

## 粗糙集属性约简判别分析方法及其应用

刘宏杰<sup>1,2</sup>, 冯博琴<sup>1</sup>, 李文捷<sup>2</sup>, 吕焕通<sup>3</sup>

(1. 西安交通大学电子与信息工程学院, 710049, 西安; 2. 新疆油田公司研究院地球物理研究所, 830013, 乌鲁木齐; 3. 新疆油田公司, 834000, 克拉玛依)

**摘要:** 为了解决统计逐步判别分析法存在的问题, 提出了一种基于粗糙集属性约简的统计判别分析方法. 首先采用粗糙集属性约简进行变量筛选, 这样可充分利用粗糙集属性约简不需要属性分布的先验信息这一特点, 再对所选择的变量进行 Bayes 判别分析训练, 建立判别函数或相应的后验概率函数, 以解决选择变量过程中存储量较大且检验变量的重要性总体服从正态分布这一主观性假设等问题. 通过对油气储层数据的实际分析表明, 所提方法不仅易于实施, 而且检验数据集的判别准确率高于统计逐步判别分析法, 同时可节省预测成本, 提高预测速度.

**关键词:** 粗糙集; 属性约简; 判别分析; 储层预测

**中图分类号:** TP302 **文献标识码:** A **文章编号:** 0253-987X(2007)08-0939-05

## Discrimination Method of Rough Set Attribute Reduction and Its Applications

Liu Hongjie<sup>1,2</sup>, Feng Boqin<sup>1</sup>, Li Wenjie<sup>2</sup>, Lü Huantong<sup>3</sup>

(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 2. Geophysical Research Institute, Research Institute of Exploration and Development, PetroChina Xinjiang Oilfield Company, Urumqi 830013, China; 3. PetroChina Xinjiang Oilfield Company, Karamay 834000, China)

**Abstract:** In order to overcome some shortcomings of the statistical stepwise discrimination method, a new discrimination method based on rough set attribute reduction is proposed, in which the attribute reduction of rough set is used to select important discriminating variables by sufficiently taking advantage of the feature that the attribute reduction of rough set needs no prior information on the distribution. Then, with the selected variables, Bayesian discrimination procedure is used to build the discrimination function or compute the posterior probability so as to solve the problems that the storage size is much bigger and the importance of discriminating variables wholly obey the normal distribution from the subjective assumption in the statistical stepwise discrimination method. Through applying the proposed method to the real-world dataset of reservoir of oil and gas, it is demonstrated that the proposed method not only is easy to be implemented with higher correct discrimination rate, but also can decrease the computational cost and increase the prediction speed.

**Keywords:** rough set; attribute deduction; discrimination analysis; reservoir prediction

统计判别分析<sup>[1-2]</sup>是多元分析方法中较为成熟的一类方法, 目前已在社会科学和自然科学的各个领域获得了广泛的应用. 在统计判别分析中, 并非是指标变量越多, 判别分析的效果就越好, 冗余的指标变量反而会降低分类的准确率. 统计判别分析主要

是利用逐步判别法选择变量, 以 Wilks 的  $\Lambda$  统计量作为各指标分类效果的重要性度量, 但该方法要检验的变量重要性需假定变量总体服从正态分布<sup>[1]</sup>, 其结果有一定的主观性, 而且存储量比较大. 目前, 粗糙集理论越来越多地运用到了数据挖掘之中<sup>[3]</sup>,

粗糙集的属性约简<sup>[3-5]</sup>不需要属性分布的任何先验信息,因此其应用结果更为客观.虽然利用粗糙集理论也可以建立判别准则,但目前的研究仍集中于离散属性变量的情况,对于连续属性变量的情形尚缺乏有效的方法.

考虑到粗糙集属性约简的客观性、有效性,尤其是分类变量是连续的情形,本文将粗糙集属性约简方法与统计判别分析方法相结合,提出了基于粗糙集属性约简的判别分析方法,即先基于粗糙集的属性约简进行变量选择,再利用统计 Bayes 判别分析方法建立判别准则.

## 1 方法的原理与模型

### 1.1 粗糙集属性约简方法

粗糙集的属性约简主要是为了求得最小的属性集,这样在保证属性集满足一定分类精度的情况下,使其属性个数最小.粗糙集的基本概念和属性约简理论<sup>[4,6]</sup>描述如下.

在粗糙集理论中,称  $S = \langle U, A, F, D, G \rangle$  为目标信息系统或决策表,其中:  $U = \{x_1, x_2, \dots, x_n\}$  为对象集,  $U$  中的每个  $x_i (i \leq n)$  称为一个对象;  $A = \{a_1, a_2, \dots, a_m\}$  为属性集,  $A$  中的每个  $a_j (j \leq m)$  称为一个属性;  $F = \{f_j: j \leq m\}$  为  $U$  和  $A$  的关系集,  $f_j: U \rightarrow V_j (j \leq m)$ ,  $V_j$  为属性  $a_j$  的值域;  $D = \{d_1, d_2, \dots, d_p\}$  为目标属性集或者决策属性集;  $G = \{g_j: j \leq p\}$  为  $U$  和  $D$  的关系集,  $g_j: U \rightarrow V_j' (j \leq p)$ ,  $V_j'$  为属性  $d_j$  的值域.

属性的约简就是在不改变决策表分类能力的情况下使参与分类的属性最少,也就是对属性集做等价有效的简化.

对于决策表  $S = \langle U, A, F, D, G \rangle$ , 若  $R_A \subseteq R_D$ , 则称目标信息系统是协调的, 否则为不协调的, 对于协调的信息系统, 若存在  $B \subseteq A$  使  $R_A \subseteq R_D$ , 且对任意  $b \in B, R_{B-\{b\}} \not\subseteq R_D$ , 则称  $B$  是目标信息系统的一个约简. 其中

$$\begin{aligned} R_A &= \{(x_i, x_j) \in U \times U: f_l(x_i) = f_l(x_j) (\forall a_l \in A)\} \\ R_B &= \{(x_i, x_j) \in U \times U: f_l(x_i) = f_l(x_j) (\forall a_l \in B)\} \\ R_D &= \{(x_i, x_j) \in U \times U: g_l(x_i) = g_l(x_j) (\forall d_l \in D)\} \end{aligned}$$

分别由属性集  $A, B, D$  确定的等价类. 令  $P$  和  $Q$  为  $U$  中等价关系,  $Q$  的  $P$  正域记为  $\Gamma_P(Q)$ , 即

$$\Gamma_P(Q) = \bigcup_{x \in U/Q} PX$$

式中:  $P$  表示下近似, 也就是正域.  $Q$  的  $P$  正域是  $U$  中所有根据分类  $U/P$  的信息可以准确地划分到关系  $Q$  的等价类中去的对象集合. 对于某个属性子集  $B \subseteq A, x, y \in U$ , 设  $b \in B$ , 如果  $R_B = R_{B-\{b\}}$ , 则属性  $b$  是  $B$  中可约简的, 否则是不可约简的. 由  $B$  中所有不可约简属性组成的集合称为  $B$  的核 ( $\text{core}(B)$ ).

若  $B \subseteq A, B$  中所有属性都是不可约简的, 有  $R_B = R_A$ , 则  $B$  是  $A$  的一个约简. 显然, 所有约简集的交集  $\text{core}(A) = \bigcap \text{red}(A)$ . 在实际决策应用中只要求取相对最简约简集  $B$ , 即满足  $\gamma_B(D) = \gamma_A(D)$ , 其中  $D$  为决策属性,  $\gamma_A(D) = |\Gamma_A(D)|/|U|$ .

属性约简算法步骤如下.

(1) 选取完备的条件属性.

(2) 对决策属性进行泛化和连续属性离散化处理<sup>[7]</sup>.

(3) 先计算不可分辨关系<sup>[8]</sup>  $R_A, R_{A-\{a_i\}} (i=1, 2, \dots)$ , 然后求核  $\text{core}(A)$ . 若  $R_{A-\{a_i\}} \neq R_A$ , 则  $a_i$  是不可约简的属性, 由所有不可约简属性构成核.

(4) 求出最简约简集. 设  $B = \text{core}(A)$ , 判断是否满足条件  $\gamma_B(D) = \gamma_A(D)$ , 若成立, 则  $B$  为所求的最简约简集; 否则, 在  $B$  中添加  $A$  中的一个可约简的属性, 再进行判断, 直至满足条件为止.

### 1.2 Bayes 判别分析模型

统计判别分析的基本方法<sup>[1-2,9]</sup>是: 设有  $g$  个总体  $G_1, G_2, \dots, G_g$ , 各总体的协方差矩阵分别为  $\Sigma_1, \Sigma_2, \dots, \Sigma_g$ , 均值向量为  $u_1, u_2, \dots, u_g$ , 对给定的一个新样品  $x$ , 判断它属于哪个总体. 本文采用等误判损失下的 Bayes 判别法.

(1) 给出先验概率. 设有  $g$  个总体  $G_1, G_2, \dots, G_g$ , 各总体先验概率为  $q_1, q_2, \dots, q_g$ . 最常见的先验概率确定方法有 2 种: ①若训练样本是在自然情况下观测取得的, 则一般指定  $q_i$  为  $G_i$  的训练样本容

量  $m_i$  和总体样本  $\sum_{i=1}^g m_i$  的比  $\{q = q_i \mid q_i = \frac{m_i}{m} (i = 1, \dots, g), m = \sum_{i=1}^g m_i, \sum_{i=1}^g q_i = 1, q_i > 0\}$ ; ②若无任何先验知识, 则指定先验概率服从均匀分布, 即  $\{q = q_i \mid q_1 = q_2 = \dots = q_i = 1/g (i = 1, \dots, g), \sum_{i=1}^g q_i = 1, q_i > 0\}$

(2) 判别方法. ①若  $\Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$ , 则广义平方距离函数  $D_j^2(x) = D^2(x, G_j) = (x - u_j)^T \Sigma^{-1} (x - u_j) - 2 \ln q_j$ , 其中  $j = 1, 2, \dots, g$ ; ②若  $\Sigma_1, \Sigma_2, \dots, \Sigma_g$  不全相等, 则广义平方距离函数

$D_j^2(\mathbf{x}) = D^2(\mathbf{x}, G_j) = (\mathbf{x} - \mathbf{u}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \mathbf{u}_j) + \ln |\boldsymbol{\Sigma}_j| - 2 \ln q_j$ , 其中  $j = 1, 2, \dots, g$ .

(3) 确定后验概率. 样品  $\mathbf{x}$  属于  $G_j$  的后验概率为

$$P(G_j | \mathbf{x}) = \frac{\exp(-0.5 D_j^2(\mathbf{x}))}{\sum_{i=1}^g \exp(-0.5 D_i^2(\mathbf{x}))}$$

式中:  $D_i^2(\mathbf{x})$  是  $\mathbf{x}$  到第  $i$  个总体的广义平方距离. 采用后验概率的判别准则为: 当  $P(G_j | \mathbf{x}) > P(G_i | \mathbf{x})$  ( $i=1, \dots, g, i \neq j$ ), 则判  $\mathbf{x} \in G_j$ .

在实际应用中, 通常已知的是来自各总体的训练样本  $\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_{m_j}^{(j)}$  ( $j=1, \dots, g$ ), 这时可利用训练样本估计  $\boldsymbol{\Sigma}_j$  和  $\mathbf{u}_j$  ( $j=1, \dots, g$ ), 即

$$\hat{\mathbf{u}}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} \mathbf{x}_i$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{m_j - 1} \sum_{i=1}^{m_j} (\mathbf{x}_i^{(j)} - \hat{\mathbf{u}}_j)(\mathbf{x}_i^{(j)} - \hat{\mathbf{u}}_j)^T$$

当  $\boldsymbol{\Sigma}_j$  ( $j=1, \dots, g$ ) 相等时, 利用  $\hat{\boldsymbol{\Sigma}}_j$  联合估计  $\boldsymbol{\Sigma}$ , 即

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m - g} \sum_{j=1}^g (m_j - 1) \hat{\boldsymbol{\Sigma}}_j$$

## 2 在油气储层预测中的应用

油气储层预测是利用地震数据进行油气储层预测的新方法, 目前主要有单属性油气预测方法、模糊模式识别油气预测方法、BP 神经网络模式识别油气预测方法<sup>[10-11]</sup>等, 但这些方法都存在着各自的缺点和问题<sup>[10, 12]</sup>. 本文将基于粗糙集属性约简的判别分析方法应用于油气储层预测中, 具有良好的效果.

### 2.1 油气储层的粗糙集属性约简预测判别分析系统

粗糙集约简的判别分析识别系统如图 1 所示.

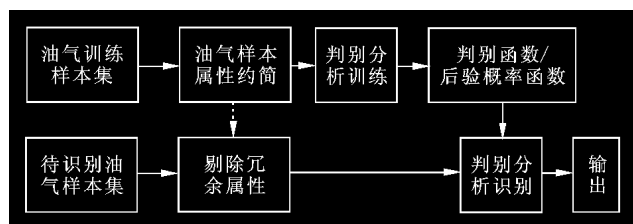


图1 基于粗糙集属性约简的油气储层预测判别分析系统

(1) 油气训练样本: 根据油气属性的物性, 随机选取样本信息.

(2) 油气样本属性约简: 根据粗糙集理论, 连续条件属性必须离散化, 因此先将连续属性离散化, 并将决策属性泛化, 再用粗糙集理论求出训练样本集的一个约简, 以剔除冗余属性.

(3) 判别分析训练: 利用约简的油气属性样本集

进行统计判别分析, 得到判别函数或相应的后验概率函数.

(4) 判别分析识别与输出: 用训练好的判别分析识别对油气属性已约简后的待识别油气样本集进行模式分类, 并输出结果.

### 2.2 实例分析

为了验证基于粗糙集约简的判别分析效果, 选取表 1 所示的实际油气数据进行储层油气预测分析, 其中  $\Phi_1$  为平均振幅,  $\Phi_2$  为能量半衰减时间,  $\Phi_3$  为瞬时频率,  $\Phi_4$  为均方根振幅,  $D$  为决策属性集.

表 1 油气储层数据

| 序号 | $\Phi_1$ /dB | $\Phi_2$ /ms | $\Phi_3$ /Hz | $\Phi_4$ /dB | $D$ |
|----|--------------|--------------|--------------|--------------|-----|
| 1  | 85.405 1     | 54.907 6     | 27.097 6     | 61.244 3     | O   |
| 2  | 66.460 0     | 39.029 4     | 25.213 6     | 53.589 2     | O   |
| 3  | 109.722 8    | 34.789 1     | 39.090 4     | 69.205 6     | O   |
| 4  | 66.849 7     | 65.520 5     | 24.254 8     | 48.198 5     | W   |
| 5  | 55.351 3     | 66.457 4     | 32.518 1     | 39.776 2     | W   |
| 6  | 40.125 6     | 77.971 8     | 33.110 4     | 24.653 5     | W   |
| 7  | 48.622 9     | 35.086 6     | 30.123 8     | 35.450 1     | W   |
| 8  | 60.547 1     | 60.230 7     | 19.382 1     | 50.398 1     | O   |
| 9  | 81.897 1     | 53.461 3     | 33.868 9     | 56.401 2     | O   |
| 10 | 69.187 6     | 31.408 2     | 37.893 5     | 50.989 5     | O   |
| 11 | 117.890 2    | 41.966 3     | 37.774 5     | 77.572 4     | O   |
| 12 | 109.092 8    | 33.212 7     | 37.625 4     | 68.642 3     | O   |
| 13 | 90.425 5     | 63.737 6     | 37.669 1     | 60.236 6     | O   |
| 14 | 40.017 0     | 38.752 2     | 36.087 9     | 27.309 0     | W   |
| 15 | 58.786 6     | 54.340 7     | 32.080 1     | 39.279 3     | W   |
| 16 | 69.185 8     | 60.538 2     | 24.927 2     | 49.489 3     | W   |
| 17 | 55.891 3     | 70.851 4     | 32.951 4     | 39.693 5     | W   |
| 18 | 34.010 6     | 78.158 2     | 27.732 0     | 21.838 6     | W   |
| 19 | 17.667 2     | 15.103 2     | 25.129 8     | 22.594 0     | W   |
| 20 | 54.355 8     | 73.527 7     | 35.410 4     | 38.471 9     | W   |
| 21 | 14.787 7     | 41.220 2     | 50.116 2     | 16.637 5     | W   |
| 22 | 25.023 6     | 52.963 8     | 65.024 3     | 16.400 3     | W   |

注: O 表示油井; W 表示干井.

(1) 油气样本集属性约简. 选取序号 1~7 的样本为训练样本集, 条件属性集为  $C = \{c_1, c_2, c_3, c_4\}$ , 其中  $c_1 \sim c_4$  分别表示  $\Phi_1 \sim \Phi_4$ , 决策属性集  $D = \{d | d \in \{1, 2\}\}$ , 其中 1、2 分别代表 O(油井)和 W(干井). 然后进行区间分割, 得到各连续属性的离散化区间如下

$$c_1: (-\infty, 69.9667), [69.9667, +\infty)$$

$$c_2: (-\infty, 52.0839), [52.0839, +\infty)$$

$$c_3: (-\infty, 30.2345), [30.2345, +\infty)$$

$$c_4: (-\infty, 49.1830), [49.1830, +\infty)$$

这样, 即得到离散化后的样本信息表, 如表 2 所示.

表2 离散化后的样本信息表

| 序号 | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $D$ |
|----|-------|-------|-------|-------|-----|
| 1  | 2     | 2     | 1     | 2     | 1   |
| 2  | 1     | 1     | 1     | 2     | 1   |
| 3  | 2     | 1     | 2     | 2     | 1   |
| 4  | 1     | 2     | 1     | 1     | 2   |
| 5  | 1     | 2     | 2     | 1     | 2   |
| 6  | 1     | 2     | 2     | 1     | 2   |
| 7  | 1     | 1     | 1     | 1     | 2   |

$C$  的不可分辨关系  $\text{ind}(C)$  有下列等价类

$$U/\text{ind}(C) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6\}, \{7\}\}$$

因为

$$U/\text{ind}(C - \{c_1\}) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6\}, \{7\}\}$$

$$U/\text{ind}(C - \{c_2\}) = \{\{1, 3\}, \{2\}, \{4, 7\}, \{5, 6\}\}$$

$$U/\text{ind}(C - \{c_3\}) = \{\{1\}, \{2\}, \{3\}, \{4, 5, 6\}, \{7\}\}$$

$$U/\text{ind}(C - \{c_4\}) = \{\{1\}, \{2, 7\}, \{3\}, \{4\}, \{5, 6\}\}$$

所以

$$U/\text{ind}(C - \{c_1\}) = U/\text{ind}(C)$$

则核为

$$\text{core}(C) = \bigcap \text{red}(C) =$$

$$\{\text{ind}(C - \{c_1\})\} = \{c_2, c_3, c_4\}$$

因为  $U/C = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6\}, \{7\}\}$ , 决策属性的分类  $L = \{y_1, y_2\} = U/\text{ind}(D) = \{\{1, 2, 3\}, \{4, 5, 6, 7\}\}$ , 所以决策属性分类  $L$  关于属性集  $C$  的近似精度为

$$\gamma_C(L) = \frac{|\underline{p}(y_1) + \underline{p}(y_2)|}{|U|} = \frac{(1+1+1) + (1+1+1+1)}{7} = 1$$

式中:  $\underline{p}(y_1)$  和  $\underline{p}(y_2)$  分别表示  $y_1$  和  $y_2$  的正域,  $y_1$  和  $y_2$  分别表示第 1 类和第 2 类。

又因为

$$U/\{\text{core}(C)\} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6\}, \{7\}\}$$

$$L = \{y_1, y_2\} = \{\{1, 2, 3\}, \{4, 5, 6, 7\}\}$$

所以, 决策属性分类  $L$  关于核的近似精度为

$$\gamma_{\text{core}(C)}(L) = \frac{|\underline{p}y_1 + \underline{p}y_2|}{|U|} = \frac{(1+1+1) + (1+1+1+1)}{7} = 1$$

说明核是完备的, 因此得  $\{c_2, c_3, c_4\}$  为最佳约简。

(2) 建立判别准则。根据前面的约简结果, 利用

分类变量  $D$  和数值变量  $(c_2, c_3, c_4)$  进行 Bayes 判别分析。在此假定  $\Sigma_1 = \Sigma_2 = \Sigma$ , 先验概率按比例分配, 即  $q_1 = 0.5, q_2 = 0.5$ , 利用 SAS 系统进行计算并用回代法检验对训练样本的判别效果, 得到如表 3 所示的回判结果, 所有 7 个样品均回判正确。

表3 样品回判结果

| 实际情况 | 回判结果 |     | 总数 |
|------|------|-----|----|
|      | 组 1  | 组 2 |    |
| 组 1  | 3    | 0   | 3  |
| 组 2  | 0    | 4   | 4  |

注: 回判结果表示将实际属于第  $i(i=1, 2)$  组的样品判归为第  $j(j=1, 2)$  组的样品数。

(3) 检验数据集的判别结果及其与其他方法的比较。从表 1 中选取样本号 8~22 的样本作为检验样本集, 采用所建立的判别分析系统进行识别, 其结果如表 4 所示, 表中  $N$  表示样品序号,  $G_1$  表示所属总体,  $P(1|j)$  和  $P(2|j)$  表示将第  $j$  个样品分别判入总体 1 和 2 的后验概率,  $\bar{G}_2$  表示判入总体, 其中  $j=1, 2, \dots, 15$ 。由表 4 可知, 识别结果与实际情况相比, 识别正确率为  $15/15=100\%$ 。

表4 采用粗糙集约简的结果

| $N$ | $G_1$ | $P(1 j)$ | $P(2 j)$ | $\bar{G}_2$ |
|-----|-------|----------|----------|-------------|
| 1   | 1     | 0.557 5  | 0.442 5  | 1           |
| 2   | 1     | 0.869 9  | 0.130 1  | 1           |
| 3   | 1     | 0.775 6  | 0.224 4  | 1           |
| 4   | 1     | 0.999 8  | 0.000 2  | 1           |
| 5   | 1     | 0.997 9  | 0.002 1  | 1           |
| 6   | 1     | 0.920 7  | 0.079 3  | 1           |
| 7   | 2     | 0.003 2  | 0.996 8  | 2           |
| 8   | 2     | 0.050 2  | 0.949 8  | 2           |
| 9   | 2     | 0.459 3  | 0.540 7  | 2           |
| 10  | 2     | 0.027 6  | 0.972 4  | 2           |
| 11  | 2     | 0.000 1  | 0.999 9  | 2           |
| 12  | 2     | 0.003 0  | 0.997 0  | 2           |
| 13  | 2     | 0.016 7  | 0.983 3  | 2           |
| 14  | 2     | 0.000 1  | 0.999 9  | 2           |
| 15  | 2     | 0.000 0  | 1.000 0  | 2           |

为了比较本文所提出的方法的判别效果, 我们仍以 1~7 号样品为训练样本, 分别采用基于全部 4 个分类变量的等先验概率下的 Bayes 判别法和逐步判别法建立判别准则, 并对其余 15 个样品进行判别, 结果如表 5 和表 6 所示。结果表明: 基于全部 4 个分类变量的等先验概率下的 Bayes 判别法将其中 14 号、15 号属于第二类的样品错判为第一类, 其正确识别率为  $13/15=86.67\%$ , 而逐步判别法只选择

了第 4 个变量,将 9 号应为第二类的样品错判为第一类,其正确识别率为  $14/15=93.33\%$ . 实验结果表明,本文提出的基于粗糙集约简的判别分析法在油气储层预测中具有更高的准确性.

表 5 不采用粗糙集约简的结果

| $N$ | $G_1$ | $P(1 j)$ | $P(2 j)$ | $G_2$ |
|-----|-------|----------|----------|-------|
| 1   | 1     | 0.583 8  | 0.416 2  | 1     |
| 2   | 1     | 0.973 4  | 0.026 6  | 1     |
| 3   | 1     | 0.996 1  | 0.003 9  | 1     |
| 4   | 1     | 1.000 0  | 0.000 0  | 1     |
| 5   | 1     | 0.996 3  | 0.003 7  | 1     |
| 6   | 1     | 0.995 7  | 0.004 3  | 1     |
| 7   | 2     | 0.002 6  | 0.997 4  | 2     |
| 8   | 2     | 0.016 8  | 0.983 2  | 2     |
| 9   | 2     | 0.147 6  | 0.852 4  | 2     |
| 10  | 2     | 0.119 8  | 0.880 2  | 2     |
| 11  | 2     | 0.000 0  | 1.000 0  | 2     |
| 12  | 2     | 0.002 5  | 0.997 5  | 2     |
| 13  | 2     | 0.195 6  | 0.804 4  | 2     |
| 14  | 2     | 0.686 2  | 0.313 8  | 1     |
| 15  | 2     | 0.949 0  | 0.051 0  | 1     |

表 6 采用逐步判别分析法的结果

| $N$ | $G_1$ | $P(1 j)$ | $P(2 j)$ | $G_2$ |
|-----|-------|----------|----------|-------|
| 1   | 1     | 0.589 2  | 0.410 8  | 1     |
| 2   | 1     | 0.894 9  | 0.105 1  | 1     |
| 3   | 1     | 0.630 9  | 0.369 1  | 1     |
| 4   | 1     | 0.999 8  | 0.000 2  | 1     |
| 5   | 1     | 0.996 9  | 0.003 1  | 1     |
| 6   | 1     | 0.963 7  | 0.036 3  | 1     |
| 7   | 2     | 0.001 5  | 0.998 5  | 2     |
| 8   | 2     | 0.050 3  | 0.949 7  | 2     |
| 9   | 2     | 0.522 7  | 0.477 3  | 1     |
| 10  | 2     | 0.056 5  | 0.943 5  | 2     |
| 11  | 2     | 0.000 3  | 0.999 7  | 2     |
| 12  | 2     | 0.000 4  | 0.999 6  | 2     |
| 13  | 2     | 0.040 0  | 0.960 0  | 2     |
| 14  | 2     | 0.000 1  | 0.999 9  | 2     |
| 15  | 2     | 0.000 1  | 0.999 9  | 2     |

3 结束语

本文通过粗糙集方法对变量进行约简,提出了基于粗糙集属性约简的判别分析方法,并在油气储层预测中得到了实际应用,进一步说明了本文方法不但简单易行,而且具有很高的正确识别率,同时具

有成本低、处理速度快等优点,从而为油气预测提供了一条有效途径.

参考文献:

[1] 范金城,梅长林. 数据分析[M]. 北京:科学出版社, 2002.

[2] 梅长林,周家良. 实用统计分析方法[M]. 北京:科学出版社,2002.

[3] Lingras P J, Yao Y Y. Data mining using extensions of the rough set model [J]. Journal of the American Society for Information Sciences, 1998, 49(5): 415-422.

[4] 张文修,吴伟志,梁吉业,等. 粗糙集理论与方法[M]. 北京:科学出版社,2003.

[5] Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis [J]. European Journal of Operational Research, 2001,129(1):1-47.

[6] Pawlak Z. Rough sets: theoretical aspects of reasoning about data[M]. Dordrecht, Netherlands: Kluwer Academic Publishers, 1991:1-53.

[7] 夏克文,沈钧毅,李昌彪. 样本信息处理中一种属性约简方法的研究[J]. 西安交通大学学报, 2005,39(6): 558-561.

Xia Kewen, Shen Junyi, Li Changbiao. Study on attribute reduction method in sample information processing [J]. Journal of Xi'an Jiaotong University, 2005,39(6): 558-561.

[8] 史忠植. 知识发现[M]. 2 版. 北京:清华大学出版社, 2004.

[9] 高惠璇. 应用多元统计分析[M]. 北京:北京大学出版社, 2005.

[10] 陈遵德. 储层地震属性优化方法[M]. 北京:石油工业出版社, 1998.

[11] 郭同翠,刘明新,郑晓东,等. 地震数据识别孔隙流体和孔隙度[J]. 天然气勘探与开发, 2004,27(3):18-20.

Guo Tongcui, Liu Mingxin, Zhang Xiaodong, et al. Recognition of pore-fluid and porosity by seismic data [J]. Natural Gas Exploration and Development, 2004, 27(3):18-20.

[12] 陈遵德,陈富贵. 地震数据油气预测中的属性优化方法[J]. 石油物探, 1998,37(4):32-40.

Chen Zunde, Chen Fugui. An attribute optimization method in oil and gas prediction with seismic data [J]. Geophysical Prospecting for Petroleum, 1998,37(4): 32-40.

(编辑 刘杨 苗凌)