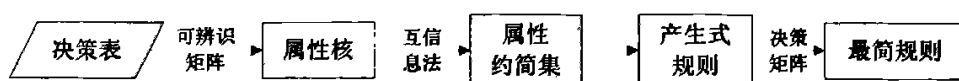


图 4-5 最简规则提取流程

基于粗糙集约简思想，本研究应用辨识矩阵提取决策表的相对属性核，然后应用互信息法找到属性约简集，研究流程如图 4-5 所示。将决策表中重复的属性删除，得到 15 个样本信息，重新赋样本号，其中 1-4 号决策属性为 0，5-6 号决策属性为 1，7-15 号决策属性为 2，对于决策属性值不同的对象提取其条件属性值不同的条件属性集合，获得如表 4-3 所示的辨识矩阵。

表 4-3 可辨识矩阵

样本	5	6	7	8	9	10	11	12	13	14	15
1	$a_1a_2a_7$	a_1a_6 a_7a_8	a_2a_5 a_6a_7	a_1a_5 a_7	$a_1a_2a_5$ $a_6a_8a_9$	$a_1a_2a_7$ a_8a_9	$a_1a_2a_3$ $a_6a_7a_9$	a_1a_7 a_8a_9	a_1a_6 a_8a_9	$a_1a_3a_5$ a_7a_8	$a_1a_7a_9$
2	a_1a_2 a_6a_8	a_1a_7	$a_2a_5a_8$	a_1a_5 a_6a_9	$a_1a_2a_5$ a_7a_9	a_1a_2 a_6a_9	$a_1a_2a_3$ a_8a_9	a_1a_6 a_9	$a_1a_7a_9$	a_1a_3 a_5a_6	a_1a_6 a_8a_9
3	a_2a_5 a_7a_8	a_5a_6	$a_1a_2a_5$ a_7a_8	a_7a_8	$a_2a_6a_9$	a_2a_5 a_7a_9	$a_2a_3a_5a_6$ $a_7a_8a_9$	a_5a_7 a_9	$a_5a_6a_9$	a_5a_7	a_5a_7 a_8a_9
4	$a_2a_5a_6$ a_7a_8	a_5	a_1a_2 a_7a_8	a_6a_7 a_8	a_2a_9	$a_2a_5a_6$ a_7a_9	$a_2a_3a_5$ $a_7a_8a_9$	a_5a_6 a_7a_9	a_5a_9	$a_5a_6a_7$	$a_5a_6a_7$ a_8a_9
5			$a_1a_5a_6$	a_2a_5	$a_5a_6a_7$ a_8a_9	a_8a_9	$a_5a_6a_9$	a_2a_6 a_9	$a_2a_6a_7$ a_8a_9	a_2a_3 a_5a_6	a_2a_9
6			$a_1a_2a_5$ a_7a_8	a_5a_6 a_7a_8	$a_2a_5a_9$	a_2a_6 a_7a_9	$a_2a_3a_7$ a_8a_9	a_6a_7 a_9	a_9	a_3a_5 a_6a_7	a_6a_7 a_8a_9

从可辨识矩阵可以看出，属性核为 $\{a_5\}$ 和 $\{a_9\}$ 。表 4-3 中

$$L_{1,5}=a_1 \vee a_2 \vee a_7$$

$$L_{1,9}=a_1 \vee a_6 \vee a_7 \vee a_8$$

.....

$$L_{6,15} = a_6 \vee a_7 \vee a_8 \vee a_9$$

将这些表达式进行合取得到合取表达式 L,

$$L = L_{1,5} \wedge L_{1,6} \wedge \dots \wedge L_{6,15}$$

对 L 进行转换, 最终得到析取范式 L',

$$L' = (a_1 \wedge a_5 \wedge a_7 \wedge a_9) \vee (a_5 \wedge a_6 \wedge a_7 \wedge a_9) \vee (a_5 \wedge a_7 \wedge a_8 \wedge a_9) \vee (a_1 \wedge a_3 \wedge a_5 \wedge a_8 \wedge a_9)$$

可以看出, 基于可辨识矩阵的属性约简结果不是唯一的, 而且如果样本数量较大, 析取表达式的转换将相当困难。从信息熵的角度考察属性约简, 对属性进行重要性度量并作为算法的启发信息, 可以避免逻辑运算中的组合爆炸问题, 找到最优属性约简集。

根据式子(3-7)、(3-8)和(3-9), 计算原决策表条件属性集 A 和决策属性集 D 之间的互信息为 $I(A;D) = 1.3383$, 而相对属性核集 C 和决策属性集 D 之间的互信息为 $I(C;D) = 0.7479$ 。由于 $I(C;D) < I(A;D)$, 说明属性核集 C 的分类能力低于属性集 A 的分类能力, 所以属性核集 $\{a_5, a_9\}$ 不是决策信息系统的约简集。

对于每个属性 $a \in A - C$, 应用式子(3-10)计算属性重要度, 即互信息的增量, 得到表 4-4。

表 4-4 条件属性重要度

属性	a_1	a_2	a_3	a_4	a_6	a_7	a_8
SGF(a,C,D)	0.3238	0.1400	0.0570	0	0.0066	0.3238	0.1400

选取互信息增量最大的属性 a_1 和 a_7 , 与相对属性核组成新的约简集

$$R = C \cup \{a_1, a_7\}$$

此时, 计算条件属性集 R 和决策属性集 D 的互信息

$$I(R;D) = 1.3383$$

即 $I(R;D) = I(A;D)$, 满足算法终止条件, 说明属性约简集 R 与条件属性集 A 的分类能力相同。应用互信息法对信息熵进行计算, 可以直接获得属性最少的约简集 $\{a_1, a_5, a_7, a_9\}$, 即不整合、泥岩、硅化和褐铁矿化。

信息系统决策表仅仅包含了全域中部分例子集的信息, 而且所选择的属性集合对于表征这个样例本身不一定是充分的。因此这个问题中有很多不确定因素, 可采用可变精度粗糙集模型进行分析。

根据本研究对决策属性的分类, 可以针对中间决策值进行分析。假定一个

概念 $Y = \{O_1, O_2, O_{12}, O_{17}\}$, 则

令 $\beta = 1$:

概念 Y 的 β 正域为 $POS_R^\beta(Y) = \{o_1, o_2\}$

概念 Y 的 β 负域为 $NEG_R^\beta(Y) = \{o_3, o_4, o_5, o_6, o_7, o_8, o_9, o_{10}, o_{11}, o_{13}, o_{14}, o_{15}, o_{18}\}$

概念 Y 的 β 边界域为 $BND_R^\beta(Y) = \{o_{12}, o_{16}, o_{17}\}$;

令 $\beta = 0.6$:

概念 Y 的 β 正域为 $POS_R^\beta(Y) = \{o_1, o_2, o_{12}, o_{16}, o_{17}\}$

概念 Y 的 β 负域为 $NEG_R^\beta(Y) = \{o_3, o_4, o_5, o_6, o_7, o_8, o_9, o_{10}, o_{11}, o_{13}, o_{14}, o_{15}, o_{18}\}$

概念 Y 的 β 边界域为 $BND_R^\beta(Y) = \emptyset$ 。

表 4-5 是一个针对概念 Y 的约简结果, Y 列为 1 的样例属于概念 Y 的 β 正域, Y 列为 0 的样例属于概念 Y 的 β 负域。这里, $\beta = 0.6$ 。

表 4-5 约简结果

序号	样例	属性				Y	Y 样例数	$\neg Y$ 的样例数
		a_1	a_5	a_7	a_9			
1	O_1, O_2	1	1	1	0	1	2	0
2	O_{12}, O_{16}, O_{17}	1	1	0	0	1	2	0
3	O_3	0	0	1	0	0	0	1
4	O_4	0	1	0	0	0	0	1
5	O_6	0	1	1	0	0	0	1
6	O_9, O_{18}	1	0	0	0	0	0	2
7	O_7	1	0	0	1	0	0	1
8	O_5, O_{14}	1	0	1	0	0	0	2
9	O_{13}	1	1	0	1	0	0	1
10	$O_8, O_{10}, O_{11}, O_{15}$	1	1	1	1	0	0	4

4.6.2 规则生成

根据约简后得到的信息系统, 对于约简结果中的每行样例子集 X_i , 可以直接得到如下形式的概率决策规则:

(1) $Des(X_i) \xrightarrow{C_i} Des(Y)$, if $P(Y|X_i) \geq \beta$;

(2) $Des(X_i) \xrightarrow{C_i} Des(\neg Y)$, if $P(Y|X_i) \leq 1 - \beta$ 。

其中, C_i 是规则的可信度因子, 在(1)式中等于 $P(Y|X_i)$, 在(2)式中等于 $1 - P(Y|X_i)$ 。

从表中可以得到如下规则:

$$\begin{aligned}
 &(a_1 = 1) \wedge (a_5 = 1) \wedge (a_7 = 1) \wedge (a_9 = 0) \xrightarrow{1} (d = 1) \\
 &(a_1 = 1) \wedge (a_5 = 1) \wedge (a_7 = 0) \wedge (a_9 = 0) \xrightarrow{0.67} (d = 1) \\
 &(a_1 = 0) \wedge (a_5 = 0) \wedge (a_7 = 1) \wedge (a_9 = 0) \xrightarrow{1} (d \neq 1) \\
 &(a_1 = 0) \wedge (a_5 = 1) \wedge (a_7 = 0) \wedge (a_9 = 0) \xrightarrow{1} (d \neq 1) \\
 &(a_1 = 0) \wedge (a_5 = 1) \wedge (a_7 = 1) \wedge (a_9 = 0) \xrightarrow{1} (d \neq 1) \\
 &(a_1 = 1) \wedge (a_5 = 0) \wedge (a_7 = 0) \wedge (a_9 = 0) \xrightarrow{1} (d \neq 1) \\
 &(a_1 = 1) \wedge (a_5 = 0) \wedge (a_7 = 0) \wedge (a_9 = 1) \xrightarrow{1} (d \neq 1) \\
 &(a_1 = 1) \wedge (a_5 = 0) \wedge (a_7 = 1) \wedge (a_9 = 0) \xrightarrow{1} (d \neq 1) \\
 &(a_1 = 1) \wedge (a_5 = 1) \wedge (a_7 = 0) \wedge (a_9 = 1) \xrightarrow{1} (d \neq 1) \\
 &(a_1 = 1) \wedge (a_5 = 1) \wedge (a_7 = 1) \wedge (a_9 = 1) \xrightarrow{1} (d \neq 1)
 \end{aligned}$$

这些规则里面一些条件属性是冗余的, 还需要通过值约简进行进一步简化。这里, 采用决策矩阵的方法进行值约简, 为了方便记录和分析, 分别用字母 a, b, c, e 表示属性 a_1, a_5, a_7, a_9 , 可得到决策矩阵如表 4-6 所示。

表 4-6 决策矩阵

序号	样本 个数	d=1		d=0			d=2					析取范式
		1	2	4	5	6	3	7	8	9	10	
1	2			ac	a	bc	ab	bce	b	ce	e	$a \wedge b \wedge e$
2	3			a	ac	b	abc	be	bc	e	ce	$a \wedge b \wedge e$
4	1	ac	a				bc	abe	abc	ae	ace	$(a \wedge b) \vee (a \wedge c)$
5	1	a	ac				b	abce	ab	ace	ae	$a \wedge b$
6	1	bc	b				ac	e	c	b	ce	$b \wedge c \wedge e$
3	2	ab	abc	bc	b	ac						$(a \wedge b) \vee (b \wedge c)$
7	1	bce	be	abe	abce	e						e
8	2	b	bc	abc	ab	c						$b \wedge c$
9	1	ce	e	ae	ace	b						$b \wedge e$
10	4	e	ce	ace	ae	ce						e

对每一行的样例子集, 将各属性匹配相应的属性值, 得到决策规则, 将相同的规则进行合并, 得到如下简化决策规则:

$$\begin{aligned}
 &(a_1 = 1) \wedge (a_5 = 1) \wedge (a_9 = 0) \xrightarrow{0.8} (d = 1) \\
 &(((a_1 = 0) \wedge (a_5 = 1)) \vee ((a_1 = 0) \wedge (a_7 = 0))) \xrightarrow{1} (d = 0)
 \end{aligned}$$

$$\begin{aligned}
 &(a_5 = 0) \wedge (a_7 = 0) \wedge (a_9 = 0) \xrightarrow{1} (d = 0) \\
 &((a_1 = 0) \wedge (a_5 = 0)) \vee ((a_5 = 0) \wedge (a_7 = 1)) \xrightarrow{1} (d = 2) \\
 &a_9 = 1 \xrightarrow{1} (d = 2) \\
 &(a_5 = 1) \wedge (a_9 = 1) \xrightarrow{1} (d = 2)
 \end{aligned}$$

4.6.3 试验结果

由研究区 18 个样本所得的规则可以看出, 矿体围岩及夹石为泥岩, 硅化和褐铁矿化是与金矿化关系极密切的蚀变类型, 蚀变程度高的单元含金量高, 位于整合地层, 并且围岩非泥岩, 蚀变无硅化或褐铁矿化的地段基本无矿。

结合本研究区的实际勘探情况, 本研究区金矿产于那洒破背斜北翼或倾伏端、加里东不整合面附近的硅化岩带及其与断裂带重合叠加部位; 金矿化富集地段以硅化、粘土化为主, 多种蚀变叠加时, 金矿化较强。目前发现的(锑)金矿体均赋存于坡松冲组底部(D_{1ps}^1)地层中, 区内 D_{1ps}^1 石英砂岩普遍硅化, 特别是裂隙发育地段硅化强, 局部地段(ZK25101 等部位)使石英砂岩硅化成石英(含水石英), 金矿体即产于硅化蚀变带内; 褐铁矿化主要在 D_{1ps}^1 地层中, 与硅化分布范围一致, 一般在硅化地段, 褐铁矿化越强, 金矿化也越强。

4.6.4 讨论

基于粗糙集提取的特征矿化信息基本与实际勘探信息相符。金矿(化)体的富集具明显的岩性选择性, 含矿岩性主要是硅化石英砂岩。属性石英砂岩同样与金矿化关系密切, 由于这一属性普遍存在于研究区地质单元中, 其在信息表中属性值均取值为 1, 对于不可分辨关系没有起到区分作用, 基于粗糙集思想无法提取。由于其包含的信息量小, 因此在进行量化分析时可忽略。

第5章 成矿地质单元的定量预测

矿产信息是各种成矿相关信息的综合体现,为了有效地提取成矿预测综合信息,有必要客观地筛选原始观测信息,突出成矿密切相关的致矿因子。粗糙集不需要数据的附加信息或先验知识,在知识库分类能力不变的前提条件下,删除无关或不重要的属性,能对决策系统进行有效约简。本研究探索基于粗糙集理论进行集成化预测模型研究的新方法,基于粗糙集思想提取与成矿密切相关的特征矿化信息,获取最佳变量组合及区间值,并将其作为参量建立预测模型,分别应用特征分析和人工神经网络模型对成矿地质单元进行定量预测,选取成矿有利单元,表明该方法能够有效降噪,简化模型,可为靶区预测提供准确的依据。

5.1 研究流程

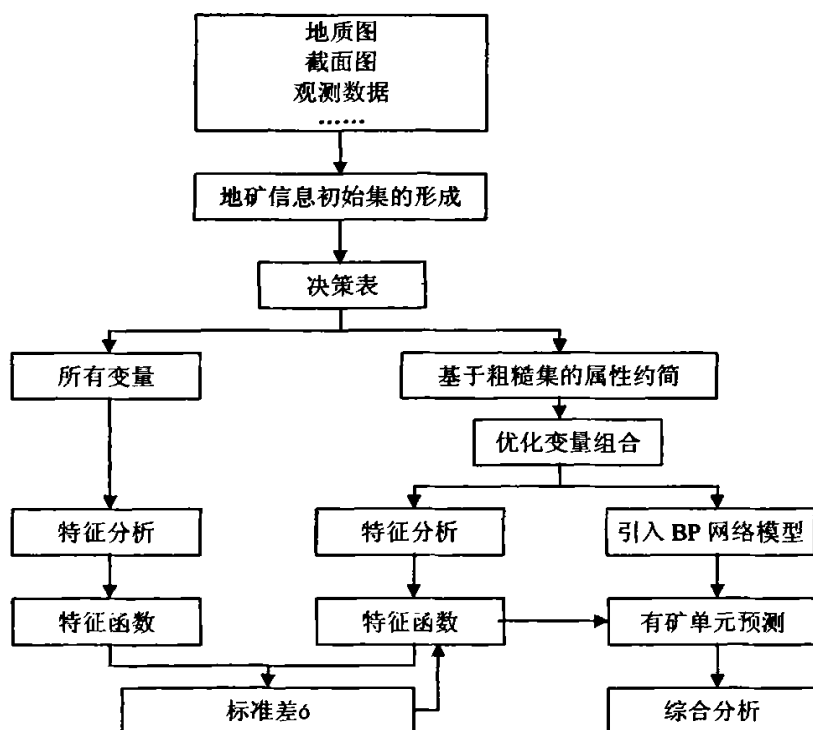


图 5-1 工作流程图

(1) 选择地质变量，建立找矿标志总体。提取研究区各种地质异常，组成初始数据集。

(2) 基于粗糙集方法的数据预处理。对于非数值属性进行编码，数值化数据集，对连续型属性值选择合适的方法离散化，形成决策表。

(3) 属性约简。直接从数据上进行分析，删除冗余属性和非特征属性，求出属性核。

(4) 建立靶区预测模型。

$$P = f(x_1, x_2, \dots, x_n) = \sum_{i=1}^n a_i x_i \quad (5-1)$$

其中

P —关联度；

a_i —各控矿标志权值；

x_i —各控矿因素分权。

(5) 预测。基于最优变量组合构建特征函数，对地质单元进行成矿预测。

5.2 粗糙-特征分析的应用

特征分析法是一种多元统计分析方法，通过研究区内已知地质单元的研究，查明地质变量之间的内在联系并确定它们的找矿意义，从而建立起特定类型矿床的定量模式^[52,53]。这种模式是该类型矿床共性的体现，是反映该类型矿床特征的成矿因素的特定组合。这种特定组合表现为若干个特征变量的最佳加权线性组合，称其为特征模型。预测时，将预测对象的地质特征与模型相比，用它们的相似程度表示预测对象的成矿可能性，据此圈定出有利成矿的各级远景区。

5.2.1 特征分析方法

目前通常使用的特征分析模型中变量权的确定方法有三种：乘积矩阵矢量长度法、乘积矩阵主分量法和概率矩阵主分量法。

5.2.1.1 乘积矩阵矢量长度法

此方法也称为平方和法，基本思想是变量与其它变量的关联性越强，变量就越重要。通过计算各地质变量的向量长来评价变量的重要性，向量长越大则该变量与矿化的关系越密切。

n 个地质单元的 m 个地质特征构成一个 $m \times n$ 矩阵 A

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \quad (5-2)$$

各元素 a_{ij} ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$) 为 1 或 0, 每一行是一个地质特征向量, 向量长为各元素平方和的平方根, 即

$$Li = \sqrt{\sum_{j=1}^n a_{ij}^2} \quad (5-3)$$

设 $B = AA'$, 则称乘积矩阵 B 中各行的向量长为逻辑向量。计算逻辑向量的长, 既考虑了某变量出现对成矿的意义, 同时又考虑了该变量和其它每一变量同时两两出现对成矿的意义。

5.2.1.2 乘积矩阵主分量法

在乘积矩阵主分量法中, 设 $B = A'A$, 给变量赋权的做法是求出矩阵 B 的最大特征值, 用它所对应的特征向量的分量作为响应的权。

设 λ 是 B 的特征值, ξ 是对应于 λ 的特征向量, 于是

$$B\xi = \lambda\xi$$

两边同时左乘 ξ' , 得

$$\xi'B\xi = \xi'A'A\xi = (A\xi)'(A\xi) = \lambda$$

将 A 按地质单元表示为向量形式, 有

$$A\xi = \begin{pmatrix} a_1\xi \\ a_2\xi \\ \vdots \\ a_n\xi \end{pmatrix} \quad (5-4)$$

$A\xi$ 的各个分量 $a_i\xi$ 是第 i 个单元在特征向量 ξ 上的投影长, 数量积 $(A\xi)'(A\xi) = \lambda$ 是各单位在特征向量 ξ 上的投影平方和。典型特征方向, 就是最大特征值所对应的特征向量的方向, 在这个方向上, 单元投影平方和最大, 或者说单元点有最大的离散程度。这个特征向量的各个分量, 即它在各原变量上的坐标, 表现了变量与特征向量之间的密切程度, 或者说它反映了变量在造成这个特征方向上所起的作用, 可作为各变量的权系数。乘积矩阵主分量法是一种比较常用的权系数确定方法。

5.2.1.3 概率矩阵主分量法

概率矩阵主分量法考虑变量之间的匹配概率，概率矩阵的主对角线规定为1，非对角线元素是两变量之间的匹配概率 P_{ij} 。变量的匹配表示了两两变量之间的关联，匹配数的大小反映了变量在各单元上有相同取值的情况。匹配概率定义为原始矩阵任意两变量的取值不变，而位置任意排列时所得匹配数不超过现实观测到的匹配数的概率。各属性的变量权就是概率矩阵最大特征值对应的特征向量。

5.2.2 模型的应用

设上述算法得到变量的权为 a_1, a_2, \dots, a_n ，这些权值的大小和符号反映了变量的重要性。将变量按绝对值大小排列，就可以区分变量的找矿意义。建立矿床的定量模型为

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n \quad (5-5)$$

y 被称为关联度，是地质单元在 n 个变量上的得分，反映矿化信息的一个综合指标。关联度越大说明某单元与模型的相似程度越高，即越具备该类型矿床的特征。结合某种方法或根据实际情况确定区分成矿远景好坏的临界值，就可以预测未知单元。因此，我们可以通过对模型单元关联度的研究得出有利成矿的数值范围并形成找矿标准，确定出有利成矿的远景地区。

5.2.3 建模参量的比较分析

5.2.3.1 基于最优变量建模

基于粗糙集方法，总结出综合找矿标志 $\{a_1, a_5, a_7, a_9\}$ ，使用特征分析方法进行定位及定量预测，以本研究区勘测程度高的单元作为成矿预测的模型单元，采用特征分析中的乘积矩阵主分量法，确定相应各变量的变量权，进而建立起本研究区典型矿床的特征分析模型如下：

$$Y = 0.6338 \cdot a_1 + 0.5570 \cdot a_5 + 0.4509 \cdot a_7 + 0.2911 \cdot a_9 \quad (5-6)$$

5.2.3.2 所有变量建模

若所有变量参与建模，采用同样的方法确定相应变量的变量权，进而建立起本研究区典型矿床的特征分析模型如下：

$$\begin{aligned} Y = & 0.4665 \cdot a_1 + 0.2150 \cdot a_2 + 0.0710 \cdot a_3 + 0.5399 \cdot a_4 + 0.3934 \cdot a_5 + 0.2693 \cdot a_6 \\ & + 0.3338 \cdot a_7 + 0.2329 \cdot a_8 + 0.2160 \cdot a_9 \end{aligned} \quad (5-7)$$

5.2.3.3 比较分析

对于这两个模型, 分别将研究区模型单元的地质标志取值逐一带入(5-6)式和(5-7)式, 即得到各单元的关联度 Y 值。本研究为了查明在地质环境中预测矿产的自然分类状况, 便于比较分析, 应用式子(4-1)极差变换将关联度和实际勘查品位值归一化处理, 分别表示成矿概率估计值和成矿概率参照值, 计算结果列于表 5-1。

表 5-1 估计值与参考值对照表

地质单元	模型(1)		模型(2)		勘查值	
	关联度	归一化(%)	关联度	归一化(%)	金品位	归一化(%)
3	1.6417	80.36	2.1815	67.87	1	50.00
4	1.6417	80.36	2.1815	67.87	1	50.00
6	0.4509	0.00	1.5909	33.76	2	100.00
7	0.557	7.16	1.1662	9.23	0	0.00
8	1.0847	42.77	1.5731	32.73	2	100.00
11	1.0079	37.59	1.5364	30.61	0	0.00
12	0.9249	31.99	1.7067	40.45	2	100.00
13	1.9328	100.00	2.1646	66.89	2	100.00
18	0.6338	12.34	1.0064	0.00	0	0.00
19	1.9328	100.00	2.7378	100.00	2	100.00
20	1.9328	100.00	1.9496	54.48	2	100.00
22	1.1908	49.93	1.6691	38.28	1	50.00
23	1.4819	69.57	1.8851	50.75	2	100.00
24	1.0847	42.77	1.4112	23.38	2	100.00
28	1.9328	100.00	2.1825	67.93	2	100.00
30	1.1908	49.93	1.6691	38.28	0	0.00
31	1.1908	49.93	1.6691	38.28	1	50.00
35	0.6338	12.34	1.2757	15.55	0	0.00

特征分析法中线性变换过程是将原线性空间分解成一些特征向量相关的子空间, 从而求取单元投影平方和最大的方向, 包含了不整合和断裂、不整合和辉绿岩等不同地质现象同时存在下对成矿的有力程度。从原始的地质信息收集到地质数据的二值化处理过程, 许多地质体或地质现象缺乏严格定义, 难免存

在一些抽象性和模糊性，多变量参与建模会产生较大的误差积累和传播，使得一些不重要的地质属性影响了投影方向的选取。基于粗糙集思想提取特征矿化信息本身是对成矿联系最密切的变量，在一定程度上剔除了部分噪音数据，能够减小误差的影响，达到更好的拟合效果。

估计值与参照值之间的误差采用公式

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{\text{估}} - x_{\text{参}})^2}$$
 (5-8)

最优组合变量参与建模，得 $\sigma = 0.3959$ ；全部变量参与建模，得 $\sigma' = 0.3986$ 。图 5-2 为估计值与参照值对照图，其中估计值 1 为约简变量参与建模预测结果，估计值 2 为所有变量参与建模预测结果，参照值为研究单元勘查值，估计值与参照值之间的误差表现为 $\sigma < \sigma'$ ，表明用约简后的组合变量{a₁, a₅, a₇, a₉}可以取代众多变量对研究区进行预测。

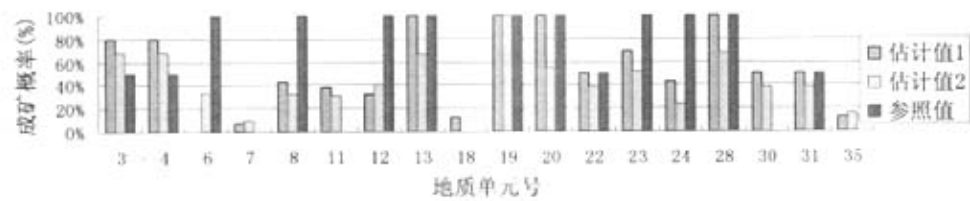


图 5-2 估计值与参照值对照图

5.2.4 成矿单元预测

表 5-2 为基于最优变量的成矿单元预测结果。

表 5-2 地质单元预测

地质单元	关联度	地质单元	关联度	地质单元	关联度
1 号	1.6417	16 号	0.9249	32 号	1.0847
2 号	1.6417	17 号	1.9328	33 号	1.9328
5 号	1.6417	21 号	0.6338	34 号	1.1908
9 号	0.4509	25 号	1.9328	36 号	1.1908
10 号	0.557	26 号	1.9328	37 号	0.6338
14 号	1.0847	27 号	1.1908		
15 号	1.0079	29 号	1.4819		

根据实际勘测情况，关联度 $Y < 1.1908$ 的地质单元可认为无勘探价值， $1.1908 \leq Y < 1.6417$ 的地质单元可认为成矿概率预测为低，有 $Y \geq 1.6417$ 的地质单元可

认为成矿概率预测为高。预测结果为：1、2、5、17、25、26、33号地质单元的勘探价值较高；27、29、34、36号地质单元的勘探价值较低；9、10、14、15、16、21、32、37号地质单元无勘探价值。

粗糙集可以客观地筛选诸多变量，突出与成矿密切相关的信息，得到最佳变量组合，在此基础上构建特征函数，能够有效的降噪，找到最有利找矿标志或找矿标志的组合数值区间，为靶区预测提供客观依据。

5.3 粗糙-神经网络模型的应用

粗糙集方法模拟人的抽象逻辑思维，是基于不可分辨思想对知识进行简化，从数据中推理逻辑规则作为知识系统的模型，而神经网络方法模拟形象直觉思维，利用非线性映射思想，用网络结构表达输入与输出关联知识的隐函数编码。神经网络一般不支持具有语义形式的输入，无法确定哪些知识是冗余的，哪些知识是有用的，因此，对样本进行网络训练时，由于样本数量大，属性维数高，往往造成网络规模加大，训练过程变得复杂而漫长。粗糙集理论可以对定性、定量或者混合信息进行分析，确定知识表达中不同属性重要性，使知识表达空间简化。目前，神经网络方法被广泛地应用到各个领域的非线性问题中，在解决地质问题也表现出其独特的优势^[54,55]。传统的矿产资源定量预测评价方法主要是统计学方法，包括多元统计和地质统计学。这种方法在解决资源定量预测问题中需要一些假设和简化的约束条件作为建立数学模型的基础，如样本正态分布等，这些假设往往与实际地质情况有一定偏差，因而所得结果也就与地质事实有一定差距。人工神经网络方法不要求有假设条件，也不要求事先搞清楚控矿条件和找矿标志与资源体及资源量之间的定量关系，人工神经网络通过训练或运算，自己会找出标志与资源之间的定量关系，且常常是非线性的，并以隐式方式存储于网络权阵中。人工神经网络中用于解决地矿资源判别分类问题的模型主要是BP模型。该模型采用S学习规则进行有监督分类，它需要相当数量的已知样本进行模型训练，以便找出且记忆输入模式与分类类别之间的映射关系。通常把需要分类的地质对象的条件集合或特征组合作为BP网络的输入模式，并给出期望输出模式(分类或预测类型)。经训练后，BP网络就具有了判别分类的能力。

5.3.1 BP 网络

BP网络是一种具有三层或三层以上神经元的多层前向网络，也是在模式识

别和分类方面发展最早、研究最多、应用最为广泛的一类人工神经网络模型。作为促进人工神经网络研究第二次热潮开始的一个重要方面，多层前向网络在人工神经网络研究的发展过程中占有重要的地位。它提供了描述复杂非线性映射和分类的一般性方法。这一观点的理论基础来源于 K. Funahashi, G. Cybenko 和 K. Hornik 等人关于多层前向网络的映射能力所做的研究工作，他们一致的研究结果表明：只要网络规模足够大，即网络中隐含层节点数足够多，多层前向网络（甚至只要一个隐含层）能够以任意精度逼近（或表达）维度空间上的任意的连续函数^[56]。

BP 网络按有导师学习的方式进行训练，训练模式包括若干对输入模式和期望的目标输出模式。网络训练时，由输入信息的正向传播和误差的反向传播两个过程组成。在正向传播过程中，输入信息从输入层到隐含层再到输出层进行逐层处理，每一层神经元的状态只影响下一层神经元的状态，如果输出层的输出与给出的样本希望输出不一致，则计算出输出误差，转入误差反向传播过程。将误差沿原来的连接通路返回。通过修改各层神经元之间的权值，使得误差达到最小。这两个传播过程在网络中反复运行，使网络误差不断减小，从而网络对输入模式的响应的正确率也不断提高，当网络误差不大于目标误差时，网络训练结束。

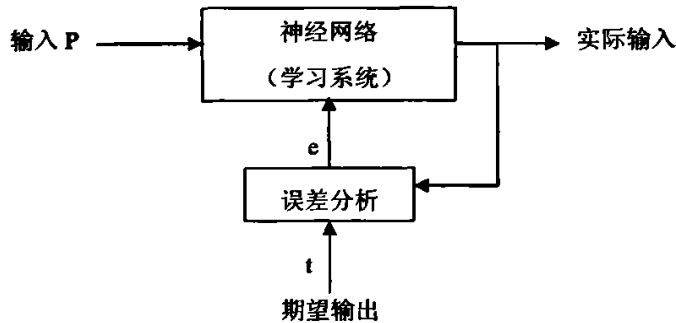


图 5-3 BP 网络工作原理

5.3.2 功能的映射关系

输入观察矢量 $X = [x_1, x_2, \dots, x_N]$ 是一个 N 维随机矢量。 $X \in R^N$ ，即其每一个分量可以取任意实数，一个 X 相当于实 N 维空间中的一个点。

设所有 X 可以分成 M 类，记为 $C_j, j = 1, 2, \dots, M$ 。与某个类别 C_j 相应的所有随机输入矢量 X 的集合在 R^N 中构成一个集合 R_j 。分类功能就是要求能够根据

某个输入 X 属于第 C_j 类时, 令输出矢量 Y 的第 j 个分量等于 1, 而其他分量等于 0, 数学表达如下:

$$\text{若 } X \in R_j, \text{ 则 } y_k = \begin{cases} 1, & \text{当 } k = j \\ 0, & \text{当 } k \neq j \end{cases}$$

事实上, 对一个实际的神经网络, 严格要求实现此形式会造成系统实现的困难, 而且也没有必要。其实, 只要尽量接近于 1 或 0 就可以了, 因此:

$$\text{若 } y_j = \max_k \{y_k\}, \text{ 则 } X \in R_j$$

5.3.3 网络结构

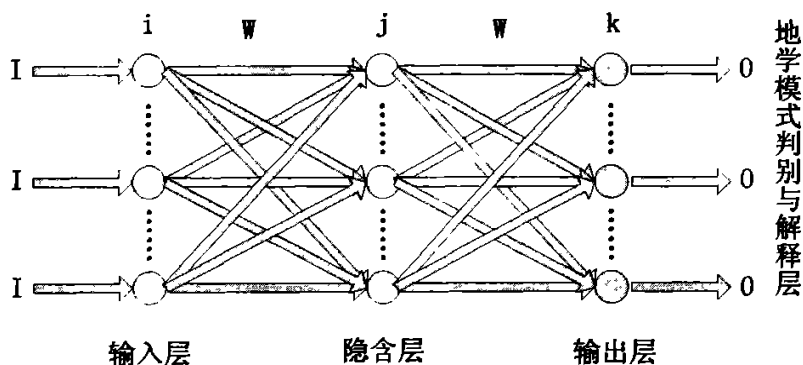


图 5-4 BP 网络结构

(1) 激励函数

神经元对输入信号处理通常分为求和与函数运算两个过程。其中函数运算主要通过激发函数对求和结果进行运算完成的, 基本作用是对输入、输出进行函数转换。BP 网络多采用 S 型函数, 包括正切 S 型函数和对数 S 型函数。本实验是对地质单元成矿可能性的判定, 即成矿概率的分析, 因此, BP 网络的隐含层神经元采用正切 S 型激发函数, 输出层神经元采用对数 S 型函数, 使得输出结果在 $[0, 1]$ 的范围内。

(2) 输入、输出层

BP 网络输入输出层维数是根据地质单元的属性描述来决定的。本模型中, 输入维数为粗糙集针对地质信息约简后的维数, 为 4, 可以方便表示为一个四维输入向量, 即不整合、泥岩、硅化和褐铁矿化, 网络中的输入节点数为 4。对地质单元成矿概率的判定, 分为三种情况, 有矿 (含金品位高)、有矿 (含金品位低) 和无矿 (含金品位为零), 因此, 输出为三维向量, 输出层节点数为 3。

(3) 隐含层及节点

隐含层起抽象的作用，即能够从输入提取特征。增加隐含层可以增加神经元网络的处理能力，但训练也将复杂化。对于一般的模式识别问题，三层网络可以很好的解决问题^[57]，因此，本研究即采用三层（输入层、隐层、输出层）前向网络对地质数据进行学习，并运用训练的网络对预测单元进行分类，达到成矿预测的目的。三层网络中，隐含层神经元个数 n_2 和输入层神经元个数 n_1 之间有以下近似关系 $n_2=2 n_1+1$ ，因此网络隐含层的神经元个数近似为 9。

5.3.4 基本算法

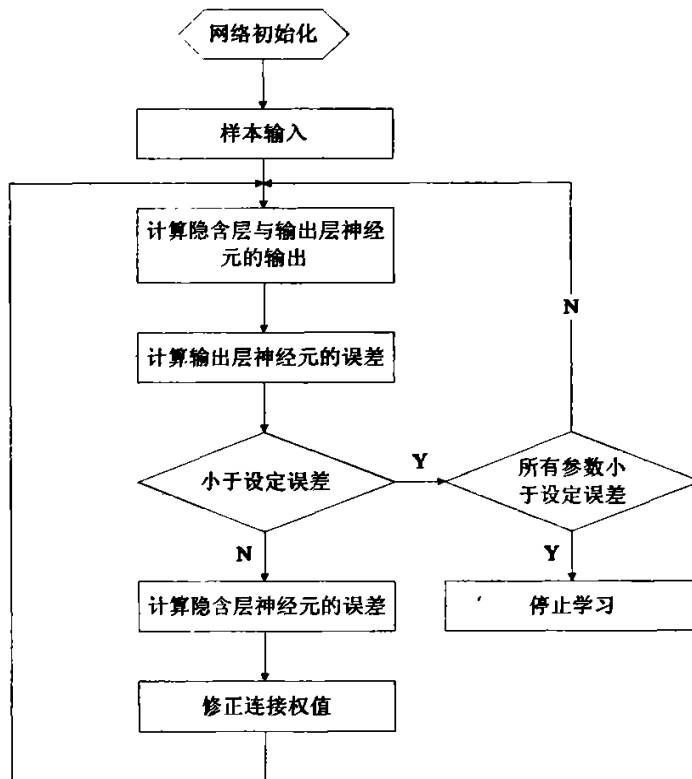


图 5-5 BP 网络算法流程图

训练流程如图 5-5 所示：

样本：（输入向量，理想输出向量）

权初始化：为避免权值调整方向同向，即权值同时增加或同时减小，应该选取均匀分布的小随机数，大概为 $(-2.4/F, 2.4/F)$ ，其中 F 为所连单元的输

入端个数。

- (1) 从样本集中取一个样本 (X_p, Y_p) ，将 X_p 输入网络；
- (2) 计算相应的实际输出 o_p ： $o_p = f_2(f_1(X_p \varpi_{ji}) \varpi_{kj})$
- (3) 计算实际输出 o_p 与相应的理想输出 Y_p 的差；
- (4) 网络关于第 p 个样本的误差测度：

$$E_p = \frac{1}{2} \sum_k (t_{pk} - o_{pk})^2$$

- (5) 网络关于整个样本集的误差测度：

$$E = \sum_p E_p$$

(6) 如果 E 小于预先设定的某一精度 ε ，则训练结束，否则调整连接权值，重新进行样本学习，直到 $E < \varepsilon$ 为止。

输出层： $\varpi_{kj}(n+1) = \varpi_{kj} + \Delta \varpi_{kj}$

隐含层： $\varpi_{ji}(n+1) = \varpi_{ji} + \Delta \varpi_{ji}$

5.3.5 成矿单元预测

以研究区勘探资料为依据，勘探程度高的地质单元为模型单元，运用 BP 神经网络，对研究区地质单元进行 BP 算法分类预测。分类预测的结果可以对该矿区勘探程度低的区域起到一定的指导作用，为地质矿产工作提供新的基础理论依据。

根据第四章中阐述的金矿床的综合信息以及基于粗糙集理论提取的特征矿化信息，得到金矿控矿变量组合：不整合、泥岩、硅化和褐铁矿化。

BP 网络的参数设置为：

网络训练目标误差： $\varepsilon = 0.01$ ；

最大循环次数为 1000；

学习速率为 0.1

表 5-3 BP 网络输入输出参数值

序号	单元号	特征矿化信息				成矿概率		
		不整合	泥岩	硅化	褐铁矿化	零	低	高
1	3	1	1	1	0	0	1	0
2	4	1	1	1	0	0	1	0
3	6	0	0	1	0	0	0	1

续表 5-3

序号	单元号	特征矿化信息				成矿概率		
		不整合	泥岩	硅化	褐铁矿化	零	低	高
4	7	0	1	0	0	1	0	0
5	8	1	0	1	0	0	0	1
6	11	0	1	1	0	1	0	0
7	12	1	0	0	1	0	0	1
8	13	1	1	1	1	0	0	1
9	18	1	0	0	0	1	0	0
10	19	1	1	1	1	0	0	1
11	20	1	1	1	1	0	0	1
12	22	1	1	0	0	0	1	0
13	23	1	1	0	1	0	0	1
14	24	1	0	1	0	0	0	1
15	28	1	1	1	1	0	0	1
16	30	1	1	0	0	1	0	0
17	31	1	1	0	0	0	1	0
18	35	1	0	0	0	1	0	0

图 5-6 为误差曲线下降图。

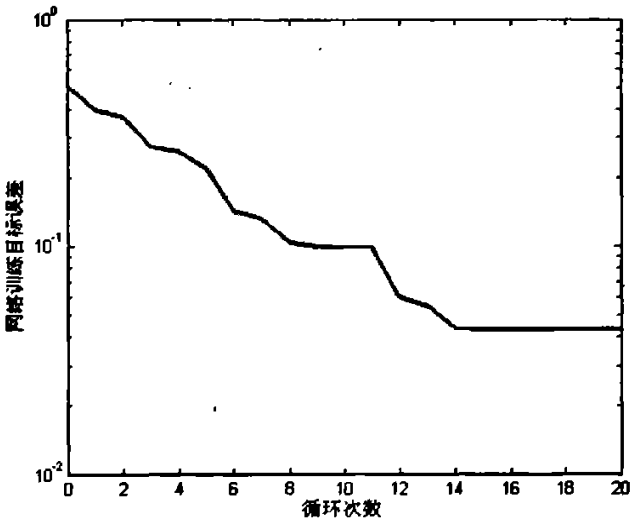


图 5-6 误差曲线下降图

迭代 15 次后，目标误差趋于稳定，最小误差：0.0432099

将预测单元地质数据作为 BP 网络预测的输入样品数据，输入到训练好的 BP 网络中，得到 BP 预测结果如表 5-4 所示。

表 5-4 地质单元预测结果

序号	单元号	成矿概率			序号	单元号	成矿概率		
		零	低	高			零	低	高
1	1, 2	0.0000	0.0000	1.0000	10	25	0.0000	0.0000	1.0000
2	5	0.0000	0.0000	1.0000	11	26	0.3333	0.6667	0.0000
3	9	0.0000	0.0000	1.0000	12	27	0.0000	0.0000	1.0000
4	10	1.0000	0.0000	0.0000	13	29	0.0000	0.0000	1.0000
5	14	0.0000	0.0000	1.0000	14	32	0.0000	0.0000	1.0000
6	15	0.0000	1.0000	0.0000	15	33	1.0000	0.0000	0.0000
7	16	0.0000	1.0000	0.0000	16	34	1.0000	0.0000	0.0000
8	17	0.0000	0.6848	1.0000	17	36	1.0000	0.0000	0.0000
9	21	0.0057	0.0000	0.0000	18	37	1.0000	0.0000	0.0000

由预测结果可知，1、2、5、9、14、17、25、27、29、32 号单元的成矿概率预测为高，应该进行进一步勘探以查明矿床分布；15、16、26 号单元成矿概率预测为低，可综合考虑地质情况和专家分析，确定进一步勘探的价值；10、21、33、34、36、37 号单元无勘探价值。

从大量观察和实验数据获取知识、表达知识、推理决策规则是智能信息处理的重要任务，特别是对于不准确、不完整的知识，粗集理论方法和人工神经网络方法都显示了无穷的魅力。粗集理论从训练数据中推理规则，定义条件属性和决策属性间的依赖关系，即输入空间与输出空间的映射关系是通过简单的决策表简化得到的，而且通过去掉冗余属性，可以大大简化知识的表达空间维数。利用粗糙集的知识约简理论作为神经网络模型的前置装置，能够有效减低神经网络的复杂结构，缩短神经网络的训练时间，提高模型的预测精度。

5.4 小结

目前的许多地学问题的研究中往往包含一些经验性的成分，所建立的数学模型具有似然性。为了有效地提取特征矿化信息进行成矿预测，必须采用一定的数据模型对各类无直观规律的数据集进行整理、分析，把握数据分布的规律

性。地质成矿问题定量化分析主要是将地质模型转换为数学模型，关键问题主要体现在以下三个方面：

1. 准确定义地质体或地质现象；
2. 在解决各类地质问题中，给出数量准则；
3. 正确处理地质数据。

地质数据本身具有噪音强、混合性强、区域性强等特点，本研究中地质单元的划分以及模型单元的选取，存在着一定的抽象性和随机性，在地质变量的选取方面也较单一，只采用了地质构造方面的信息，没有直接演算地球物理、地球化学相关数据，难免造成地质数学方法和手段的精确性、严格性与原始地质资料的描述性、概略性之间的矛盾。

成矿单元预测的思想主要是将模型单元与预测单元进行类比分析，特征分析强调的是地质变量之间的内在联系，神经网络模型强调的是地质变量与成矿判别之间的映射。在实验结果上，成矿概率预测为低的地质单元为不确定单元，需进一步分析，其余单元中，只有 9、14、32、33 号地质单元预测结果不符，有 78.9%的结果相吻合，说明粗糙集理论集成化预测模型研究方法具有一定参考意义，综合两种方法，1、2、5、17、25 号地质单元有进一步勘探价值。

对地质构造资料进行成矿分析，能够在一定程度上反映研究区地质现象与地质成矿之间的联系和规律，验证本研究中地矿定量化分析这一探索性工作的实效和价值，其解决实际地质问题的有效性还有待于更多的生产实践的检验。

第6章 总结与展望

6.1 全文总结

随着地质生产和研究工作的大量开展,正确而迅速地处理数据,最大限度地获取有用信息已成为矿产资源预测所面临的新问题,为了准确地定义地质体或地质现象,给出数量的准则,建立数学模型进行成矿预测已成为当今隐伏矿体寻找的重要手段。找矿模型是在成矿规律研究的基础上,通过对矿床(体)的地质、物探、化探、遥感诸方面信息显示特征的充分发掘及综合分析,从中优选出那些有效的、具单解性的信息作为找矿标志,并在确定了找矿标志和找矿方法的最佳组合后才建立起来的。因此,在进行信息合成产生新知识时,只有采用一定的数据模型对各类无直观规律的数据集进行整理、分析,把握数据分布的规律性,才能进行不同种类的信息集成工作。一般成矿作用,通常理解为多种地质作用相互耦合叠加的结果。通过信息关联而确定的有用信息的叠合部位或信息浓集区,则被认为是成矿可能性最大的空间地段。这种成矿可能性最大的空间地段的认识的得出即是信息提取的一种物化表现。

本研究主要做了两方面的工作,一是基于粗糙集思想研究地质过程中各种因素及其相互关系,提取与矿化密切相关的特征信息;二是研究适合地质任务和地质数据特点的数学分析方法,建立地质单元成矿预测的数学模型。具体工作内容如下:

(1) 学习和应用粗糙集理论对地矿信息进行分析。根据收集资料提取相关信息,组成数据集,构建决策表。利用粗糙集对属性值进行约简,找出各种地矿信息对成矿的关联度和重要度,得到规则知识,并给出合理的解释。这个过程减少了地质信息的空间维数,简化系统,同时又不损失与研究对象有直接和间接联系的主要信息。

(2) 地质单元的成矿定量预测。粗糙集理论的计算方法是知识的表达和约简,可以描述对象组成的集合之间的关系,从而提取规则知识。要实现矿产信息的定量分析,必须构造连续特征函数,即要查明各种控矿因素和找矿标志的找矿信息量。基于地质数据的特点,数据类型都是二态变量,本研究选取传统的特征分析法和目前非线性问题研究热点之一的神经网络两个模型进行分析。BP神经网络和特征分析法都是通过对已知的矿床的成矿因素的研究,选择控制单元,提取地质变量,所以它们对地质单元的划分、变量的选取可以是一致的,

这为非线性和线性分析的有机结合奠定了坚实的基础,对提高矿产预测的可靠性有很大的实际意义。

本研究将粗糙集方法与特征分析以及神经网络模型相结合,对研究区地质单元进行了成矿定量预测。结果表明,粗糙—特征分析模型可以很好利用尽量少的地质信息进行成矿单元预测,与全变量参与建模相比,产生较小的预测误差,降低了地质数据的信息维数达到同样甚至更好的预测效果。利用粗糙—BP神经网络模型对成矿单元进行预测,粗糙集对原始数据的处理后,简化了样本的属性数据,同时能较好的反映系统的实际情况。BP 人工神经网络可方便的对各种复杂的非线性关系进行拟合,提高预测精度。粗糙-BP神经网络预测模型强化了建模的灵活性,根据实验结果来看可以很好的预测成矿地质单元。

6.2 本研究的创新点

通过对研究区真实数据进行试验,运用粗糙集理论提取重要控矿地质因素,结合特征分析方法和人工神经网络模型,找到控矿信息之间的关系以及对成矿作用的贡献率,如各因素标志的信息量或信息权,找到各因素标志最有利成矿或找矿的数值区间,直接从数值上评定矿点性质,为靶区预测提供科学的依据,本研究的创新点主要有以下两点:

(1) 应用粗糙集理论提取矿化信息,将多类型数据综合分析,克服人为干预。粗糙集不依赖任何先验知识,同时考虑定性和定量因素,通过对变量赋值或者离散化形成决策表,进而对决策表进行分析,完全基于数据评估地矿因素的贡献率,挖掘各个地质因素之间的联系和规则。利用粗糙集对诸多地质属性进行约简,对地质变量进行筛选,突出成矿密切相关的变量,得到最佳变量组合,有效地提取成矿预测信息。

(2) 以粗糙集约简思想为基础,以约简的矿化特征信息为输入数据,分别采用传统的统计预测方法以及人工智能建模对成矿单元进行定量预测,实现了线性理论中的特征分析与非线性理论中的神经网络的矿产资源预测评价,实验证明,预测结果能准确地反映出实际情况,指导实际操作。从思想上,迎合了地质问题定性向定量发展的大趋势;从方法上,在分形理论、灰色理论和地质统计在地矿分析领域成熟应用的今天,选择粗集方法作为理论基础是一种新的尝试。

6.3 展望

矿产资源定量预测的数学模型是由地质概念模型发展而成的定量模型，是在定性研究和定性预测基础上的定量预测，因此数学模型的建立必须以坚实的地质分析所建立的正确地质概念模型为基础。本文试采取粗糙集这一新方法提取矿化特征信息，结合数学建模，对地质单元进行成矿预测，在研究过程中，存在许多不成熟不完善的地方，需要在以下几个方面作进一步学习和研究：

(1) 变地质模型为数学模型，抽掉了地质学的具体内容，抽象地研究各种地质现象的数量关系和空间形式，而地质作用往往是在漫长的时间、复杂的介质并以巨大规模进行的，因此，各种数学模型实验具有一定的局限性。应用数学方法解决地质问题，必须充分了解和考虑地质任务、方法的特点和存在的问题。本研究收集整理了研究区地质构造数据，将地质信息直接转换成二态数据，适用于粗糙集理论进行分析，没有直接演算大量的地球化学、地球物理相关数据，在某种程度上影响并降低了预测准确度。因此，必须研究粗糙集理论和多智能体技术对矿产预测数据的离散化方法，结合地质单元不同形式的地质数据进行综合分析，提高预测精度。

(2) 经学者们研究，人工神经网络已经发展了多种改进算法，本研究采用最基本的 BP 神经网络模型，基于较少的样本及实验数据，可以达到较好的效果，如果运用于实践，将面临更复杂更大量的训练数据，可能会造成网络训练陷入错误的工作模式，如可能落入局部极小点，使算法不收敛，在某些区域内连接权值调整缓慢的情况，因此，应该学习和研究改进的神经网络算法，如附加动量项修正权值等，研究更具实际价值的成矿预测模型。

矿产资源定量预测是对取样鉴定和地质制图两大传统方法的革新和改造，随着研究者对地质过程性质认识的不断深化，数学理论和方法持续进步，非线性数学思想与方法在地质中渗透与应用，成矿预测数学模型的建立将更为有效利用已有地质资料来科学推测未知矿体，这是矿产勘查方法体系和工程实践的关键，因此，建立合理而有效的成矿预测数学模型具有十分重要的意义。

参考文献

- [1] Cox D P, Singer D A. Mineral deposit models[M]. U.S. Geological Survey Bulletin, 1986
- [2] 肖克炎, 朱裕生, 宋国耀. 矿产资源 GIS 定量评价[J]. 中国地质, 2000, 7: 29~32
- [3] 弓小平, 王世称, 杨兴科, 等. 地质矿产预测信息化相关问题的探讨[J]. 地质找矿论丛, 2005, 20(1): 66~70
- [4] Chen Jianping, Wang Gongwen, Hou Changbo. Quantitative Prediction and Evaluation of Mineral Resources Based on GIS: A Case Study in Sanjiang Region, Southwestern China[J]. Natural Resources Research, 2005, 14(4): 285~294
- [5] Ye Shuisheng, Wang Shicheng, Li Deqiong. Application of GIS in Mineral Resource Prediction of Synthetic Information[J]. Journal of China University of Geosciences, 2003, 14(3): 234~241
- [6] 赵鹏大, 池顺都. 当今矿产勘探问题的思考[J]. 中国地质大学学报, 1998, 23(1): 70~74
- [7] 赵鹏大, 孟宪国. 地质学的量化问题[J]. 中国地质大学学报, 1992, 17: 51~56
- [8] Singer Donald A. Some Suggested Future Directions of Quantitative Resource Assessment[J]. Journal of China University of Geosciences, 2001, 12(1): 40~44
- [9] Wang Shicheng, Ye Shuisheng, Zhou Dongdai. Theory and Method of Mineral Resource Prediction Based on Synthetic Information[J]. Journal of China University of Geosciences, 2003, 14(3): 207~214
- [10] 赵鹏大. “三联式”资源定量预测与评价——数字找矿理论与实践探讨[J]. 地球科学——中国地质大学学报, 2002, 27(5): 482~488
- [11] 万丽, 王庆飞, 高帮飞. 成矿预测中的非线性数学方法[J]. 地质找矿论丛, 2006, 21 (1): 45~48
- [12] 陈建平, 张寿延, 汤军, 等. 非传统矿产资源定量预测的理论思考[J]. 地球物理学进展 2006, 17(2): 342~347
- [13] Sinclair, A.J., G.J. Woodsworth. Multiple regression as a method of estimating exploration potential in an area near Terrace[J]. Economic Geology, 1970, 65: 998-1003
- [14] Geoffroy, J.D., Wignall, T.K.. Statistical decision in regional exploration: application of regression and Bayesian classification analysis in the southwest Wisconsin zinc area[J]. Economic Geology, 1970, 65: 769-777
- [15] 翟裕生, 邓军, 崔彬, 等. 成矿系统及综合地质异常[J]. 现代地质, 1999, 13(1): 100~103
- [16] 肖斌, 潘懋, 赵鹏大, 等. 矿产资源定量评价在火山岩铀资源中的应用研究[J]. 矿床地质, 2001, 20(3): 285~290

- [17] 李随民, 姚书振, 周宗桂. 矿产资源定量预测的研究现状[J]. 地质找矿论丛, 2007, 22(1): 9~12
- [18] Pawalai Kraipeerapun, Chun Che Fung, Warick Brown, et al.. Uncertainty in Mineral Prospectivity Prediction[J]. Neural Information Processing, 2006, 4233: 841~849
- [19] Brown, W.M., et al.. Artificial neural networks: a new method for mineral prospectivity mapping[J]. Australian Journal of Earth Sciences, 2000, 47: 757~770
- [20] 王丽, 冯山. 基于模糊粗糙集的两种属性约简算法[J]. 计算机应用, 2006(3): 635~637
- [21] 黄艳新, 于哲舟, 胡成全, 等. 粗糙集模糊神经网络味觉信号识别系统[J]. 吉林大学学报, 2004, 22(3): 236~243
- [22] 任永功, 王杨, 闫德勤. 基于遗传算法的粗糙集属性约简算法[J]. 小型微型计算机系统, 2006, 27(5): 862~865
- [23] 刘富春. 变精度集对粗糙集模型中的属性约简[J]. 计算机工程与应用, 2006(5): 8~11
- [24] 钟根元, 王方华, 范小军, 等. 基于变精度粗糙集模型的定量数据挖掘算法[J]. 上海交通大学学报, 2004, 38(5): 764~767
- [25] 何维, 马廷淮. 基于可变精度的加权粗糙集模型在约简中的研究[J]. 电子科技大学学报(社科版), 2006(8): 146~149
- [26] 周平, 丁进良, 岳恒等. 基于相似粗糙集的案例特征权值确定新方法[J]. 信息与控制, 2006, 35(3): 329~334
- [27] 王永茂, 刘克勤. 粗糙集理论在电力系统燃料管理中的应用[J]. 现代电力, 2002, 19(5): 88~93
- [28] 黄艳新, 于哲舟, 胡成全, 等. 粗糙集模糊神经网络味觉信号识别系统[J]. 吉林大学学报, 2004, 22(3): 236~243
- [29] 王霞, 唐德善. 基于粗糙模糊集的区域水资源系统的评价方法[J]. 水利规划与设计, 2006(1): 31~34
- [30] Guocheng Pan, Deverle P. Harris. Decomposed and Weighted Characteristic Analysis for the Quantitative Estimation of Mineral Resources[J]. Mathematical Geology, 1992, 24(7): 807~823
- [31] 周勃, 费朝阳, 刘绍宇, 等. 基于粗糙集理论的室内环境模糊评价方法[J]. 暖通空调 HV&AC, 2006, 36(4): 21~24
- [32] 姜云, 吴立新, 车德福, 等. 基于粗糙集和 GeoMo3D 的城市岩土参数重要性评估与可视化[J]. 地理与地理信息科学, 2005, 21(6): 19~21
- [33] 侯运炳, 潘启新. 基于粗糙集决策方法的矿山不确定多属性问题的研究[J]. 煤炭工程, 2005(4): 52~54
- [34] 王向阳, 蔡念, 杨杰, 等. 基于近似精度和条件信息熵的粗糙集不确定性度量方法[J]. 上海交通大学学报, 2006, 40(7): 1130~1135

- [35] 石晓, 赵庆飞. 基于粗糙集理论的模糊综合评判权值确定[J]. 西南石油学院学报, 2001, 23(3): 16~18
- [36] Zdzislaw Pawlak. In Pursuit of Patterns in Data Reasoning from Data: The Rough Set Way[R]. In: J. Alpigini, J.F. Peters, A. Skowron, N. Zhong, Eds., Rough Sets and Current Trends in Computing: Third International Conference, 2002: 1~9
- [37] Zdzislaw Pawlak. Reasoning about Data: A Rough Set Perspective[R]. In: L. Polkowski, A. Skowron, Eds., Rough Sets and Current Trends in Computing: First International Conference, 1998: 25~34
- [38] 袁艳斌, 汪新庆, 刘刚. 三峡坝区工程地质信息系统集成开发研究[J]. 地球科学——中国地质大学学报, 1999, 24(5): 542~544
- [39] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001
- [40] 陈晓红, 陈岚. 基于粗糙集理论的知识约简及应用实例[J]. 大学数学, 2003 19(4): 68~73
- [41] 冷永刚. 粗糙集理论约简算法的研究[D]. 杭州: 电子科技大学硕士论文, 2004
- [42] 葛丽. 粗糙集在海量科学数据挖掘中的应用[D]. 杭州: 电子科技大学硕士论文, 2004
- [43] 刘永军. 大数据集的属性选择算法的研究与实现[D]. 沈阳: 东北大学硕士论文, 2005
- [44] 阎桦. 基于粗糙集的数据挖掘约简算法的研究与应用[D]. 重庆: 西南大学硕士论文, 2006
- [45] 刘道华, 原思聪, 江祥奎, 等. 基于知识的神经网络产生式规则的获取方法研究[J]. 西安建筑科技大学学报(自然科学版), 2007, 39(3): 423~428
- [46] 杨雪银. GIS 空间数据模型与数据存储初探[J]. 西南林学院学报, 2003, 23(2): 53-60
- [47] 袁峰, 周涛发, 岳书仓. GIS 矿产资源预测中的证据权重法[J]. 黄金地质, 2003, 9(3): 75~77
- [48] 赵鹏大, 胡旺亮, 李紫金. 矿床统计预测[M]. 北京: 地质出版社, 1983
- [49] 王世称, 成秋明, 范继璋. 金矿资源综合信息评价方法[M]. 长春: 吉林科学技术出版社, 1990
- [50] Guo Jiayuan. Rough Set---Based Approach to Rule Generation and Rule Induction[J]. International Journal of General Systems, 2002, 31(6): 601~617
- [51] Pawlack Z. Rough Sets[J]. Communications of the ACM, 1995, 38(11): 89~95
- [52] 谢贵明, 范继璋. 吉林省珲春东部地区金矿综合信息找矿模型及找矿靶区预测[J]. 黄金科学技术, 2000, 8(5): 20-27
- [53] 黄晓乃, 畅文生, 杨建明, 等. 某铀矿床成矿预测[J]. 矿业研究与发展, 2003, 23(6): 37-39