文章编号: 1000 - 2634(2007)06 - 0027 - 04

测试井数据资料的数据清洗技术研究*

张 允,姚 军,王子胜
(中国石油大学石油工程学院,山东 东营 257061)

摘要: 针对测井和试井资料中存在数据质量的问题,提出了一种基于聚类分析和神经网络预测技术的数据清洗新方法。该方法首先检测测试井数据中存在空缺项的记录数据,对无空缺数据项的记录数据采用模糊聚类分析技术进行数据分类,再对各类数据分别进行蚁群聚类分析和神经网络学习并矫正噪声数据。将该数据清洗方法运用到试井分析中进行检验,取得了良好的效果。为提高测试井数据质量进行正确的解释评价提供了保证。

关键词: 测井数据;试井数据;数据清洗;模糊聚类;蚁群算法;神经网络
中图分类号: TE353; TP274 **文献标识码**: A

引 言

测井在油气田勘探开发中起着重要的作用,它可用来发现油气层、进行储层评价和油气资源评价;而试井是了解油藏动态的重要手段,其目的是通过油气井的测试资料来评价油井或油藏的生产动态,所以,测试井数据资料质量的优劣直接关系到测试井分析结果的准确与否[1]。而由于目前对这类数据基本上采用手工处理或者是简单的计算机处理方法,其数据质量仍不尽如人意,甚至有的数据不得不丢弃。为了进行准确和可靠的测井或试井分析,需要有高质量的测试井数据体,在这种情况下对测试井数据体的数据清洗显得尤为重要。

数据质量问题主要包括 2类:一类是与数据模式有关,另一类则是和实际数据有关。模式层的数据质量问题主要是由于结构设计不合理和缺乏属性间的完整性约束造成的,它可以通过设计程序来自动发现,但由于模式设计更多地涉及到对数据本身的理解,所以,一般都需要手工去实现清洗。

实例层的数据清洗包括重复记录清洗和属性数据清洗。"重复记录清洗"是用来检测和消除重复记录的;"属性数据清洗"就是通过填写空缺值,平滑噪声数据,识别、矫正、删除孤立点,并解决不一致问题来提高数据质量,从而保证测井或试井分析的顺利进行[2]。

1 基于模糊聚类和神经网络技术的数据清洗模型

数据清洗用于解决测试井数据体中数据质量问题,这是一个非常复杂和繁琐的过程,需要经过不同的方法进行处理才能得到高质量的测试井数据体。根据目前油田上测试井数据体的实际情况以及测试井解释和评价的需要[3],并在参考国外文献中数据清洗方法[4-6]的基础上提出了如图 1所示的测井和试井资料数据清洗的逻辑结构。

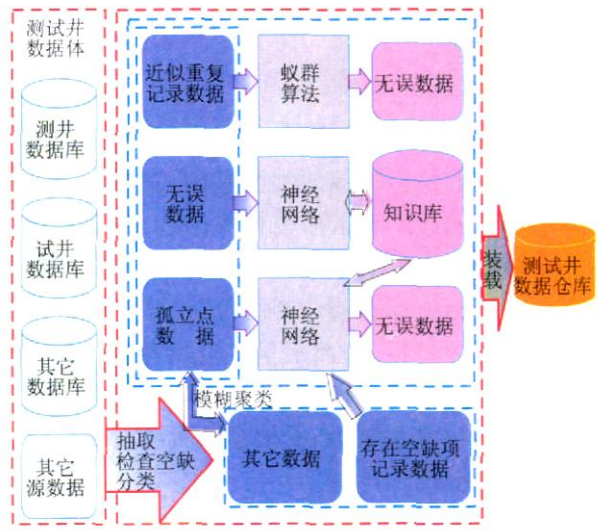

图 1 测井和试井数据资料清洗的逻辑结构

* 收稿日期: 2007 - 06 - 12
基金项目: 国家重点基础研究发展计划 973项目"碳酸盐岩缝洞型油藏数值模拟研究"(2006CB202405)。
作者简介: 张 允(1978 -),男(汉族),山东临沂人,博士研究生,从事油气田开发理论与系统工程研究。

从图 1中可以看出,对测井和试井数据资料的清洗可分为 3步:

(1) 从源数据中抽取数据检测存在空缺数据项的记录,并且把数据分为 2类,一类为存在空缺项记录数据,其它数据为另一类。

(2) 对去除存在空缺项记录的数据项进行模糊聚类分析,然后通过分析将其分为无误数据、孤立点数据、近似重复记录数据 3类[7]。其中孤立点数据既有错误的数据也有正确的数据;重复记录数据主要是近似重复记录的数据,因为在关系数据库系统中,只有当 2条记录的所有属性值都完全相同时才认为是重复数据;而所谓近似重复对象是指表现形式不同但语义上相同的对象,从狭义角度来看,如果 2条记录在某些属性上的值相等或足够相似,则认为这 2条记录互为近似重复记录,所以完全重复记录数据在关系数据库中是不存在的,这主要来源于其它的数据源。

(3) 对无误的数据进行神经网络学习并把获得的知识放入知识库,同时,知识库也指导神经网络学习;再通过神经网络来识别孤立点,对由模糊聚类分析出孤立点数据中的错误数据进行识别和修正,并检测和消除重复记录的问题,分析预测和填补空缺值,继而得到无误数据;最后把无误测试井数据体装载到测试井数据仓库中供测试井解释评价使用。

2 模糊聚类算法实现测试井数据体的预处理

数据聚类是对测试井数据体清洗研究的重要内容。模糊聚类分析以相似性为基础,主要用于研究样本的分类问题。为了研究样本间的关系,需要选择一个能反映研究对象之间亲疏关系的统计量,即反映样本间相似程度的分类统计量,将该统计量作为分类的数量指标,从而定量地进行分类。常用的分类统计量主要有距离和相似系数 2大类[8],结合距离系数与相似系数,采用相似度来进行测试井数据体的数据清洗。

设样本集 **X**有 m 个样本: $x_i\ (i = 1, 2, \cdots, m)$,则 $\mathbf{X} = [x_1, x_2, \cdots, x_m]^{\mathrm{T}}$;每个样本有 n个属性,第 i个样本的第 k个属性记为 x_{ik},则 $x_i = [x_{i1}, x_{i2}, \cdots, x_{in}]$。

(1) 计算样本 x_i的相似系数

相似系数法有多种,本文采用相关系数法,则第 i个样本 x_i与第 j个样本 x_j之间的相似系数 r_{ij}为

$$r_{ij} = \frac{\sum_{k=1}^{n}\left[\left|x_{ik} - \overline{x_i}\right|\left|x_{jk} - \overline{x_j}\right|\right]}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \overline{x_i})^2 \sum_{k=1}^{n}(x_{jk} - \overline{x_j})^2}} \tag{1}$$

式中

$$\overline{x_i} = \frac{1}{n}\sum_{k=1}^{n} x_{ik}$$

$$\overline{x_j} = \frac{1}{n}\sum_{k=1}^{n} x_{jk}$$

(2) 计算距离系数

根据测试井数据体,采用欧氏距离公式计算距离系数,则第 i个样本与第 j个样本之间的欧氏距离为

$$d_{ij} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{2}$$

(3) 计算相似度

为了更好地确定样本之间的相似程度,把近似重复样本记录数据和孤立点数据区分出来,特引入包含相似系数和距离系数 2个参数的相似度[9],即

$$s_{ij} = \frac{1}{2}(r_{ij} + d_{ij}) \tag{3}$$

(4) 层次聚类谱系图

在数据清洗过程中,模糊相似矩阵 R就被改造成了一个模糊等价关系矩阵 R^*。对满足传递性的模糊分类关系的 R^*进行聚类处理,给定不同置信水平的 λ,求 R_λ^*阵,这样最后就得到动态聚类谱系图。

(5) 样本集分类

根据得到的动态聚类谱系图,确定 λ在某一范围 (如 $\lambda > 0.85$) 的自成一类的单个样本为孤立点样本,λ在某一范围 (如 $\lambda < 0.1$) 的非自成一类的样本为近似重复记录样本数据,去除这 2部分的样本为无误样本,这样就把数据初步分成了孤立点数据、近似重复记录数据和无误数据 3类。

3 蚁群算法实现近似重复记录数据的处理

经过模糊聚类算法得到的近似样本集的样本数大大减小,而蚁群算法[10]是利用全局搜索最优解,然后合并或消除重复样本,所以正是利用该算法的这个优点来对近似重复记录的数据进行清洗处理[11]。

设近似重复记录样本集为 $\mathbf{Y} = [x_1, x_2, \cdots, x_r]^{\mathrm{T}}$ 的一个划分为 $\mathbf{C}(c_1, c_2, \cdots, c_s)$,选取 s个聚类中心,$Z_1, Z_2, \cdots, Z_s$,取其聚类半径 $r_1', r_2', \cdots, r_s'$,总的类

内离散度和为

$$J = \sum_{j=1}^{s}\sum_{x_i\in C_j} d'(x_i, Z_j) \tag{4}$$

式中

Z_j—C_j的聚类中心, $j = 1, 2, \cdots, s$;

$d'(x_i, Z_j)$—样本 x_i到其聚类中心的距离, $i = 1, 2, \cdots, r$。

在第 i个样本处设置 1个蚂蚁,该样本分配给第 j个聚类中心 Z_j,蚂蚁就在样本 i到聚类中心 Z_j的路径上留下外激素 τ_{ij},第 i个蚂蚁选择聚类中心 Z_j的概率为

$$p_{ij} = \frac{\tau_{ij}}{\sum_{j=1}^{s}\tau_{ij}} \tag{5}$$

更新方程为

$$\tau_{ij}^{new} = \rho\tau_{ij}^{old} + \frac{Q}{d'(x_i, Z_j)} \tag{6}$$

式中

ρ—强度的持久性系数,一般取 0.5～0.9;

Q—正常数。

每只蚂蚁从聚类中心出发,在整个解空间中搜索到下一个样本点后,返回聚类中心;再从聚类中心出发,在整个解空间中搜索到另一个样本点,再返回聚类中心,当搜索到 L个样本点(该聚类原来的样本点总数)返回聚类中心后,就认为蚂蚁完成了一个路径的搜索。为使蚂蚁在同一路径的搜索中不重复搜索同一个样本点,给每只蚂蚁设置一个禁忌表 $tabu_j(k)$ $(j = 1, 2, \cdots, s; k = 1, 2, \cdots, L_j)$。规定:如果 $tabu_j(k)$ 的值为 1,则结点 k是可以选择的搜索样本点,当蚂蚁选择了结点 k后,就将 $tabu_j(k)$ 置为 0,此时蚂蚁就不能选择结点 k。当聚类中心确定时,聚类的划分按下面的最邻近法则进行,若样本 x_i, j满足:

$$\min_{k=1, 2, \cdots, L_j} x_i - Z_j \le r_j'$$

则 x_i属于第 j类。

最后,根据蚁群聚类算法得到的结果对非自成一类的聚类进行合并或消除,从而达到对近似重复记录数据清洗的目的。

4 神经网络算法实现孤立点数据和空缺数据的处理

为了实现对孤立点数据和空缺数据的准确处理,首先必须对无误数据进行神经网络学习,把从无误数据中学习的规则和模式等知识放入知识库,然后再通过这些知识对孤立点数据进行检测[12],确属错误数据的就把错误数据中的错误数据项去除,最后通过神经网络技术对其进行清洗,主要是利用神经网络技术对空缺数据进行预测[13]。

本文采用三层感知器 BP算法,设输入向量为 x,期望输出向量为 y,输入层和输出层为 m 个节点,无偏置元的隐藏层为 n 个节点,记第 $k(k = 1, 2)$层第 j个神经元的输入为 u_j^k,输出为 v_j^k,步长为 δ,学习率为 η,动量因子为 α,迭代次数为 t,神经元的功能函数为阀值等于 0的 Sigmoid函数

$$v_j^{(k)} = f(u_j^{(k)}) = \frac{1}{1 + \exp(-u_j^{(k)})}$$

,则:

记样本输入 $x = u^{(0)}$,输出 $y = v^{(2)}$,第 t步的权系数 $\omega(t) = \{\omega_{ij}^{(k-1)}(t) \mid k = 1, 2\}$,则:

输出层

$$v_l^{(2)} = f(u_l^{(2)}),\ u_l^{(2)} = \sum_{j=1}^{r} v_j^{(1)}\omega_{jt}^{(1)}(t)$$

$$l = 1, 2, \cdots, m' \tag{7}$$

隐藏层

$$v_j^{(1)} = f(u_j^{(1)}),\ u_j^{(1)} = \sum_{i=1}^{m} v_i^{(0)}\omega_{ij}^{(0)}(t)$$

$$j = 1, 2, \cdots, n' \tag{8}$$

输入层

$$v_i^{(0)} = f(u_i^{(0)}),\ u_i^{(0)} = x_i,\ i = 1, 2, \cdots, m' \tag{9}$$

输出层与隐藏层之间的权值

$$\omega_{jt}^{(1)}(t+1) = \omega_{jt}^{(1)}(t) + \eta\delta_l^{(2)} v_j^{(1)} + \alpha\Delta\omega_{jt}^{(1)}(t) \tag{10}$$

$$\eta_l^{(2)} = v_l^{(2)}(1 - v_l^{(2)})(y_l^{(2)} - v_l^{(2)}) \tag{11}$$

输入层与隐藏层之间的权值

$$\omega_{ij}^{(0)}(t+1) = \omega_{ij}^{(0)}(t) + \eta\delta_j^{(1)} v_i^{(0)} + \alpha\Delta\omega_{ij}^{(0)}(t) \tag{12}$$

$$\delta_j^{(1)} = v_j^{(1)}(1 - v_j^{(1)})\delta_l^{(2)}\omega_{jt}^{(1)}(t) \tag{13}$$

误差指标函数

$$E = \frac{1}{2}\sum_{j=1}^{r}\sum_{l=1}^{m}(y_l^{(j)} - v_l^{(j)})^2 \tag{14}$$

经 t步迭代后神经网络在无误数据的训练样本集中学习结束,从而网络获得一组最佳权值,然后存入知识库。在对存在空缺数据进行预测时,这组最佳权值就是预测模型的参数,最后再通过神经网络模型对空缺值预测来填补空缺值。

5 应用分析

试井是了解油藏开发动态特性的主要手段之一,它可以为油气田开发提供大量的基础数据,如地

层原始压力、地层渗透率、目前地层平均压力、井间地层的连通性、表皮系数、油气田开发不同阶段的地层参数的变化、不同开采条件下地层中的流体分布等等。这些数据是油田开发方案的部署以及油田开发方案调整的重要依据,是合理经济地开发油气田的重要保障。因此,必须保证这些试井数据的可靠和准确,但由于在试井过程及记录数据的过程中会出现一些偏差甚至错误,如图 2所示是未经清洗的试井数据得出来的试井曲线。

该数据显然存在明显的问题,利用上述方法进行数据清洗,可以从图 3所示的试井曲线中看出这时的试井数据是合理的、有效的。可见,该数据清洗方法是可行的,它为测井和试井解释或评价提供了很好的质量保证。

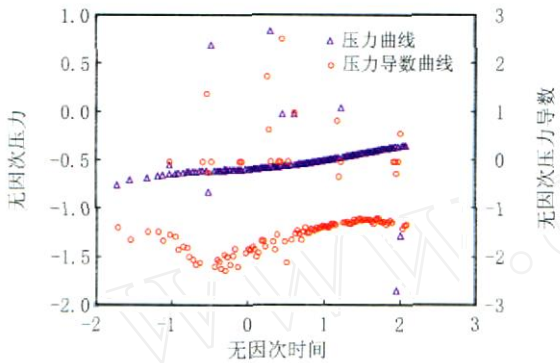

图 2 双对数坐标系下的试井曲线
(试井数据未经数据清洗)

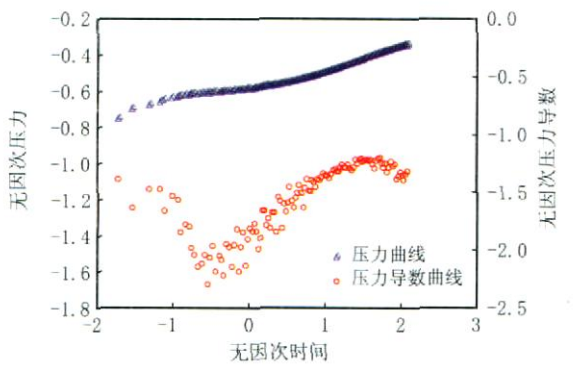

图 3 双对数坐标系下的试井曲线
(试井数据经过数据清洗)

6 结 论

(1) 测试井数据清洗的过程和方法,即先利用模糊聚类算法进行测试井数据体预处理,再利用蚁群算法对近似重复数据处理,最后利用神经网络算法对孤立点数据和空缺数据处理。

(2) 把该方法应用到试井分析中,其实验结果表明该方法达到了预期效果,高质量的测试井数据体为进行测试井以后正确的解释评价提供了保证。

参考文献:

[1] 姜汉桥,姚军,姜瑞忠. 油藏工程原理与方法 [M]. 山东东营:石油大学出版社, 2003.

[2] 王石,李玉忱,刘乃丽,等. 在属性级别上处理噪声数据的数据清洗算法 [J]. 计算机工程, 2005, 31 (9): 86 - 87.

[3] 吴洪彪,刘立明,陈钦雷,等. 四维试井理论研究 [J]. 石油学报, 2003, 24 (5): 57 - 62.

[4] Tamrapami D, Theodore J, John W, et al. Exploratory data mining and data cleaning[J]. Computers and Mathematics with Applications, 2003, 46 (5 - 6): 980.

[5] Lup L W, Li L M, Wang L T. A knowledge- based approach for duplicate elimination in data cleaning[J]. Information Systems, 2001, 26 (8): 585 - 606.

[6] Jan V D B, Solveig A C, Roger E, et al. Data cleaning: detecting, diagnosing, and editing data abnormalities [J]. Open Access, Freely Available Online, 2005, 2 (10): 0966 - 0970.

[7] Liu H, Shah S, JiangW. On- line outlier detection and data cleaning[J]. Computers and Chemical Engineering, 2004, 28 (9): 1635 - 1647.

[8] Tang L, Huang P Z, Xie W X. A new method of FCM considering the distribution of data [J]. Geometric and Information Science of Wuhan University, 2003, 28 (4): 476.

[9] Estivill- Castro V, Yang J. Fast and robust general purpose clustering algorithms [J]. Data Mining and Knowledge Discovery, 2004, 8 (2): 127 - 150.

[10] 杨欣斌,孙京诰,黄道. 基于蚁群聚类算法的离群挖掘方法 [J]. 计算机工程与应用, 2003, 39 (9): 12 - 13.

[11] 郭志懋,俞荣华,田增平,等. 一个可扩展的数据清洗系统 [J]. 计算机工程, 2003, 29 (3): 95 - 96.

[12] Manuel C L, Joaquín B, Francisco J, et al. Outlier detection and data cleaning in multivariate non- normal samples: the PAELLA algorithm [J]. Data Mining and Knowledge Discovery, 2004, 9 (2): 171 - 187.

[13] 欧邦才. 基于 BP神经网络的经济预测方法 [J]. 南京工程学院学报 (自然科学版), 2004, 2 (2): 12 - 13.

(编辑:宋艾玲)

ror, but also change the modality of construction and the deformation in phase axis in subjacent layers. Furthermore, it could change the frequency of reflective event. They are the pseudo phenomena caused by migration. To correctly identify them is very important for testing the correctness of velocity model and seismic data explanation.

Key words: velocity error; depth error; prestack migration; post-stack migration; migration pseudo phenomenon

THE TECHNOLOGY OF WELL LOGGING AND WELL TESTING DATA CLEANING

ZHANG Yun, YAO Jun, WANG Zi-sheng (School of Petrolum Engineering, China University of Petroleum, Dongying Shandong 257061, China). *JOURNAL OF SOUTHWEST PETROLEUM UNIVERSITY, VOL.* 29, *NO.* 6, 27 - 30, 2007 *(ISSN* 1000 - 2634, *IN CHINESE)*

Abstract: In terms of the quality problem of well logging and well testing data, the authors of this paper put forward a new data cleaning method based on clustering analysis and nerve network prediction technology with the method, the missing data in well logging data and well testing data are firstly checked up, clustering analysis adopted to classify the data that don't exist missing data, and then ant colony clustering analysis and neutral network study are carried on to correct the wrong data and fill in the missing data. The data cleaning method is proved to have good effects in the application of well testing analysis.

Key words: well logging data; well testing data; data cleaning; fuzzy clustering; ant colony algorithm; neutral network

RESERVOIR PARAMETER PREDICTION OF NEURAL NETWORK BASED ON PARTICLE SWARM OPTMIZATION

WANG Wen-juan[1], CAO Jun-xing[1], Zhang Yuan-biao[2], WANG Xiao-quan[1] (1. Chengdu University of Technology, Chengdu Sichuan 610059, China; 2. Packaging Engineering Institute, Jinan University, Zhuhai Guangdong 519070, China). *JOURNAL OF SOUTHWEST PETROLEUM UNIVERSITY, VOL.* 29, *NO.* 6, 31 - 33, 2007 *(ISSN* 1000 - 2634, *IN CHINESE)*

Abstract: A Predictive reservoir model with the self-adoption and complicated nonlinear property is set up. Because Multi-layer Feed Forward Neural Networks BP Algorithm exists weakness of getting bogged down in the local optima, stronger robustness and global convergence of PSO Algorithm, this research makes use of the particle swarm optimization (PSO) to improve the neural network, then, on the basis of the Luodai gas field in Sichuan Province, by the computation methods of PSO of the neural network, the reservoir characters (such as porosity, permeability) is forecasted, also the precision is tested and is compared with traditional computation methods of BP and LMBP, by which a obvious geography efficiency superior to traditional explanation methods is obtained, the shortcoming base from BP and LMBP algorithm are effectively overcome.

Key words: artificial neural networks; porosity; permeability; particle swarm optimization; the forecast of the layer parameters

GEOCHEMISTRY AND GENESIS OF CRUDE OIL OF TAIZHOU FORMATION IN NORTHERN JIANGSU BASIN

YUAN Ji-hua[1,2], LIU Guang-di[1] (1. Key Laboratory for Hydrocarbon Accumulation, Ministry of Education, China University of Petroleum, Changping Beijing 102249, China; 2. Research Center on Petroleum Resources Strategy of Ministry of Land and Resources, Beijing 100034, China). *JOURNAL OF SOUTHWEST PETROLEUM UNIVERSITY, VOL.* 29, *NO.* 6, 34 - 38, 2007 *(ISSN* 1000 - 2634, *IN CHINESE)*

Abstract: Taizhou Formation in Northern Jiangsu Basin is still a new petroleum exploration area now. By analyzing the fractional composition of saturate hydrocarbon and aromatic hydrocarbon and the geochemical parameter of the