文章编号：0253-2697(2018)02-0240-07 DOI:10.7623/syxb201802013

特约来稿

油气勘探开发常用数据挖掘算法优选

李大伟 石广仁

（中国石油勘探开发研究院 北京 100083）

摘要：迈入大数据时代的石油工业，需要充分挖掘石油工业大数据的巨大潜在价值。虽然数据挖掘已经在许多行业取得了丰硕的成果，但在油气勘探开发领域的应用还处于初始阶段，这主要由于油气勘探开发的数据及其应用具有自己的特殊性。数据挖掘常用的算法可分为回归、分类、聚类、估计、预测、关联分析等。其中的回归、分类是最成熟、应用最多的算法。但是对于具体的研究对象、不同的研究问题和数据源，不同的回归和分类算法又具有各自的适用性，因此需要针对具体问题优选适合该数据集的算法。以塔河油田的试油数据为例，以地层系数和油层分类为分析挖掘对象，详细解析了常用回归、分类算法的适用性。研究发现，对于常见的石油行业数据和研究对象：①最优的回归算法是反向传播神经网络(BPNN)，其次为支持向量机回归(*R*-SVM)和多元回归分析(MRA)；②最优的分类算法是支持向量机分类(*C*-SVM)，其次为贝叶斯逐步判别(BAYSD)；③ MRA 和 BAYSD 可以用于数据降维，BAYSD 的降维效果更好；④ R 型聚类分析(RCA)可以用于数据降维，Q 型聚类分析(QCA)可以用于样本约简；⑤在做具体的数据挖掘应用研究时一定要针对具体数据集对所用算法进行优选。

关键词：大数据；数据挖掘；回归；分类；数据清洗；优选；地层系数；油层分类

中图分类号：TE318 **文献标识码**：A

Optimization of common data mining algorithms for petroleum exploration and development

Li Dawei Shi Guangren

(*PetroChina Research Institute of Petroleum Exploration and Development*, *Beijing* 100083, *China*)

Abstract: For the petroleum industry in the big data period, it is necessary to fully exploit the great potential value of big data in the petroleum industry. Although data mining has achieved remarkable results in many industries, its application in the field of hydrocarbon exploration and development is still in its initial stage, which mainly lies on the particularity of the data and its specific applications in hydrocarbon exploration and development. The common algorithms in data mining can be divided into regression, classification, clustering, estimation, prediction, association analysis and so on. Among them, regression and classification are the most mature and most widely used algorithms. However, for specific research objects as well as different research questions and data resources, different regression and classification algorithms have their own applicability, thus it is required to optimize the appropriate algorithm for data sets aiming at specific problems. Taking the oil test data of Tahe oilfield as an example, and formation factor and reservoir classification as the mining objects, the applicability of common regression and classification algorithms is analyzed in detail. The results show that for common petroleum industry data and study objects, the optimal regression algorithm is the back propagation neural network (BPNN), followed by support vector machine regression (*R*-SVM) and multivariate regression analysis (MRA); the optimal classification algorithm is the support vector machine classification (*C*-SVM), followed by Bayesian stepwise discrimination (BAYSD); MRA and BAYSD can also be used for data dimensionality reduction, and the latter is better; R-type clustering analysis (RCA) can also be used for data dimensionality reduction, while Q cluster analysis (QCA) can be adopted for sample reduction; in the research of specific data mining applications, the algorithm must be optimized according to specific data set.

Key words: big data; data mining; regression; classification; data cleaning; optimization; formation factor; oil layer classification

引用：李大伟，石广仁. 油气勘探开发常用数据挖掘算法优选[J]. 石油学报，2018，39(2)：240-246.

Cite: LI Dawei, SHI Guangren. Optimization of common data mining algorithms for petroleum exploration and development[J]. Acta Petrolei Sinica, 2018, 39(2): 240-246.

根据大数据定义的一般标准，石油工业已迈入了大数据时代。以中国石油天然气集团公司为例，经过“十五”—“十二五”，已有约 70 个大型信息系统完成了建设并上线运行，其中仅“勘探与生产技术数据管理系

基金项目：国家重大科技专项“全球油气资源评价与选区选带研究”(2016ZX05029)资助。

第一作者及通信作者：李大伟，男，1969 年 5 月生，1991 年获中国地质大学(武汉)学士学位，1996 年获中国地质大学(北京)博士学位，现为中国石油勘探开发研究院高级工程师，主要从事海外勘探开发信息化建设、应用与数据挖掘工作。Email：leedw@petrochina.com.cn

统”(A1 系统)中，就管理着约 1500 TB 的数据和约 30×10^4 口井的结构化数据[1]。尚未入库、分散在各个单位和个人的数据更是难以统计。但是，从目前已入库数据的应用情况看，主要还局限于数据存储、管理、简单的报表和一般的查询应用，远远没有发挥这些宝贵数据资产的价值，现在所管理和利用的数据仅是冰山一角。如何有效管理和利用石油工业的大数据，是管理人员、研究人员和信息人员都非常关注的问题，而数据挖掘(data mining)正是解决这一问题的有效途径之一，其可以将数据转换成有用的信息和知识，从而实现从“大数据”到“大信息”、“大知识”的跨越。

“数据挖掘”是指从大量数据中揭示出隐含的、先前未知、并具有潜在价值信息的过程，是人工智能和数据库领域研究的热点问题之一[2-4]。数据挖掘已经在许多行业(如通讯、金融、电子商务等)都得到了广泛应用，但在石油工业的应用还处于初始阶段[5-10]，这主要是由于石油工业的数据及其应用具有特殊性(如数据量大、数据格式多、存储分散、非结构化数据多、研究对象非均质性强、专业应用软件多等)，从而影响了数据挖掘的应用[9]。从中国知网[11]可以统计出中国数据挖掘研究的情况：2002 年之后进入快速发展期，到 2012 年达到高峰，近几年虽然有所减少，但仍是许多领域的研究热点；从研究领域看，绝大部分集中在计算机相关学科，与石油等地质矿产冶金行业有关的论文还很少，只占总量的约 1.2%。中国石油行业的数据挖掘还没有形成系统的理论体系和实用的挖掘平台，只有一些具体的应用和少量的算法优化。目前国内外已经有不少成熟、实用甚至开源的数据挖掘软件平台(Weka、SPSS、Rapid Miner、Matlab、TipDM 等)。仅在 Weka(3.9.1 版本)中就有 3 大类、约 130 种不同的算法，其中关联算法 6 种、聚类算法 12 种、分类算法 110 种，还有预处理算法约 80 种。在应用方面，虽然国外的一些知名石油公司做了许多尝试、并获得了良好的回报[10]，但研究程度还远远不够。随着石油工业大数据的发展和应用，以及各种相关算法的不断完善[12]，数据挖掘会在石油工业的许多业务领域发挥越来越重要的作用[13-16]，同时结合其他相关新技术、新思维(云计算、认知计算、物联网、虚拟现实)等，新的挖掘技术和体系也会不断提出[17-20]，必将成为油气公司降本增效的重要途径，以及数字油田、智能油田建设的重要工具。因此，在石油工业开展数据挖掘研究和应用的空间和前景都非常大，而且如何利用好现有成熟的数据挖掘算法和平台应是一个重点研究方向。

数据挖掘常用的算法可分为回归、分类、聚类、估计、预测、关联分析等，其中回归、分类是最成熟、应用最多的算法[9]。但是对于具体的研究对象、研究问题和数据源，不同的回归和分类算法又具有不同的适用性。常用的回归算法有 3 种：多元回归分析(MRA)、反向传播神经网络(BPNN)和支持向量机回归(*R*-SVM)。常用的分类算法有 5 种：支持向量机分类(*C*-SVM)、决策树(DTR)、朴素贝叶斯(NBAY)、贝叶斯判别分析(BAYD)以及贝叶斯逐步判别分析(BAYSD)。对于同一套数据源和同一类问题，这 8 种算法使用的已知参数相同，要预测的未知量也相同，其区别在于具体的算法和计算结果。此外，MRA、BPNN 和 *R*-SVM 是用于处理实数的，而 *C*-SVM、DTR、NBAY、BAYD 和 BAYSD 算法得到的是整数。其中，只有 MRA 是线性的，而其他 7 种算法是非线性的，这是由于 MRA 建立的是一个线性方程，而其他 7 种算法建立的方程是非线性的。

由于 DTR 算法的使用非常复杂[9,21]，而 BAYSD 算法要优于 BAYD，因此本文只讨论其他几类常用数据挖掘回归和分类算法，并进行比较，为在实际应用中进行算法优选提供一定借鉴。

1　回归与分类算法简介

1.1　模型及方法

假设有 n 个学习样本，每个样本有 $m+1$ 个参数 $(x_1,x_2,\cdots,x_m,y_i^*)$ 的成组观察值 $(x_{1i},x_{2i},\cdots,x_{mi},y_i^*)$ $(i=1,2,\cdots,n)$；$n>m$，且在实际应用中 n 一般要远远大于 m，这样才能保证预测结果的精度和代表性。可以将 m 个参数的 n 个学习样本定义为 n 个向量，即学习样本的表达式为：

$$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{im},y_i^*)\quad(i=1,2,\cdots,n)\tag{1}$$

令 $\boldsymbol{x}_0$ 为 $(x_{i1},x_{i2},\cdots,x_{im})$ 中一个向量的一般形式。MRA、BPNN、NBAY 和 BAYSD 算法的原理是一样的，就是要建立一个表达式 $y=y(\boldsymbol{x}_0)$，并使其值最小：

$$\min(y)=\sum_{i=1}^{n}[y(\boldsymbol{x}_0)-y_i^*]^2\tag{2}$$

另外，*R*-SVM 和 *C*-SVM 是要建立一个方程 $y=y(\boldsymbol{x}_0)$，使得基于支持向量点的分类间隔最大化，以得到最优的分隔线。当然，由于这几种算法使用的具体方法不同，所得结果的精度也是不一样的。

$y=y(\boldsymbol{x}_0)$ 就是在学习过程中所得到的拟合公式。不同算法所得到的拟合公式是不一样的。在本文中，将 y 定义为一个单变量。

具体的计算流程为：①学习：使用 n 个学习样本得到拟合公式；②检验：将 n 个学习样本代入拟合公式中，得到各自的预测值 $(y_1,y_2,\cdots,y_n)$，以检验该算法的拟合度；③预测：将预测样本所表示的 k 个预测样本

代入拟合公式中，得到各自的预测值（y_{n+1}，y_{n+2}，…，y_{n+k}）。

其中，预测样本为：

$$\boldsymbol{x}_i^* = (x_{i1}, x_{i2}, \cdots, x_{im}) \quad (i = n+1, n+2, \cdots, n+k) \tag{3}$$

1.2 误差分析

当使用 MRA、BPNN、*R*-SVM、*C*-SVM、NBAY 和 BAYSD 这 6 种算法时，为了表示学习样本和预测样本的预测变量 y 的计算精度，采用相对误差绝对值 R_i、平均相对误差绝对值 $\bar{R}$ 和总平均相对误差绝对值 $\bar{R}^*$。

对于每个样本，其相对误差绝对值 R_i 为：

$$R_i = |(y_i - y_i^*)/y_i^*| \times 100\% \tag{4}$$

需要说明的是：y_i^* 不能取 0 值，否则会发生浮点溢出。因此，对于回归算法，要删除 $y_i^* = 0$ 的样本；对于分类算法，y_i^* 的值应当为正整数。

对于所有的学习样本或预测样本，其平均相对误差绝对值 $\bar{R}$ 为：

$$\bar{R} = \sum_{i=1}^{N_s} R_i / N_s \tag{5}$$

式(5)中，对于学习样本，$N_s = n$；对于预测样本，$N_s = k$。

对于学习样本，R_i 和 $\bar{R}$ 称为拟合误差，用于反映学习过程的拟合度；在本文中，将 $\bar{R}_i$ 命名为 $\bar{R}_1$。对于预测样本，R_i 和 $\bar{R}$ 称为预测误差，用于反映预测过程的精确度；在本文中，将 $\bar{R}$ 命名为 $\bar{R}_2$。

对于所有的样本，将总平均相对误差绝对值 $\bar{R}^*$ 定义为：

$$\bar{R}^* = [\bar{R}_1 + \bar{R}_2]/2 \tag{6}$$

如果没有预测样本，则有 $\bar{R}^* = \bar{R}_1$。这是一个非常重要的指标，用于确定预测结果是否可用。根据经验，如果该值小于 10%，计算结果基本可信，否则就需要寻找更好的算法。

1.3 非线性和精度分析

因为 MRA 是一种线性算法，对于某一个所要研究的问题，$y = y(\boldsymbol{x})$ 的值表示了该问题的非线性程度。

对于线性算法（MRA）和非线性算法（BPNN、*R*-SVM、*C*-SVM、NBAY 和 BAYSD），其值表示了该算法对于所研究问题得到结果的精度。

对于 BPNN 来说，其所得到的 $y = y(\boldsymbol{x})$ 是一个隐式表示式（无法用常规的数学公式表达），而其他 4 种算法得到的 $y = y(\boldsymbol{x})$ 是显式表示式（可以用常规的数学公式表达）。

2 数据预处理

数据挖掘预处理是指在数据挖掘之前对原始数据进行一些必要的前期处理工作。这是因为用于数据挖掘的原始数据通常存在各种各样的问题（如不正确、不及时、不完整、不一致、不安全、不规范等），无法直接进行数据挖掘，或挖掘结果不理想，所以数据预处理在整个数据挖掘中起着非常重要的作用。常用的数据预处理有数据清洗、集成、变换、消减等[1]。限于篇幅，本文仅对其中的数据清洗和数据消减进行简述。

2.1 数据清洗

在实际应用中的数据一般都是有“噪音”的，存在不完整、不一致等问题。数据清洗就是对原始数据中的缺失数据、重复数据、异常数据、错误数据等进行处理，解决数据中的人为误差，提高数据挖掘质量。

2.2 数据消减

对于石油工业的数据分析挖掘，由于涉及数据量大、变量多等因素，通常都需要花费大量时间对这些复杂的数据进行分析，这有时会使得分析挖掘难以进行（特别是做交互式数据挖掘时）。数据消减就是在保证数据的完整性和挖掘结果可靠性的前提下，减少用于挖掘的数据属性和样本数，以提高挖掘的速度和精度。对于中、小型数据集，使用一般的数据预处理就可以了，但对于大型数据集（特别是“大数据”）通常会需要进行数据消减。

数据消减的主要策略包括[9]：①数据聚合：如建立数据立方体聚合，这种聚合主要用于构建数据立方体数据仓库；②检测并删除那些无关的、弱相关或冗余的属性或维度；③数据压缩：使用编码技术（例如最小编码长度或小波）来压缩数据集；④数据块消减：使用更简单的数据表达形式（如参数模型，以及聚类、采样等非参数模型）来代替原始的数据。

降维处理和样本消减处理是 2 种常用的数据消减方法，在相关领域的应用也比较广泛[9,21]。通过降维，可以减少不相关变量的数量。通过样本消减处理，可以减少样本的个数。由于 MRA 和 BAYSD 这 2 种算法都能够给出预测因变量 y 与各个不相关自变量（x_1，x_2，…，x_m）的亲疏关系，所以其可以用于降维处理。因为 MRA 属于线性相关分析，而 BAYSD 属于非线性相关分析，所以对于非线性强的应用问题，用 BAYSD 进行降维处理效果会更好[22]。

通过石广仁等的研究还发现[9,22]，R 型聚类分析（RCA）可以作为一种先导的降维工具，而 Q 型聚类分析（QCA）可以作为一种先导的样本消减工具。需要注意的是，RCA 和 QCA 都属于线性算法。这里所谓的“先导”是指降维或样本消减的结果正确与否，还需要使用非线性工具（如将 BPNN 用于回归问题，*C*-SVM 用于

分类问题)进行二次验证,以确定能消减多少个样本或多少个无关的变量。之所以要进行二次验证,是由于地质学研究对象和所涉及数据的复杂性,在大多数情况下各种地质学数据之间的关系都是非线性的。

3　应用实例

3.1　研究目的与样本

笔者利用实际试油数据,通过一定的流程来选择合适的算法,预测地层的流动能力并进行油层分类,以确定最佳的数据挖掘算法,并论证本文的观点以及实现过程。该研究结果对于缺乏试油数据的地区具有很大的应用价值;也可以推广到其他数据挖掘案例,用于指导油气勘探开发的数据挖掘工作。

算例中的 13 个样本数据来自于塔里木盆地塔河油田[23],每个样本都有 7 个独立变量(原油黏度、原油日产量、油嘴尺寸、油压、原油密度、气油比和含水率)的试油结果数据(表 1)。其中,因变量 y^* 为地层系数和油层分类(表 2)。

在本文中,使用 12 个样本作为学习样本,1 个作为预测样本,且每个样本具有 7 个独立自变量(表 1)。

表 1　塔河油田地层流动能力分析输入数据

Table 1　Input data of formation flow capacity analysis in Tahe oilfield

样本类型	样本编号	井号	自变量							因变量 y^*	
			原油黏度 x_1/(mPa·s)	原油日产量 x_2/(t·d^{-1})	油嘴尺寸 x_3/mm	油压 x_4/MPa	原油密度 x_5/(g·cm^{-3})	气油比 x_6	含水率 x_7/%	地层系数/(m·μm^2)$^{-1}$	油层分类
学习样本	1	TK631	35.6	133.98	6.0	9.56	0.96	0	8.50	19.50	3
	2	S65	8.0	298.00	8.0	7.30	0.95	0	0.29	23.00	3
	3	TK413	78.5	81.57	4.0	7.80	0.96	68.00	0.08	3.90	4
	4	TK404	78.0	219.00	6.9	11.69	0.95	43.90	0	67.77	2
	5	TK409	28.3	160.00	7.9	5.16	0.96	52.90	0.10	26.07	2
	6	S67	11.7	413.00	8.0	10.35	0.97	0	0.20	102.50	1
	7	S74	22.6	136.00	6.0	9.28	0.98	22.00	0	82.11	2
	8	TK609	25.5	116.00	7.0	6.10	0.96	0	7.41	21.21	3
	9	TK313	1.9	201.70	6.0	17.00	0.90	88.00	0.50	126.00	1
	10	TK607	24.6	203.30	8.0	9.60	0.95	37.78	0.03	191.31	1
	11	TK444	9.2	197.90	6.0	8.75	0.99	0	0	130.19	1
	12	TK629	22.6	83.40	6.0	4.30	0.98	0	0.11	6.99	3
预测样本	13	TK442	17.9	232.00	7.0	7.50	0.98	0	7.13	(4.80)	(3)

注:油层分类中,1—高产油层;2—中产油层;3—低产油层;4—干层。括号中的数字不是输入数据,只是作为预测和实际对比。

表 2　根据地层系数确定的油层分类

Table 2　Oil layer classification confirmed by formation factor

油层分类	地层系数/(m·μm^2)$^{-1}$
高产油层	>100
中产油层	(25,100]
低产油层	[4,25]
干层	<4

3.2　输入参数

输入参数包括 12 个学习样本和 1 个预测样本的各个已知变量值 $x_i(i=1,2,3,4,5,6,7)$,以及 12 个学习样本的预测变量 y^* 值(表 1)。需要说明的是:对于回归计算,y^* 值为地层系数;对于分类计算,y^* 值为油层分类。

3.3　回归计算

3.3.1　学习过程

使用表 1 中的 12 个学习样本和 *R*-SVM、BPNN 和 MRA 这 3 种算法进行回归计算。首先建立了 7 个独立变量($x_1,x_2,x_3,x_4,x_5,x_6,x_7$)所对应的地层系数的 3 个拟合关系式;然后将表 1 中的 12 个学习样本的 $x_1,x_2,x_3,x_4,x_5,x_6,x_7$ 的值分别代入这 3 个拟合公式,从而得到每个学习样本的地层系数。表 3 为 *R*-SVM、BPNN 和 MRA 这 3 种算法的最终预测结果。

3.3.2　预测过程

将表 1 中所列出的 1 个预测样本的 x_i 值分别代入 *R*-SVM、BPNN 和 MRA 的拟合公式,从而得到该预测样本的地层系数(表 3)。同时,计算了每种算法所得结果的误差(表 4),其中 MRA 算法的 $\bar{R}^*=516.65\%$,所以对于算例而言,其预测得到的 y 值与其相关的 7 个独立变量($x_1,x_2,x_3,x_4,x_5,x_6,x_7$)之间具有非常强的非线性关系。用 *R*-SVM 和 MRA 算法所得解的精度非常低,用 BPNN 得到的结果精度中等。

鉴于本文算例是一个非线性关系很强的问题,所以 *R*-SVM、BPNN 和 MRA 这 3 种算法都不适用。

3.4　分类计算

3.4.1　学习过程

使用表 1 中的 12 个学习样本以及 *C*-SVM、NBAY、BAYSD 和 MRA 算法进行分类计算。首先建

立了 7 个独立变量(x_1,x_2,x_3,x_4,x_5,x_6,x_7)所对应的油层分类的 4 个拟合关系式,然后将表 1 中的 12 个学习样本的 x_1,x_2,x_3,x_4,x_5,x_6,x_7 的值分别代入 4 个拟合公式,从而得到每个学习样本的油层分类。表 5 列出了 *C*-SVM、NBAY、BAYSD 和 MRA 这 4 种算法的最终学习结果。

表 3 塔河油田地层系数预测结果

Table 3 Prediction results of formation factor in Tahe oilfield

样本类型	样本编号	试油结果	回归算法					
			R-SVM		BPNN		MRA	
		$y^*/(\mathrm{m}\cdot\mu\mathrm{m}^2)^{-1}$	$y/(\mathrm{m}\cdot\mu\mathrm{m}^2)^{-1}$	R_i/%	$y/(\mathrm{m}\cdot\mu\mathrm{m}^2)^{-1}$	R_i/%	$y/(\mathrm{m}\cdot\mu\mathrm{m}^2)^{-1}$	R_i/%
学习样本	1	19.50	45.38	132.7	19.50	0.2	30.10	54.6
	2	23.00	46.95	104.1	41.70	81.2	33.20	44.2
	3	3.90	46.17	1083.9	3.90	0	−21.90	662.4
	4	67.77	47.23	30.3	70.30	3.7	92.00	35.7
	5	26.07	46.61	78.8	27.40	5.0	59.10	126.8
	6	102.50	47.44	53.7	140.00	36.2	103.00	0.4
	7	82.11	46.86	42.9	85.00	3.5	118.00	43.7
	8	21.21	45.26	113.4	28.80	35.9	7.90	62.8
	9	126.00	47.77	62.1	136.00	7.9	136.00	7.9
	10	191.31	47.22	75.3	191.00	0	119.00	37.7
	11	130.19	46.94	63.9	144.00	10.5	107.00	18.2
	12	6.99	46.11	559.4	6.60	5.6	17.60	151.0
预测样本	13	4.80	45.74	833.1	4.82	0.5	49.40	929.5

表 4 用 3 种回归算法计算塔河油田地层系数的结果比较

Table 4 Comparison of formation factor calculation results by three regression algorithms in Tahe oilfield

算法	拟合公式	平均相对误差绝对值/%			预测量 y 与各独立变量的相关性(由大到小排序)	计算时间(在 Intel Core 2 PC 机上)/s	计算精度
		$\bar{R}_1$	$\bar{R}_2$	$\bar{R}^*$			
R-SVM	非线性,显式	200.05	852.99	526.51	(计算不出)	3	非常低
BPNN	非线性,隐式	15.81	0.51	8.16	(计算不出)	30	中等
MRA	线性,显式	103.78	929.52	516.65	x_4,x_3,x_5,x_6,x_1,x_2,x_7	<1	非常低

表 5 塔河油田油层分类的预测结果

Table 5 Prediction results of oil layer classification in Tahe oilfield

样本类型	样本编号	试油结果	分类算法						MRA	
			C-SVM		NBAY		BAYSD			
		y^*	y	R_i/%	y	R_i/%	y	R_i/%	y	R_i/%
学习样本	1	3	3	0	1	66.7	3	0	3	0
	2	3	3	0	1	66.7	3	0	3	0
	3	4	4	0	2	50.0	4	0	4	0
	4	2	2	0	2	0	2	0	2	0
	5	2	2	0	2	0	2	0	2	0
	6	1	1	0	1	0	1	0	1	0
	7	2	2	0	2	0	2	0	1	50.0
	8	3	3	0	1	66.7	3	0	3	0
	9	1	1	0	1	0	1	0	1	0
	10	1	1	0	2	100.0	1	0	1	0
	11	1	1	0	1	0	1	0	1	0
	12	3	3	0	1	66.7	3	0	3	0
预测样本	13	3	3	0	1	66.7	1	66.7	2	33.3

注:试油结果中,1—高产油层;2—中产油层;3—低产油层;4—干层。MRA 计算得到的 y 通过四舍五入由实数转换成了整数。

3.4.2 预测过程

将表 1 中所列出的 1 个预测样本的 x_1,x_2,x_3,x_4,x_5,x_6,x_7 的值分别代入 *C*-SVM、NBAY、BAYSD 和 MRA 的拟合公式,从而得到该预测样本的油层分类。表 5 列出了 *C*-SVM、NBAY、BAYSD 和 MRA 这 4 种算法的预测结果。其中 MRA 算法的 $\bar{R}^* = 18.8\%$(表 6),所以对于本文算例,其预测得到的 y 值及其相关的 7 个独立变量(x_1,x_2,x_3,x_4,x_5,x_6,x_7)之间具有较强的非线性关系。用 *C*-SVM 所得解的精度非常高,不仅 $\bar{R}_1 = 0$,而且 $\bar{R}_2 = 0$,因此其总平均相对误差绝对值 $\bar{R}^* = 0$。如前所述,本文算例具有强非线性关系,所以只有 *C*-SVM 算法适用,而 NBAY 和 BAYSD 算法所得解的精度很低,所以并不适用。

综上所述,通过对比不同算法的计算结果可以发现:①由于地层系数具有非常强的非线性,*R*-SVM、BPNN、MRA 这几种回归算法都不适用;②由于油层分类具有较强的非线性,所以只有 *C*-SVM 算法是适用的,而 NBAY、BAYSD 这 2 种算法所得解的精度很低,也并不适用。

表 6　用 4 种分类算法计算塔河油田油层分类的结果比较

Table 6　Comparison of oil layer classification results by four classification algorithms in Tahe oilfield

算法	拟合公式	平均相对误差绝对值/%			预测量 y 与各独立变量的相关性(由大到小排序)	计算时间(在 Intel Core 2 PC 机上)/s	计算精度
		$\bar{R}_1$	$\bar{R}_2$	$\bar{R}^*$			
C-SVM	非线性，显式	0	0	0	计算不出	5	很高
NBAY	非线性，显式	34.72	66.7	50.7	计算不出	<1	很低
BAYSD	非线性，显式	0	66.7	33.3	$x_1, x_7, x_4, x_3, x_5, x_6, x_2$	1	很低
MRA	线性，显式	4.17	33.3	18.8	$x_2, x_4, x_1, x_5, x_3, x_6, x_7$	<1	强非线性

4　结　论

(1) 对于常用数据挖掘算法：最优的回归算法是 BPNN，其次为 R-SVM 和 MRA；最优的分类算法是 C-SVM，其次为 BAYSD；MRA 和 BAYSD 还可以用于数据降维，且 BAYSD 的降维效果更好；R 型聚类分析(RCA)可以用于数据降维，Q 型聚类分析(QCA)可以用于样本约简。

(2) R-SVM、BPNN 和 MRA 并不适用于本文算例所用数据集的回归处理，而 C-SVM 则适用于该数据集的分类处理。因此在做具体的数据挖掘应用研究时一定要针对具体问题选择适合其数据集的算法。

符号注释：n—学习样本数量；m—样本中独立变量数量；$\boldsymbol{x}_i$—第 i 个样本向量；$\boldsymbol{x}_0$—$(x_{i1}, x_{i2}, \cdots, x_{im})$中一个向量的一般形式；$x_{ij}$—第 i 个样本的第 j 个独立变量；y_i^*—第 i 个样本的观测值；y_i—第 i 个样本的计算值；R_i—第 i 个样本的相对误差绝对值；$\bar{R}$—平均相对误差绝对值；$\bar{R}^*$—总平均相对误差绝对值；N_s—学习或预测样本数；k—预测样本的个数；x_1—原油黏度，mPa·s；x_2—原油日产量，t/d；x_3—油嘴尺寸，mm；x_4—油压，MPa；x_5—原油密度，g/cm^3；x_6—气油比；x_7—含水率。

参　考　文　献

[1] 李大伟，熊华平，石广仁，等. 基于全球典型油气田数据库的数据挖掘预处理[J]. 大庆石油地质与开发，2016，35(1)：66-70.
LI Dawei，XIONG Huaping，SHI Guangren，et al. Preprocessing of the data tapping based on global typical oil and gas field database[J]. Petroleum Geology and Oilfield Development in Daqing，2016，35(1)：66-70.

[2] HAN J W，KAMBER M. Data mining：concepts and techniques[M]. 2nd ed. San Francisco：Morgan Kaufmann，2006.

[3] MAIMON O，ROKACH L. Data mining and knowledge discovery handbook[M]. 2nd ed. New York，USA：Springer，2010.

[4] 戴红，常子冠，于宁. 数据挖掘导论[M]. 北京：清华大学出版社，2015.
DAI Hong，CHANG Ziguan，YU Ning. Introduction to data mining[M]. Beijing：Tsinghua University Press，2015.

[5] 李功权，陈恭洋，吴东胜，等. 油气储层建模中的空间数据挖掘技术[J]. 石油天然气学报，2006，28(8)：47-49.
LI Gongquan，CHEN Gongyang，WU Dongsheng，et al. Spatial data mining in petroleum reservoir modeling[J]. Journal of Oil and Gas Technology，2006，28(8)：47-49.

[6] 李洪奇，郭海峰，郭海敏，等. 复杂储层测井评价数据挖掘方法研究[J]. 石油学报，2009，30(4)：542-549.
LI Hongqi，GUO Haifeng，GUO Haimin，et al. An approach of data mining for evaluation of complex formation using well logs[J]. Acta Petrolei Sinica，2009，30(4)：542-549.

[7] 严丽，王燕，范树平. 多元回归分析方法预测川东北礁滩相储层产能[J]. 新疆石油天然气，2011，7(4)：37-79.
YAN Li，WANG Yan，FAN Shuping. Output prediction of reef-bank facies in northeastern-Sichuan Basin using multiple regression analysis[J]. Xinjiang Oil & Gas，2011，7(4)：37-79.

[8] 钟仪华，王丹，李晴晴，等. 基于数据挖掘的低品位油藏经营评价指标分析[J]. 数据挖掘，2014，4(4)：32-37.
ZHONG Yihua，WANG Dan，LI Qingqing，et al. The analysis of evaluation index of low-grade reservoir operation based on data mining[J]. Hans Journal of Data Mining，2014，4(4)：32-37.

[9] 石广仁. 地学数据挖掘与知识发现[M]. 北京：石油工业出版社，2012.
SHI Guangren. Data mining and knowledge discovery for geoscientists[M]. Beijing：Petroleum Industry Press，2012.

[10] 檀朝东，陈见成，刘志海，等. 大数据挖掘技术在石油工程的应用前景展望[J]. 石油工程技术，2015(1)：49-51.
TAN Chaodong，CHEN Jiancheng，LIU Zhihai，et al. Application prospect of big data mining technology in petroleum engineering[J]. Petroleum Engineering Technology，2015(1)：49-51.

[11] 中国知网. [EB/OL]. [2017-01-01]. http：//www. cnki. net.
China National Knowledge Infrastructure. [EB/OL]. [2017-01-01]. http：//www. cnki. net.

[12] LUO Weiping，LI Hongqi，SHI Ning. Semi-supervised least squares support vector machine algorithm：application to offshore oil reservoir[J]. Applied Geophysics，2016，13(2)：406-415.

[13] 张莹，潘保芝. 松辽盆地火山岩岩性识别中测井数据的选择及判别方法[J]. 石油学报，2012，33(5)：830-834.
ZHANG Ying，PAN Baozhi. Selection and identification of log-

ging data for lithology recognition of volcanic rocks in Songliao Basin[J]. Acta Petrolei Sinica,2012,33(5):830-834.

[14] 张尘. 数据挖掘技术在石油勘探中的应用研究[J]. 中国石油和化工标准与质量,2014(6):49.
ZHANG Chen. Application of data mining in petroleum exploration[J]. China Petroleum and Chemical Standard and Quality, 2014(6):49.

[15] 尚福华,原野,王才志,等. 基于知识库的解释模型智能优选测井数据处理方法[J]. 石油学报,2015,36(11):1449-1456.
SHANG Fuhua,YUAN Ye,WANG Caizhi,et al. A logging data processing method of intelligent optimization logging interpretation model based on knowledge-base[J]. Acta Petrolei Sinica, 2015,36(11):1449-1456.

[16] DURLOFSKY L J,DIMITRAKOPOULOS R. Smart oil fields and mining complexes[J]. Mathematical Geosciences, 2017, 49(3): 275-276.

[17] 王建君. 分布式数据挖掘研究[J]. 电子商务,2017(7):41-42.
WANG Jianjun. Study on distributed data mining[J]. E-Business Journal,2017(7):41-42.

[18] 杨震宇. 基于 MapReduce 框架下的数据挖掘方法研究[J]. 中国高新技术企业,2017(4):8-10.
YANG Zhenyu. Study on data mining methods based on MapReduce frame[J]. China High-Tech Enterprises,2017(4):8-10.

[19] 杨涛. 基于决策树算法的石油基础数据挖掘系统应用研究[J]. 电子设计工程,2016,24(18):16-18.
YANG Tao. Oil based data mining system based on decision tree algorithm applied research[J]. Electronic Design Engineering, 2016,24(18):16-18.

[20] 檀朝东,项勇,赵昕铭,等. 基于大数据的油气集输系统生产能耗时序预测模型[J]. 石油学报,2016,37(增刊 2):158-164.
TAN Chaodong, XIANG Yong, ZHAO Xinming, et al. Energy consumption prediction and application in oil and gas gathering and transferring system production based on large data[J]. Acta Petrolei Sinica,2016,37(S2):158-164.

[21] 喻思羽,李少华,何幼斌,等. 基于样式降维聚类的多点地质统计建模算法[J]. 石油学报,2016,37(11):1403-1409.
YU Siyu,LI Shaohua,HE Youbin,et al. Multiple-point geostatistics algorithm based on pattern scale-down cluster[J]. Acta Petrolei Sinica,2016,37(11):1403-1409.

[22] SHI Guangren,ZHU Yixiang,MI Shiyun,et al. A big data mining in petroleum exploration and development[J]. Advances in Petroleum Exploration and Development,2014,7(2):1-8.

[23] 康志宏,郭春华,伍文明. 塔河碳酸盐岩缝洞型油藏动态储层评价技术[J]. 成都理工大学学报:自然科学版,2007,34(2):143-146.
KANG Zhihong, GUO Chunhua, WU Wenming. Technique of dynamic descriptions to the crack and cave carbonate rock reservoir in the Tahe oil field,Xinjiang,China[J]. Journal of Chengdu University of Technology:Science & Technology Edition,2007, 34(2):143-146.

(收稿日期 2017-01-12 改回日期 2018-01-17 编辑 王培玺)

(上接第 200 页)

[14] 梁丹,冯国智,谢晓庆,等. 聚合物驱阶段注采动态特征及影响因素分析[J]. 特种油气藏,2014,21(5):75-78.
LIANG Dan,FENG Guozhi,XIE Xiaoqing,et al. Analysis on features and influencing factors of injection-production performance during poymer flooding[J]. Special Oil & Gas Reservoirs,2014, 21(5):75-78.

[15] 何春百,冯国智,康晓东,等. 海上油田聚合物驱注入方式室内对比评价研究[J]. 石油地质与工程,2014,28(2):136-138.
HE Chunbai,FENG Guozhi,KANG Xiaodong,et al. Comparison and evaluation on injection patterns of polymer flooding oversea oilfield[J]. Petroleum Geology and Engineering, 2014, 28(2): 136-138.

[16] 王敬,刘慧卿,汪超锋,等. 聚合物驱数学模型的若干问题[J]. 石油学报,2011,32(5):857-861.
WANG Jing,LIU Huiqing,WANG Chaofeng,et al. Discussions on some problems about the mathematical model of polymer flooding[J]. Acta Petrolei Sinica,2011,32(5):857-861.

[17] 穆文志. 粘弹性聚合物驱数值模拟研究[D]. 大庆:大庆石油学院,2005.
MU Wenzhi. The numerical simulation study of viscoelastic polymer displacement[D]. Daqing:Daqing Petroleum Institute,2005.

[18] 杨晶,朱焱,李建冰,等. 聚合物驱剖面返转时机及其影响因素[J]. 油田化学,2016,33(3):472-476.
YANG Jing,ZHU Yan,LI Jianbing,et al. Influence factors of profile reversal opportunity during polymer flooding[J]. Oilfield Chemistry,2016,33(3):472-476.

[19] 陈晨,董朝霞,高玉莹,等. 盐水组成对极性组分在石英表面吸附的影响[J]. 石油学报,2017,38(2):217-226.
CHEN Chen, DONG Zhaoxia, GAO Yuying, et al. Effects of brine composition on quartz surface absorption of polar components[J]. Acta Petrolei Sinica,2017,38(2):217-226.

[20] 谭锋奇,许长福,王晓光,等. 砾岩油藏水驱与聚合物驱微观渗流机理差异[J]. 石油学报,2016,37(11):1414-1427.
TAN Fengqi,XU Changfu,WANG Xiaoguang,et al. Differences in microscopic porous flow mechanisms of water flooding and polymer flooding for conglomerate reservoir[J]. Acta Petrolei Sinica,2016,37(11):1414-1427.

[21] 李勇,颜照坤,李洪香,等. 断坳叠置湖盆岩相古地理研究方法——以沧东凹陷为例[J]. 石油学报,2016,37(增刊 2):39-55.
LI Yong, YAN Zhaokun, LI Hongxiang, et al. Paleogeograghy method of fault depression superposition lake basin:a case study of Cangdong sag[J]. Acta Petrolei Sinica,2016,37(S2):39-55.

[22] LIU Z,LI Y,LV J,et al. Optimization of polymer flooding design in conglomerate reservoirs[J]. Journal of Petroleum Science & Engineering,2017,152:267-274.

(收稿日期 2017-03-24 改回日期 2017-11-23 编辑 王培玺)