

GIS 环境下地质变量自动提取与地质异常的圈定

曹 瑜¹ 胡光道² 杨志峰¹ 沈珍瑶¹

¹ (北京师范大学环境科学研究所, 北京 100875)

² (中国地质大学(武汉)数学地质研究所, 武汉 430074)

E-mail: cy_610@sohu.com

摘 要 通过对研究区合理的网格单元划分, 介绍地理信息系统(GIS)提取地质变量的原理, 即利用 GIS 的空间叠加功能, 实现地质变量自动提取。在此基础上, 进行变量综合, 挖掘隐含信息。过滤与已知矿产有关的变量, 圈定地质异常, 它是研究致矿异常和发现新矿产地的基础。以云南省维西地区为实例, 按上述方法原理, 提取其地质变量, 圈定其地质异常。已发现的矿体(包括矿化点)中, 有 76.84% 矿体(含矿化点)落入异常区内。

关键词 GIS 环境 地质变量 自动提取 地质异常圈定

文章编号 1002-8331-(2003)14-0081-05 文献标识码 A 中图分类号 P208

Geo-variables Extraction and Geo-anomaly Circle Based on GIS

Cao Yu¹ Hu Guangdao² Yang Zhifeng¹ Shen Zhenyao¹

¹ State Key Laboratory of Water Environment Simulation, Institute of Environmental Science, Beijing Normal University, Beijing 100875)

² Faculty of Earth Resources, China University of Geoscience, Wuhan 430074)

Abstract: GIS is an effective tool to process and analyze spatial geological information. Since 1980s, with the mature of the GIS technology, GIS is widely used in many areas. Although a lot of works has been done for the mineral resources quantitative appraisal with GIS, few studies are conducted through extracting, screening and synthesizing geological variables using GIS for the mineral resources quantitative appraisal. In this paper, some data involving geology and geo-chemistry have been collected to establish the database. With the help of GIS technology, these data are integrated to circling geo-anomaly. When suitable grid cells are ascertained in the research area, GIS spatial overlaying functions are used to extract geo-information. Then some geological variables are synthesized to burrow implicit geological information such as composite entropy, beneficial degree of fault and center symmetry degree of fault. The analysis method of ore-finding beneficial degree in linear anomaly and plane anomaly is introduced to filtrate the geological variables relative to mineral resources. Geo-anomaly model based on GIS is expounded. The geological variables in Weixi area in the Yunnan province are taken out, and the geo-anomaly are circled in the same area.

Keywords: GIS geo-variables extraction geo-anomaly circle

90 年代初, 人类进入信息化社会, 推动地球科学即将发生新一次远比“计量革命”意义更加深远的“信息革命”。特别是 1998 年美国提出“数字化地球”的概念, 地质学的发展从过去的定性描述到定量分析, 从定量分析到数字化, 其中地理信息系统(GIS)起到决定性的作用。GIS 作为一门新兴边缘学科正在世界范围内日益广泛地得到应用, 并正在或已经成为一个独特的行业被用户欢迎。文中引进 GIS, 从信息数字化的角度, 提取地质变量, 圈定地质异常。地质变量是进行矿产资源评价、成矿靶区圈定最基础的信息, 也是进行信息综合、地质背景圈定、各种地质异常划分的前奏, 地质异常是矿床预测实现科学找矿的重要课题和成矿预测的基础(赵鹏大)。

前人工手操作提取地质变量, 不仅费时费力, 而且精度差, 准确度低, 有时无法进行。陕西某地质队对某预测区手工提取地质变量, 8 人工作了 6 个月, 才完成整幅图近 70 个变量取值

(胡光道, 1999 年), 但 GIS 方法提取地质变量只需近 2 小时完成 341 个变量 * 1350 个网格变量, 并提高了准确度。后来前人又从专家系统的角度, 提取地质变量, 虽取得进步, 但提取地质信息不完全、自动化低, 特别是在当时条件下, 难以把整个系统捆绑形成整体, 难以从项目变成产品推向市场发挥经济效益。在国家“九五”重点攻关课题(96-914-05)和国家重点基础研究规划 973 项目(G1999043605)资助下, 以中文 Win95、Visual C++ 为平台, 引入 GIS 方法, 开发整个系统, 该系统克服上述不足, 下面介绍它的核心思想。

1 网格单元划分

为了提取研究区地质变量, 首先对其网格单元划分。它是进行地质研究(如矿产资源评价、成矿靶区圈定、地质信息提取)最基础的环节。其目的是为了确定地质变量观察尺度和取值

范围,提高评价结果的准确性。如何确定最佳网格单元大小,并非易事,它必须结合实际资料水平和采用的评价模型,选择合理的单元划分方法。目前有规则网格单元划分和不规则单元划分两种方法。

1.1 规则网格单元划分

在许多地质单元研究中,通过样品的观测结果来描述总体和确定远景区,并遵循抽样的随机性及样品的代表性原则,故通过选择一定大小的网格将整个研究区划分面积相等、形状相同的单元,用作统一观测和取值范围的基本单元;同时也是进行成矿远景计算和预测评价的基本单元,因此预测单元的大小对预测效果有直接的影响,通常对预测单元的划分应该考虑的因素有:预测比例尺和精度要求;预测区地质条件复杂程度;矿点数及空间分布特征;研究区范围大小及保证统计分析所需的单元数;地质特征的空间变异性。

从矿点分布的方差 S^2 与其均值 X 的比例变化可知,单元面积越大,则单元矿点分布模型越接近负二项分布,单元面积越小,则越接近普阿松分布,因此单元大小反映了不同的抽样观测条件。条件不同,则会影响统计分析的结果,单元面积的大小目前尚无明确的划分准则,但常用的经验算法有:

(1) 经验性最佳面积^[1]

$$S=2 \times \text{预测区总面积} / \text{矿点总数}$$

(2) 单元大小能保证当矿点的分布为随机型时,落入单元内的期望矿点数等于或小于实际落入单元矿点数标准差的三倍^[1],即

$$\frac{\delta}{E} = \sqrt{\frac{(1-S)}{nS}}$$

式中 E :落入单元内的矿点数的数学期望; δ :实际落入单元内的矿点数 x 的标准差; S :单元面积(%); N :矿点总数

若 $\frac{\delta}{E} = \frac{1}{3}$ 则 $S = \frac{9}{(0+n)100}$

(3) 相应比例尺单元大小参考数据区间,而根据相应比例尺的地质图用 1 到 4 平方公里的面积为基本单元的大小,如对于 1:5 万地质图单元大小为 0.25~1 平方公里的面积比较适宜^[1]。

(4) 智能单元面积

根据 (1)(2)(3) 中的因素和经验公式,系统将以上专家知识经验形式化、具体化,采用对话框提问方式,通过对预测区基本地质特征的询问,如预测区长度范围,地质图面上的矿点数目,比例尺信息,推理计算出最佳单元面积,然后转换成相应的网格图形迭加于地质图上。

(5) 用户单元面积

主要根据用户直接提供的单元大小参数,进行屏幕图形的网格单元确定。单元大小及网格数的多少根据用户的经验和知识随意缩放、旋转,直到用户认为满意为止。

1.2 不规则单元划分

前述几种单元划分都是以满足抽样的随机性或代表性为前提,而进行的等大小规则单元划分,其优点是在统一观察变量的前提下,可以把众多的地质变量所包含的矿产资源信息最大限度地反映出来,有利于矿与非矿地质特征的判断,显然此法给成矿预测带来了方便,但它也破坏了地质规则的变化性,如二个等大的单元中提出花岗岩出露面积相同,但其中一个单元的花岗岩是由后期构造作用的外来推覆体,而另一个单元的花岗岩则是真正的侵入体,因此两个单元的地质意义与控矿意义相差极大。再者,在众多的等大单元中,通常研究程度与观察

尺度水平不尽一致,因此限制了成矿作用的客观分析和方法的应用,而按不规则单元为单元进行地质研究,相对来说是一种按地质变化的实际和地质数据变化的性质出发合理的一种新的预测途径,但另一方面它又是一种代价比较高的预测,且要求数据水平高。

实际经验表明,一般以地层边界和岩体边界作为不规则单元边界反映地质情况的实际,因此,不规则单元划分模块中有地层边界和岩体边界两种划分方法。

2 GIS 提取地质变量的原理

一般来说,大多数 GIS 把空间数据分为六大类,即点类数据、线类数据、区类数据、网格类数据、栅格类数据和图象类数据,但地质学上常使用前三类数据(即点类、线类和面类数据)。由于地质数据的特殊性,即使同一类数据,不同的空间实体具有不同的属性,为了节约存储空间、提高查询速度和管理的方便,将它们分为不同的子类,因此,地学中面类数据如地质图上的地层、岩体、脉岩、围岩蚀变、混合岩化、变质相带及特殊岩层等子类数据;点类数据分为矿点、火山构造、钻孔、各类测试数据采样点如化石同位素年龄采样点、泉水及震中等子类数据;线类数据分为断裂、褶皱轴、成矿远景区划等子类数据。经 GIS 表述的上述文件,与网格文件进行空间叠加操作,求得每个网格单元地质变量。下面以规则网格与线文件(如断裂)叠加交操作为例,说明提取线性变量的原理^[2-8]:

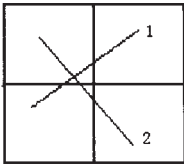


图 1 断裂分布图

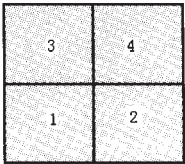


图 2 规则网格图

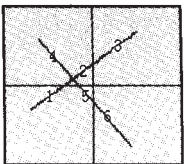


图 3 叠加图

表 1 图 1 的属性数据表

ID	长度 (cm)	名称	性质	倾向	走向
1	3.00	A 断裂	正断裂	NW	NE
2	3.05	B 断裂	逆冲断裂	SW	SE

表 2 图 2 的属性数据表

ID	周长 (cm)	面积 (cm ²)	单元号
1	6.90	3.06	1
2	6.90	3.06	2
3	6.90	3.06	3
4	6.90	3.06	4

表 3 图 3 的属性数据表

ID	长度 (cm)	名称	性质	走向	倾向	单元号
1	0.90	A 断裂	正断裂	NE	NW	1
2	0.82	A 断裂	正断裂	NE	NW	3
3	1.28	A 断裂	正断裂	NE	NW	4
4	1.20	B 断裂	逆冲断裂	SE	SW	3
5	0.50	B 断裂	逆冲断裂	SE	SW	1
6	1.35	B 断裂	逆冲断裂	SE	SW	2

其中图 1 是线类数据文件(以断裂为例)的分布图,图中数字为表 1 的 ID 编号,表 1 记录图 1 的属性数据,每条断裂有六个属性,分别为实体 ID 编号、长度、名称、走向、倾向、性质。图

2 是规则网格图 ,图中数字是表 2 的 ID 号 ,表 2 记录图 2 的属性数据 ,每个实体有四个属性 ,分别为 ID 号、面积、周长、单元号。当图 1 与图 2 作空间叠加交操作后 ,得到图 3 ,图中数字是表 3 的 ID 号 ,计算机根据规则网格和线类文件的空间数据 ,及表 1 和表 2 的属性数据计算出表 3 的数据 ,它记录图 3 的属性数据 ,每个实体有七个属性 ,分别为 ID 号、长度、名称、走向、倾向、断裂性质、单元号。以这些基本变量为基础 ,求得其它变量或综合变量 ,如网格单元中的断裂条数、断裂的优益度、中心对称度等。

点类数据、区类数据与规则网格叠加操作求每个网格点变量、面变量的原理与求线性变量几乎类似 ,不在详述。

根据以上原理 ,借助 GIS 二次开发函数库 ,在中文 Win95、Visual C++平台上 ,开发出地质变量提取子模块 ,它是国家“九五”科技攻关项目“金属矿产资源分析评价系统”中的核心模块 ,完成了地质变量 (包括综合变量)全自动实现 (原程序略^[9-11])。

3 隐含信息挖掘

通过第 2 节得到了各网格中最基本的面、线、点变量信息 ,利用已有变量信息进一步挖掘更深层次的信息 ,进行变量综合。利用网格中各面变量 (地层、岩体、脉体等)的面积百分数及研究区面变量总数 ,求出各单元的相对熵^[1] ,其计算公式为 :

$$100Hr=\frac{-100\sum_{i=1}^Np_i\ln p_i}{Hm}$$

式中 p_i —— N 元系统中第 i 成分所占百分比 ($i=1,2,\dots,N$)
 Hm ——最大可能熵 ,又称最大不确定性值

$$Hm=-\sum\frac{1}{N}\ln\frac{1}{N}=-N(\frac{1}{N}\ln\frac{1}{N})=\ln N$$

N ——所研究的 N 元系统

利用网格中性线构造 (如断裂)的长度、方向及各线性体所属方位角区间的控矿权重 (可人工赋值、可计算机赋值) ,求出网格单元的断裂优益度和中心对称度 ,计算断裂优益度的公式^[12]如下 :

$$\varepsilon=l_iw_i(1+\frac{|\sin\theta|}{2})+\sum_{i=2}^n l_i[1+(\frac{|\sin(\alpha_i-\alpha_{i-1})|}{2})]w_i$$

式中 i 是线性构造序号 α_i 是第 i 条线性构造的方位角 θ 是第一线性构造与最后一条线性构造的夹角 l_i 为第 i 条线性体在该网格单元的长度 w_i 是第 i 条线性体所属方位区间的线性构造控矿权重 ,可通过已知样本统计求出。将 ε 分别按 α_1

$\alpha_2 \dots \alpha_n$ 及 $\alpha_1 \alpha_2 \dots \alpha_n$ 分别计算出 ε_1 和 ε_2 值 , $\frac{\varepsilon_1+\varepsilon_2}{2}$ 则为该统计网格单元的断裂优益度值。该法实质是利用线性构造的夹角作为加权值求取线性构造密度。

计算断裂中心对称度的公式^[12]如下 :

$$\sigma_3=\left[\frac{\sum_{i=1}^n l_i\sin^2(\theta_i-\bar{\theta})}{\sum_{i=1}^n l_i}\right]^{\frac{1}{2}}$$

$$\bar{\theta}=\frac{\sum_{i=1}^n l_i\theta_i}{\sum_{i=1}^n l_i}$$

式中 $\bar{\theta}$ 为平均方位 ,并与各线性体成锐夹角 ;

θ_i 是某一网格第 i 条线性体在地理坐标上半空间的方位角 $-90^\circ\leq\theta_i\leq90^\circ$ l_i 为某一网格第 i 条线性体的长度。

规则网格与点型实体 (矿点或矿化点)叠加后 ,得到网格单元点变量 (矿点或矿化点)的类型、规模、个数等信息。统计出预测区各已知矿种各规模之间的当量关系、价值关系 ,计算各单元的矿床当量和矿床价值等综合变量。

将几个变量综合 ,构成有特定意义的新变量。这种综合不是被综合信息的简单反映 ,也不是它们的孤立集合 ,而是具有明确意义的综合概念和结果特征 ,常能使隐含信息显性化。用“相对熵”的定量指标来考察围岩蚀变组合特征与矿化的关系 ,这对于与围岩蚀变有关的矿化预测具有指导意义 ;用断裂优益度变量评价在以构造为主要控矿因素的内生矿产中每个网格单元是否有利储矿 ;用中心对称度来研究和圈定古火山机构、小型等轴状隐伏侵入体等具有放射状断裂体系的环形构造。

4 地质异常的圈定

地质异常理论是赵鹏大教授近年提出的新理论 ,以往地质人员多论及物化探异常在成矿中的作用 ,而忽略了地质异常的研究和分析 ,研究地质异常是实现科学找矿的重要课题 ,是成矿预测的基础 ,并提出圈定成矿可能地段、找矿可行地段和找矿有利地段的任务。下面介绍有先验模型的地质异常提取方法。

有先验模型的地质异常提取 :其基本思想是通过模型单元的控矿因素、找矿标志和地质特征等地质变量权系数的计算 ,进而建立数学模型 ,以表达模型单元地质条件与矿化有利度的关系 ,将其外推到预测区 ,从而达到评价预测单元矿化有利度的目的。其建模方法很多 ,并根据地质变量数据水平的不同分为四类。一是对数据水平要求较高 ,地质变量为连续的定量变量 ,如判别分析和回归分析等 ;第二类为混合数据的地质变量 ,其地质变量取值为连续型和二态离散型的混合数据 ,相应的计算方法有事件概率回归、数量化理论等 ;第三类为纯二态变量的地质变量 ,如特征分析、逻辑信息法、勃尔法和相关频数比值法等 ;第四类为根据专家经验观察到的地质变量对成矿的有利性 ,主观给定权系数 ,然后建立成矿有利度与地质变量之间函数关系式 ,如成矿建造分析 ,主观给权预测等。

由于地质情况的复杂性、变量多样性。地学数据是典型的多源空间数据 ,它们的量纲不一、形式多样 ;既有定量数据、又有定性文字描述数据 ,选择不同的变量组合、数学模型及定性变量的赋值应模拟出与之对应的结果。程序开发中 ,按以下过程进行 :模型单元的选择、GIS 提取地质变量、二级变量筛选 (计算机筛选即利用 GIS 分析已知矿产与其产出的地质环境关系 ,自动挑选。人工筛选即地质专家取舍变量)、定性变量赋值 (计算机自动赋值、人工赋值、计算机和用户混合赋值)、数据变换及预处理、数学模型及方法的选择、地质异常与非地质异常的临界值的确定、结果的模拟显示和最后结果存盘 ,圈定和评价最能反映实际情况的地质异常。地质异常分级的依据 :① 通过模型计算 Y 的大小 ;② 有利的地层、岩体部位 ;③ 有利的构造部位 ;④ 已知有矿单元。

5 实例

以云南省维西地区为实例 ,该区位于云南省西北部 ,地理
 计算机工程与应用 2003.14 83

坐标为东经 99°~100° ,北纬 26°40'~27°20' ,比例尺 1 20 万。气候寒冷 ,山高路险 ,自然条件艰苦。受交通、气候、地理位置影响 ,该区地质工作程度低 ,研究程度不高。利用上述方法原理 ,提取该区地质变量 ,圈定其地质异常 ,给后人进一步工作以参考。原始资料来自云南省地矿厅 ,包括地、物、化、遥等及 1 20 万各图件 ,根据国家、部门、行业标准进行分层 ,建立相关空间数据库 ,完成信息数字化输入。

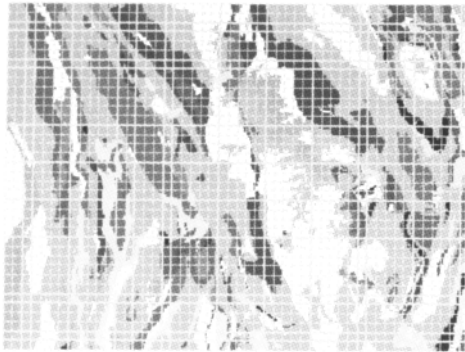


图 4 云南维西预测区基本单元为 4 平方公里规则
网格叠加于地层图上

网格单元大小确定。分析研究区影响单元大小的因素 ,比较各种经验算法的面积 ,得到最佳基本网格单元大小是 4 平方公里 ,图 4 是采用 4 平方公里为一个基本网格单元大小 ,对云南省维西地区进行规则网格单元划分 ,共 1850 个 ,网格单元号从左到右从下到上依次为 1 2 3..... ,并叠加到预测区的地层文件之上 ,该图是由 1 20 万的原图压缩而成 ,因压缩时未知比例尺 ,故图中没有标出比例尺 (下同) ,各时代地层的属性 (如岩性、时代等) 都在属性数据库中 ,因此图中无图例。

地质变量提取。根据 GIS 提取地质变量的原理 ,提取了地层、岩体、脉体、断裂和矿点或矿化点变量 ,在此基础上 ,依据 3 求得综合变量。限于篇幅 ,取 1850 个单元中前 5 个单元的地层变量 ,表 4 记录前 5 个单元中地层变量的一些信息 (组合熵、地层种类数、岩性、时代、面积、面积百分数) 。表 5 记录前 36 个单元中岩体变量的某些信息 (包含岩体名称、岩体种类数、时代、面积、面积百分数、周长) ,表中只记录有岩体的单元。表 6 列出前 23 个网格中的断裂变量的某些信息 ,无断裂通过的单元表中没有列出。其它变量表未列出。虽共提取了 341 个变量 ,表中只显示少数变量。

表 4 前 5 个单元中地层变量表

网格单元号	组合熵	地层种类数	岩性	时代	面积	面积百分数
1	0.00	1	攀天阁组	T3P	100.556	1.0000
2	0.00	1	攀天阁组	T3P	100.556	1.0000
3	13.8	2	攀天阁组	T3P	25.201	0.2506
3	13.8	2	崖依比组地一段、攀天阁组	T3C1-T3P	75.355	0.7494
4	16.00	1	崖依比组地一段、攀天阁组	T3C1-T3P	100.556	1.000
5	17.23	2	崖依比组地一段、攀天阁组	T3C1-T3P	64.234	0.6388
5	17.23	2	花开左组第一段	J2H1	36.322	0.3612

地质异常的圈定。根据圈定地质异常过程 ,采用 4 平方公里作为一个基本网格单元大小 ,利用 GIS 方法提取研究区 341 个变量 (包括综合变量) ,通过变量筛选 ,过滤出 19 个与该区矿床相关的变量 ,分别为 :面变量中的单元地层种类数、单元组合

熵、石中山组第三段、石中山组第二段、景星组第一段、上兰组、花开左组第二段和花开左组第一段 ;线变量中的单元断裂条数、断裂中心对称度、通甸断裂、阿则里—松登—车邑坪—洛脑西—喀呼儒—咪里断裂、阿多底西—史普利—里袜—油洛断裂、日拉达断裂、富隆厂—打主保—安乐街断裂、嘎嘎塘—阿木花独—驮儿罗—塘上村—腊八底—五点石断裂和维西—雪龙山断裂 ;点变量中的单元矿点数、单元矿石品位。建立研究区的多元回归地质异常圈定模型 ,通过作图法厘定地质异常与非地质异常的临界值是 32.12 ,圈定地质异常色块图如图 5。已发现的矿体 (包括矿化点) 中 ,有 76.84% 矿体 (含矿化点) 落入异常区内 ,其中中型矿全部落入一级异常区中 ;小型矿中 ,68.21% 落入一级异常区 ,12.62% 落入二级异常区 ,8.42% 落入三级异常区 ,10.75% 落入背景区 ;矿点中 ,71.23% 落入二级异常区 ,28.77% 落入背景区 ;矿化中 ,60.15% 落入二级异常区 ,39.85% 落入背景区 ,本区无大型及特大型矿。

表 5 某些单元中岩体变量表

网格单元号	面积	周长	岩体名称	年代	岩体种类数	面积百分数
32	0.089	6.756	麻栗坪、弯路子	τ6	1	0.0009
33	19.562	18.296	麻栗坪、弯路子	τ6	1	0.1945
34	100.557	55.870	喇叭山、老军山	ζπ6-ζ6	1	1.000
35	99.651	91.187	喇叭山、老军山	ζπ6-ζ6	1	0.9982
36	81.892	61.189	喇叭山、老军山	ζπ6-ζ6	1	0.8144

表 6 前 23 个有断裂信息单元的断裂线变量表

网格单元号	长度	断裂条数	断裂交点数	断裂优益度	中心对称度	断裂方向 (度)	断裂名称
5	10.500	1	0	0.00	0.00	64	周乌场
9	9.836	1	0	4.42	0.00	96	阿则里、松登、车邑坪、洛脑西、喀呼儒、咪里
13	10.998	1	0	0.00	0.00	65	富隆厂、打主保、安乐街
14	9.983	2	0	0.59	0.19	88	川阱
14	0.354	2	0	0.59	0.19	343	马拉壳
23	10.081	1	0	0.00	0.00	83	日拉达

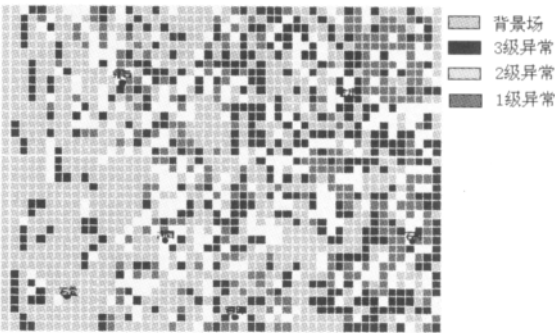


图 5 云南省维西地区地质异常分区图

6 结论

(1) 该文从信息数字化角度 ,首次利用 GIS 方法提取地质变量 ,该方法与前人工方法相比 ,节约人力、物力、财力 ,提高速度、数据精度和准确度 ;与前人半自动、不完整信息提取的计算机方法比 ,实现了自动化、完整信息提取。

(2) 完成简单变量提取后 ,引入 3 个综合变量 ,挖掘隐含信息。用“相对熵”的定量指标来考察围岩蚀变组合特征与矿化的关系 ,这对于与围岩蚀变有关的矿化的预测具有指导意义 ;用断裂优益度变量评价在以构造为主要控矿因素的内生矿产中

每个网格单元是否有利储矿,用中心对称度来研究和圈定古火山机构、小型等轴状隐伏侵入体等具有放射状断裂体系的环形构造。

(3)取舍地质变量,建立圈定地质异常模型,实施工作区地质异常的划分,该异常与该区已知矿产吻合得很好,符合客观实际,效果佳,可作为该区后续矿产资源评价预测的基础。

(4)从变量提取、地质异常圈定,贯穿信息量化这一主线,特别是地质异常的量化,与前人的定性地质异常描述相比,无疑又前进一步。(收稿日期 2002 年 7 月)

参考文献

1.赵鹏大,胡旺亮,李紫金.矿床统计预测[M].第二版,北京:地质出版社,1994-05
2.朱光,季晓燕,戎只.地理信息系统基本原理及应用[M].北京:测绘出版社,1997
3.陈俊,宫鹏.实用地理信息系统—成功地理信息系统的建设与管理[M].北京:科学出版社,1998

4.J E McCormack J Hogg.Virtual-memory tiling for spatial data handling in GIS[J].computer &Geosciences ,1997 23 (1):119~126
5.Henry T.The Spatial Data Transfer Standard[J].Cartograpy and Geographic Information System ,1994 21 (3)
6.Bonham-Carter G F.Geographic Information System for GeoScientists :Modelling with GIS[M].Pergamon ,Garry-town ,New York ,1994
7.Wilson J D.GIS Goes Global-Technology Builds Management Resources[J].GIS world ,1997 ;(10)
8.Strand E.GIS Plays a Role in Business Processing Reengineering[J].GIS World ,1995 ;(4)
9.David J Kruglinski.Inside Visual C++[M].Tsinghua University Press ,1996
10.Richard C Leinecker Jamie Nye.Visual C++ toolkit[M].Machine Industry Press ,1997
11.Nigel Tompson.Animation techniques in Win32[M].Tsinghua University Press ,1996
12.赵不亿,秦小光.遥感构造的定量分析方法[J].地质科技情报,1988 7 (1):127~135

(上接 74 页)

A₂ A₃ A₄ A₅ 为条件属性 A₅ 为决策属性。

表 1

	A ₁	A ₂	A ₃	A ₄	A ₅
1	0.171 0.277 0.500	0.203 1.899 1.990	0.082 1.830 1.990	N	0.082 2.830 3.990
2	0.105 0.253 0.448	0.203 1.899 1.990	0.082 1.830 1.990	N	0.112 2.221 3.225
3	0.105 0.253 0.448	0.203 1.899 1.990	0.082 1.830 1.990	N	0.082 2.830 3.990
4	0.095 0.193 0.379	0.526 1.233 2.012	0.082 1.830 1.990	N	0.112 2.221 3.225
5	0.105 0.253 0.448	0.526 1.233 2.012	0.350 2.211 3.445	N	0.112 2.221 3.225
6	0.171 0.277 0.500	0.526 1.233 2.012	0.350 2.211 3.445	N	0.082 2.830 3.990
7	0.105 0.253 0.448	0.526 1.233 2.012	0.082 1.830 1.990	N	0.082 2.830 3.990
8	0.171 0.277 0.500	0.203 1.899 1.990	0.082 1.830 1.990	P	0.082 2.830 3.990
9	0.105 0.253 0.448	0.203 1.899 1.990	0.082 1.830 1.990	P	0.082 2.830 3.990
10	0.095 0.193 0.379	0.526 1.233 2.012	0.082 1.830 1.990	P	0.112 2.221 3.225
11	0.105 0.253 0.448	0.526 1.233 2.012	0.350 2.211 3.445	P	0.082 2.830 3.990
12	0.171 0.277 0.500	0.526 1.233 2.012	0.350 2.211 3.445	P	0.082 2.830 3.990

因为表 1 是一个复合决策表,所以利用前文介绍的优势-等价关系来划分论域 U={1 2 ... 12}得到优势-等价类及逆优势-等价类如下:

$R_p^+ : \{1\}\{2\}\{3\}\{1\ 2\}\{1\ 2\ 3\}\{1\ 2\ 3\ 4\ 5\ 6\ 7\}\{5\ 6\}\{6\}\{1\ 2\ 7\}\{8\}\{8\ 9\}\{8\ 9\ 10\ 11\ 12\}\{11\ 12\}\{12\}.$

$R_p^- : \{1\ 2\ 3\ 4\ 7\}\{2\ 3\ 4\ 7\}\{2\ 3\ 4\ 7\}\{4\}\{5\ 7\}\{4\ 5\ 6\ 7\}\{4\ 7\}\{8\ 9\ 10\}\{9\ 10\}\{10\}\{10\ 11\}\{11\ 12\}.$

因为只有两个决策类,所以 $Cl_1^{\leq}=Cl_1=\{2\ 4\ 5\ 10\}$ $Cl_2^{\geq}=Cl_2=\{1\ 3\ 6\ 7\ 8\ 9\ 11\ 12\}$ 根据 Cl_i^{\leq} Cl_i^{\geq} 的上下近似定义有:

$\underline{C}Cl_1^{\leq}=\{4\ 10\}$ $\overline{C}Cl_1^{\leq}=\{2\ 3\ 4\ 5\ 7\ 10\}$ $Bn(Cl_1^{\leq})=\{2\ 3\ 5\ 7\}$
 $\underline{C}Cl_2^{\geq}=\{1\ 6\ 8\ 9\ 11\ 12\}$ $\overline{C}Cl_2^{\geq}=\{1\ 2\ 3\ 5\ 6\ 7\ 8\ 9\ 11\ 12\}$
 $Bn(Cl_2^{\geq})=\{2\ 3\ 5\ 7\}$

因此,根据分类质量的定义有 $\gamma=0.67$ 。
根据约简的定义可知,表 1 只有一个约简{A₁ A₄}。
所以,可提取决策表 1 的决策规则如下:

rule1 :如果 $f(A_1) \geq 0.171\ 0.277\ 0.500$,那么 $X \in Cl_2^{\geq}$ 。
rule2 :如果 $f(A_1) \geq 0.1050\ 0.253\ 0.448$),而且 $f(A_4) = P$,那么 $X \in Cl_2^{\geq}$ 。
rule3 :如果 $f(A_1) \leq 0.095\ 0.193\ 0.379$) ,那么 $X \in Cl_1^{\leq}$ 。
rule4 :如果 $f(A_1) = 0.105\ 0.253\ 0.448$),而且 $f(A_4) = N$,那么 $X \in Cl_1^{\leq}$ 或 $X \in Cl_2^{\geq}$ 。

6 结语

信息不完备是复杂决策环境中不可避免的问题^[5]。因为不完备决策表中的不确定成分太多,因此从不完备决策表中挖掘出潜在的有用知识,增加了数据挖掘的难度。传统的粗糙集理论是处理不确定信息的有效方法,但决策表中数据本身的不确定性在方法中不能得到有效处理。论文利用基于优势-等价关系的扩展粗糙集模型,结合模糊集理论知识,给出了有效的模糊决策表的数据挖掘方法。算例证明了该方法的有效性。
(收稿日期 2003 年 1 月)

参考文献

1.Slowinski R Stefanowski J.Handling Various Type of Uncertainty in the Rough Set Approach.In :Rough Sets ,Fuzzy Sets and Knowledge Discovery London Springer-Verlag ,1994 366~376
2.Greco S Matarazzo B Slowinski R.Rough Approximation of a Preference relation by dominance relations[J].Eropean Journal of Operational Research 117 ,1999 63~83
3.Salvatore Greco ,Benedetto Matarazzo ,Roman Slowinski.Rough Approximation by Dominance Relations[J].International Journal of Intelligent Systems 2002 ;17 :153~171
4.徐泽水.三角模糊数互补判断矩阵的一种排序方法[J].模糊系统与数学 2002 ;16 (1):47~50
5.赵卫东,李旗军,盛昭翰.区间值属性不完全信息下的数据挖掘[J].系统工程理论与实践 2001 ;10 (2):35~38
<http://www.cnki.net>