计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(8):9−15 [doi: 10.15888/j.cnki.csa.005985]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Spark 的油藏数据挖掘与分析①

武志军, 夏盛瑜, 王　鹏

(中国石油大学(华东) 计算机与通信工程学院, 青岛 266580)

摘　要: 为了方便油藏数据特征的分析和石油的勘探开发过程, 本文利用 Spark 并行计算框架分析油藏数据, 并通过数据挖掘算法分析油藏属性之间的潜在关系, 对油藏的不同层段进行了分类和预测. 本文的主要工作包括: 搭建 Spark 分布式集群和数据处理、分析平台, Spark 是流行的大数据并行计算框架, 相对传统的一些分析方法和工具, 可以实现快速、准确的数据挖掘任务; 根据油藏数据的特点建立多维异常检测函数, 并新增渗孔比判别属性 *Pr*; 在处理不平衡数据时, 针对逻辑回归分类提出交叉召回训练模型, 并优化代价函数, 针对决策树, 提出 KR-SMOTE 对小类别样本进行过采样扩充, 这两种方法都可以有效处理数据不平衡问题, 提高分类精度.

关键词: Spark; 数据挖掘; 异常点检测; 不平衡数据; 分类

引用格式: 武志军,夏盛瑜,王鹏.基于 Spark 的油藏数据挖掘与分析.计算机系统应用,2017,26(8):9–15. http://www.c-s-a.org.cn/1003-3254/5985.html

Reservoir Data Mining and Analysis Based on Spark

WU Zhi-Jun, XIA Sheng-Yu, WANG Peng

(Computer and Communication Engineering, China University of Pertroleum, Qingdao 266580, China)

Abstract: In order to improve the analysis of reservoir properties and oil exploration and development process, this paper analyzes data and finds relationships between reservoir properties using Spark parallel computing framework and data mining algorithm, and classifies and predicts different reservoir segments. The main work in this paper includes: building the Spark distributed clustering and data processing and analysis platform, Spark being a popular big data parallel computing framework, which can achieve fast and accurate data mining tasks compared with some traditional analysis methods and tools; establishing a multidimensional outlier detection function according to the characteristics of reservoir data and adding a new discriminant attribute Pr; proposing a cross-recall training model and optimized cost function for logistic regression classification in dealing with the imbalanced data. KR-SMOTE is used to oversample for decession tree classification that both improve the classification precision.

Key words: Spark; data mining; outlier detection; imbalanced data; classification

近年来, 随着数字油田、智慧油田和物联网的快速建设, 油田信息化得到了飞速发展, 并且随着石油勘探和开发工作的不断深入, 油田数据已经呈爆炸式增长趋势. 面对数据规模庞大、数据形式多样、数据结构复杂等问题, 传统数据库和分析预测方法已无法满足现在的需求, 由此数据挖掘相关技术就被运用到油田的生产实践中, 国内外一些知名石油公司也纷纷开始研究和利用数据挖掘算法, 为企业创造更多的经济效益[1]. 同时, 互联网行业的高速发展也涌现了一批大数据分析工具和技术, 比如 Hadoop, HBase, Hive、Mahout、Spark 等集群数据存储和分析框架, 这些框架已被广泛应用于互联网各大企业. 利用数据挖掘算法

① 收稿时间: 2016-12-09; 采用时间: 2017-02-15

来处理油藏数据, 可以从大量的原始数据中提取各种信息和知识. 目前, 对油藏数据的研究主要包括油田大数据分析平台及体系架构的构建[2,3], 神经网络预测油田产量[4-6], 支持向量机进行油藏历史拟合, 生产性能评估及采收率预测[7,8], 但这些研究都主要集中在模型构建以及理论体系研究, 未能结合实际应用框架进行数据分析. 因此, 本文主要利用 Spark 数据分析框架分析油藏数据, 并通过数据挖掘分类算法挖掘油藏参数的关系, 训练油藏地层分类模型和评估模型性能.

1 Spark 数据分析平台

Spark 是由 UC Berkeley AMP 实验室开发的开源通用并行云计算平台, 是基于 MapReduce 思想实现的分布式内存计算, 拥有 Hadoop MapReduce 所具有的优点[9,10], 但不同的是运算中间输出结果能存储在内存中, 从而不再需要读写 HDFS, 通过减少数据读取的磁盘 IO 开销来提高数据处理速度. Spark 作为一个分布式框架, 支持内存计算、迭代处理、流处理以及图计算, 所以非常适合于交互式数据分析和数据挖掘中的迭代运算.

如图 1 所示是 Spark 生态圈, 主要分为四层, 主要以 Spark Core 为核心, 从数据持久层(Hadoop 分布式文件系统 HDFS、分布式数据库 HBASE 等)读取数据, 并通过资源管理器(Spark 自带的 Standalone、MESOS、Hadoop YARN 等)调度 Job 完成 Spark 的计算任务, 同时 Spark Core 也负责内存管理, 容错机制等. Spark 的计算任务对应于不同的应用组件, 包括即席查询 Spark SQL、实时数据流处理 Spark Streaming、通用机器学习库 MLlib、图的并行处理库 GraphX 等.

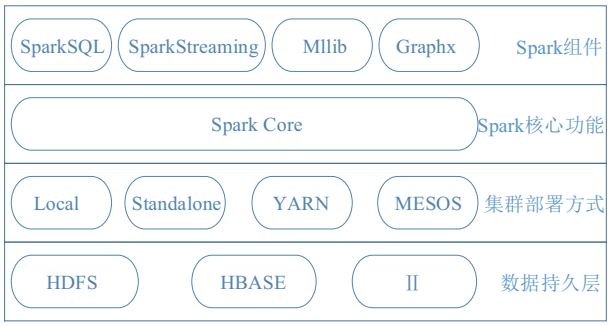

图 1 Spark 生态圈

MLlib(Machine Learning library), 是 Spark 提供的通用的机器学习库, 包含了许多常见的机器学习和统计算法, 大大简化大规模机器学习时间, 主要包括: 聚类、分类与回归、协同过滤、奇异值分解、主成分分析、特征提取和转换、随机梯度下降法等, 这些丰富的算法库为进行油藏数据分析提供了非常有力的基础条件, 本文主要通过 Spark 的 MLlib 来实现油藏数据的处理和研究工作.

2 油藏数据

2.1 油藏参数特征

为了提高油井的开采率和单井产量, 降低勘探开发生产成本, 需要对地下油藏的参数有更精确的描述和理解[11]. 油藏参数直接影响着生产井的单井产量及采收率, 油藏数据包含很多种, 如原油粘度、孔隙度、渗透率、含水饱和度、地层压力等, 每一个属性都能反映油藏的某一特性, 不同的油藏属性之间或者可能存在某种联系[12,13]. 目前油藏参数的研究内容主要包括油藏地质建模仿真、虫洞模型或者历史拟合研究[14-16], 但是由于油藏地下环境复杂难以一窥全貌, 加之油藏参数种类非常多, 导致油藏建模仿真对时间和技术的要求都非常高, 不适合在油田开发周期或短期研究中使用.

油藏参数的种类非常多, 在对油藏分类的时候无法统一标准, 通过单一参数对油藏进行分类只能针对某些特殊的油藏, 缺乏普遍性. 孔隙度代表油藏的存储能力, 渗透率表示流体的通过能力. 在油藏的参数中, 孔隙度高、渗透性好、含油饱和度高的岩层, 相应油井的开采难度会相对较低, 产量也会相应提高, 相反则开采难度就会比较大, 开采效率和采收率就会下降, 就需要采用一定的增产措施. 因此如何确定油藏参数中的关键因素和不同油藏参数之间的关系, 以及对地层精确分类是本文的主要研究内容.

2.2 数据挖掘算法

数据挖掘是从不完整或者有噪音的数据中挖据出未知的且有用的信息或者模式, 主要功能包括分类分析、关联规则分析、聚类分析、预测建模等. 本文主要对油藏地层分类, 通过回归模型和决策树模型进行分类, 找出油藏参数之间的关系, 并预测地层的类别.

2.2.1 逻辑回归分类

逻辑回归本质是线性回归, 但由于线性回归的鲁棒性很差, 不适用于定性预测值的预测, 所以对特征结果加入了一层函数映射, 将线性回归的结果通过函数 $g(z)$ 进行计算, 即 $h(x)=g(f(x))$, $g(z)$ 一般使用使用

Sigmoid 函数, 可以将结果数据范围压缩到一定的范围, $g(z)=\frac{1}{1+e^{-z}}$, 因此逻辑回归模型的方程可以写为:

$$h(x)=\frac{1}{1+e^{-\theta^{\mathrm{T}}\mathrm{X}}} \tag{1}$$

θ是各个特征的权值参数矩阵, X 是特征参数矩阵, 逻辑回归通过减小预测范围来降低对数据的敏感度. 对于二分类问题使用回归模型, 可以设定一个阈值, 当 $h(x)\geqslant 0.5$, 则预测样本为正样本, 当$h(x)\leqslant 0.5$, 则预测样本为负样本. 逻辑回归是一种监督学习方法, 设训练数据集中特征数据 $x=\{x_1,x_2\cdots x_m\}$, 标签数据 $y=\{y_1,y_2\cdots y_m\}$, m 为样本个数, 如果每个样本有 $n+1$ 个特征, 即$x\in\begin{bmatrix}x_0\\x_1\\\cdots\\x_n\end{bmatrix}$, x_0 为截距, 一般用 1 表示.

传统逻辑回归模型通过最大似然方法训练逻辑回归模型, 得到模型最优的θ权值参数. 模型的训练过程就是寻找最小的代价函数, 逻辑回归模型的代价函数可以表示为:

$$L(\theta;x,y)=\sum\frac{1}{2}(h(x)-y)^2 \tag{2}$$

对于求使代价函数最小的θ算法主要有梯度下降、共轭梯度法、拟牛顿法、BFGS 等, 本文针对多分类问题使用 LBFGS[17]求解最小$L(\theta;x,y)$, 收敛快.

2.2.2 决策树分类

决策树作为一种分类算法的优势在于模型构造过程不需要任何领域知识或参数设置, 因此方便了非领域研究人员的使用, 而且能够更好地进行领域知识发现, 已被广泛应用于医学、生物学、商业等诸多领域.

决策树是一个树形结构, 其内部节点表示一个样本特征, 最后的叶子节点则表示一个类别, 有向边表示一次属性划分. 决策树是一个局部最优算法, 从根节点出发, 每次选择最优的特征作为划分准则划分数据样本, 直到满足判定条件后划分类别. 主要步骤如下[18]:

①特征选择. 选取局部最优特征, 最优的判别标准则是通过该属性划分后每个分区的纯度, 分区的数据纯度越高, 则当前划分规则越好, 划分后的数据会尽可能属于同一个类别. 纯度的判别标准主要有信息熵、基尼指数和方差, 一般信息的期望越小, 信息增益越大, 纯度也就越高.

②生成决策树. 决策树的生成方法主要有 ID3、C4.5 和 CART, ID3 就是每次特征划分时选取增益率最大的特征属性进行划分, 而 C4.5 使用最大信息增益比作为划分标准进行特征选取, 可以避免 ID3 偏向多值属性划分, CART 使用 gini 指数作为纯度判别指标.

③剪枝. 决策树模型的训练很容易出现过拟合现象, 为了提高模型的泛化能力, 可以通过剪枝来调节决策树的复杂度. 剪枝策略包括预剪枝和后剪枝, 预剪枝是在特征划分前进行估计, 若该划分不能提高决策树的泛化能力, 则停止划分并标记当前节点为叶子节点; 后剪枝则是在决策树生成之后, 自底向上遍历, 若当前节点替换为叶子节点之后能够提高泛化能力, 则替换该节点. 设置决策树的最大深度和最大划分数也可以有效地避免过拟合.

决策树的使用有助于我们直观地分析油藏参数, 挖掘油藏参数之间潜在的规律, 并从新的角度去研究油藏特征.

2.2.3 评价指标

分类算法采用常用的判别指标: 准确率(Precision)和召回率(Recall). 准确率表示判定为类别 A 中真正为 A 类别的数据所占的比例, 用来评价结果的质量, 召回率则表示所有 A 类别数据中正确划分为 A 类别的数据所占的比例, 用来评价结果的完整性. 准确率和召回率之间相互影响, 一般情况下, 如果准确率高则召回率就低, 否则相反. 在类别不平衡数据集中, 单一判别指标并不能简单地评价模型的好坏. 本文采用 $F1$ 来衡量模型的性能:

$$F1=\frac{2*P*R}{P+R} \tag{3}$$

P 是准确率, R 是召回率, $F1$ 同时考虑准确率和召回率, 只有当两者都比较大时, $F1$ 的取值才会大, 因此 $F1$ 越大, 则该模型的预测性能越好.

3 数据预处理与算法优化

数据预处理是数据挖掘最重要的前提, 也是数据挖掘和模型训练能否成功的关键步骤. 本文选取的油藏特征数据主要包括测井深度、砂层厚度、孔隙度、渗透率、含油饱和度.

3.1 多维数据异常点检测与处理

由于油藏数据采集是地下作业, 而且上述特征参数是通过测井曲线解释而来, 设备故障、数据传输或人为操作误差等原因, 会造成数据集中存在偏离真实数据较大的数据点, 称为异常点. 异常数据的存在对我

们训练模型带来了极大的不确定性, 如果训练数据集中存在噪声数据, 可能会导致训练模型的过度拟合, 会影响测试数据集预测的准确率.

当数据非常多时人工检错就比较繁琐, 拉依达准则是常用的异常点判别统计分析方法, 也称为3σ准则. 如果数据满足正态分布, 则$p(|x-\mu|>3\sigma)\leqslant 0.003$, x 表示当前数据点, μ和σ分别是数据集的均值和标准差, p 也就是距离平均值超过3σ的概率. 由于出现$x\leqslant\mu-3\sigma$或$x\geqslant\mu+3\sigma$的概率非常小, 因此可以把这些数据当做异常数据剔除掉.

但拉依达准则只适用于服从正态分布的单一属性, 多维特征之间会存在某种隐含的关系, 所以不能简单地根据某个参数的阈值来判定某一条记录为异常数据. 已知高斯分布的密度函数为:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{x-\mu}{2\sigma^2}\right) \tag{4}$$

为了考虑油藏数据的所有特征参数, 我们通过高斯分布的密度函数建立多维异常点判别函数:

$$p(x)=\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi\sigma_j}}\exp\left(-\frac{x_j-\mu_j}{2\sigma_j^2}\right) \tag{5}$$

j 表示第 j 个特征值, 若$p(x)<\varepsilon$, 则认为该条记录为异常数据, 需要剔除. 基于密度的判别法更加直观, 可以检测出局部异常, 通过每个特征的密度函数连乘的方式实现多维数据的异常检测.

同时, 研究表明[19], 孔隙度与渗透率之间存在很好的指数相关性$y=ae^{bx}$, 当某些地层无法获取岩心数据时可用来预测油藏地层的渗透率. 根据孔隙度和渗透率的关系, 我们在数据预处理时新增了一个判别属性渗孔比:

Pr=渗透率/ln(孔隙度)

该变量可以有效地防止孔隙度和渗透率同时出现异常的情况, 提高检测效率和判别函数的鲁棒性.

3.2 不平衡数据

在数据挖掘的训练集中, 可能会存在某些类别的样本数远大于其他类别, 导致分类器的多数类别的分类精度高而少数类的分类精度低.

在油藏数据中, 由于石油生成的特殊条件, 地层类别: 水层、干层、油层、油水同层的数据不会均匀分布, 在本文使用的油藏数据集中, 类别的所有数据比例约是 2.2:2.3:2.0:1.1, 很明显第四个类别的数据相对而言比较少. 针对逻辑回归和决策树分类算法原理的不同, 本文采用两种不同的改进策略.

(1) 逻辑回归分类. 考虑到过采样容易导致过拟合和欠采样会丢失多数类重要信息的缺点, 故采用代价函数重构的方法从算法角度解决逻辑回归的不平衡数据问题. 传统逻辑回归分类的代价函数为$L(\theta;x,y)=\sum\frac{1}{2}(h(x)-y)^2$, 该代价函数只适用于类别均匀分布的数据集, 从函数中看出当样本数量增加时, 多数类别样本的误差也会累积, 而少数类别样本对代价函数的影响就会减少, 这使得该代价函数训练出来的权值偏向于多数样本, 最后少数样本的分类精确度会大大降低.

本文对逻辑回归提出改进的代价函数, 对多数类样本赋予惩罚权值, 少数类样本赋予奖励权值. 采用交叉召回权值训练, 首先将数据集分为训练集、验证集和测试集, 用验证集匹配训练集得到的模型, 改进的代价函数 LR:

$$LR(\theta;x,y)=\sum\omega_i\frac{1}{2}(h(x_i)-y_i)^2 \tag{6}$$

ω_i = 1 - R_i, ω_i是每个特征权值也就是惩罚因子, R_i则是验证集中该特征的召回率, 召回率表示该模型能正确对类别分类的概率, 用未召回率加权则可以实现各类别识别率的平衡. 通过循环训练, 直到满足最大运行次数或召回率的误差$|R_i-R_{i-1}|<\varepsilon$, ε是误差阈值.

(2) 决策树分类. 决策树在剪枝和信息增益判别阶段, 会剪除少数样本的信息或者忽略少数类别样本的信息熵增益, 导致少数类别样本被错误分类到多数样本中. 因此为了保证数据集类别的均匀分布, SMOTE 是常用的不均衡数据集过采样方法, 是在距离比较近的 N 个少数类别样本之间进行插值, 生成 N 个新的少数类别样本, 即$x_{new}=x+rand(0,1)*\left(x^{(i)}-x\right)$. $x^{(i)}$是离随机点 x 近邻中的第 i 个点, 但该方法无法区别少数样本, 而且容易添加重复冗余数据或产生边界值, 所以不一定能提高分类的精确度. 针对这些问题本文提出一种改进的 SMOTE 算法(KR-SMOTE), 因为是多维数据分析, 所以数据点都用向量表示, 算法步骤如下:

① 确定少数类别样本扩充的数量 num=最多类别样本数-最少类别样本数.

② 首先对少数类别样本使用 k-means 进行聚类, 记录每个簇的聚类中心点$c=[x_1,x_2\cdots,x_n]$, n 为样本的特征数, 本文只有一个类别是少数样本类别, 所以只有

一个聚类中心.

③ 如果只有一个聚类中心, 则从聚类中心 c 的 K 个最近的点和 K 个最远的点中随机选择两个点 p_c 和 p_f, 然后随机取两个点之间的某个点作为新增点 $p_n = rand(0,1)*|p_f - p_c|$.

如果聚类中心数大于 2, 则随机选择两个聚类中心 p_1 和 p_2, 然后选择 p_1 连线中 p_2 中任意一点 $p_n = rand(0,1)*|p_1 - p_2|$ 作为新的数据点. 两种方法计算公式一样, 只是所选取的两个点有所不同.

④ 执行③直到所选取的点的个数等于 num.

该方法通过聚类中心随机找到新增点, 保证了新增数据不会出现靠近所属类别样本的边缘数据, 另外针对少数类别中是否存在多个不同的类别做了不同的处理, 可针对异类少量样本进行数据扩充.

4 实验分析

4.1 Spark 平台搭建

在进行数据分析之前首先需要构建 Spark 分布式集群, 搭建符合实际生产应用场景的开发环境. 分布式集群的搭建需要多台计算机, 本文的实验测试环境是通过三台主机搭建的集群环境. 如表 1 所示, 是 Spark 集群的搭建环境. 本集群的部署方式是 Spark on YARN, 在一个集群中同时部署 Hadoop 和 Spark, Hadoop 为 Spark 提供分布式文件文件系统.

表 1 Spark 集群运行环境

主机名	操作系统	内存	处理器	硬盘	Hadoop版本	Spark版本
Master	Ubuntu 14.04	6 G	4核	40 G	2.7.3	2.0.1
Slave1	Ubuntu 14.04	4 G	2核	40 G	2.7.3	2.0.1
Slave2	Ubuntu 14.04	4 G	2核	40 G	2.7.3	2.0.1

Spark on Yarn 集群模式, 一方面可以同其他框架共享集群资源, 适合集群中部署多个框架系统, 互不干涉, 可维护性和扩展性非常好, 另外可以提高集群资源的利用率, 通过 YARN 管理集群资源, 实现按需分配内存、CPU 等, 支持多任务并发执行, 可用性也非常高.

4.2 模型训练与分析

实验油藏数据格式如表 2 所示, 地层类别 1, 2, 3, 4 分别表示水层、干层、油层和油水同层. 首先将数据集到 hdfs 中, 然后在 Spark 中使用异常点判别函数对油藏数据进行过滤, 之后得到整个数据集的各个特征分布情况如图 2 所示, 图中各特征数据整体分布比较均匀, 虽有几个离群的点但并没有超出正常取值的范围, 并不认为是噪声.

表 2 油藏数据格式

井名	测井深 (m)	砂层斜厚 (m)	孔隙度 (%)	渗透率 (毫达西)	含油饱和度(%)	地层类别
12A-1	821.8	3.7	37	679.8	12	1
12A-1	871	9.8	34.9	717.8	8	1
12A-1	883.1	13.7	39.7	1694.5	15.1	1
12A-2	1638.5	1.5	31.338	273.838	23.392	3
12A-2	1640	1.1	17.4	245.5	13.7	2
12A-3	1664.9	6.2	36.4	680	53.1	3
12A-3	1685	7.4	31.423	357.29	30.6	4
12A-4	1455	1	34.3	128.8	24.6	3
12A-4	1480.3	1	31.1	48.1	0	2
12A-4	1488.9	3.7	30.9	474.5	11.9	4

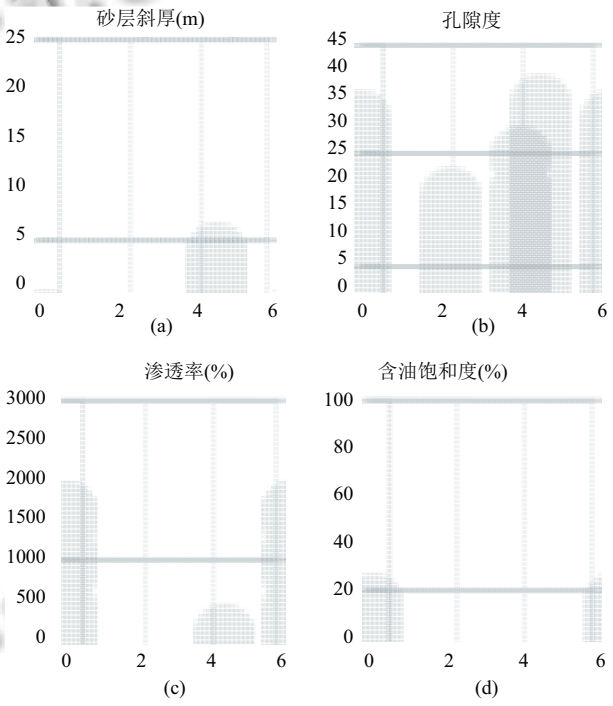

图 2 数据集分布

孔隙度与渗透性数据通过线性拟合后可得到线性逻辑回归模型, 如图 3 所示.

$$y = 0.0402e^{0.2723x} \tag{7}$$

$R^2 = 0.6506$, 拟合的程度还是比较高的, 该指数关系可用来预测油藏地层的渗透率或孔隙度, 解决数据缺失问题, 完善数据集.

本文中分类算法的主要是通过 Spark 编程实现和运行, 并根据 MLlib 库进行扩展.

为了解决数据不均衡问题, 分类模型训练前我们将数据集 15182 条记录按照 7:2:1 随机拆分成训练

集、验证集和测试集. 首先建立回归分类模型, 并对比对使用交叉召回训练前后模型的性能, 多分类算法实现基于 Spark MLlib 中的 LogisticRegressionWithLBFGS 算法. 改进后的逻辑回归分类的代价权值为[0.201, 0.215, 0.174, 0.617], 如表 3 所示, 是测试集验证该模型的混淆矩阵(confusionmatrix), 从表中看出每个类别的准确率和召回率, 图 4 是改进前后各类别的 F1 指标对比, 很明显改进后第四个类别的分类精度明显增加, F1 增加了 18.32%, 而且整体的分类精度都有所提高. 新的逻辑回归模型的整体准确率是 76.48%, 即使第四个类别的样本数相对较少, 但该类别的准确率和召回率与其他 3 个类别相差都不大, 说明交叉召回训练方法可以很好地适应不平衡数据的分类问题.

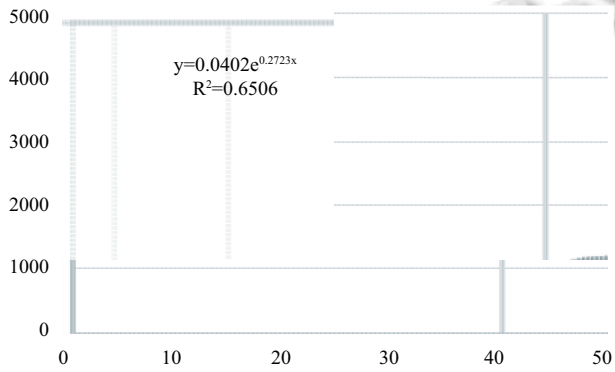

图 3 孔隙度渗透率关系

表 3 改进的逻辑回归分类混淆矩阵

类别	1	2	3	4	召回率
1	340	53	27	18	0.7763
2	41	362	47	16	0.7768
3	38	31	303	20	0.7730
4	27	14	25	156	0.7027
准确率	0.7623	0.7870	0.7537	0.7429	0.7648

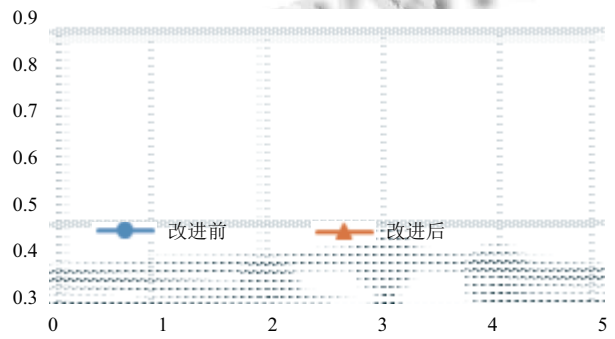

图 4 逻辑回归改进前后各类别 F1 值

逻辑回归分类整体性能还不错, 但离能正确的判别油藏性质还有一定的差距, 整体准确率有待提高, 而且逻辑回归模型的可解释性比较差. 因此我们建立比较直观的决策树分类模型.

为了分析不平衡数据解决方法的有效性, 决策树的数据集分为三类: 原始不均衡数据集、SMOTE 过采样数据集和 KR-SMOTE 过采样数据集, 决策树模型的训练是使用 Spark MLlib 中的 DecisionTreeModel, 经过一定的参数调优后选取各自准确率最高的参数组合.

图 5 所示是不同数据集下决策树的不同类别 F1 曲线, 本文的 KR-SMOTE 算法对决策树处理不平衡数据是有效的, 相对原始数据中第四个类别的 F1 明显有所提高, 并且优于传统的 SMOTE 算法.

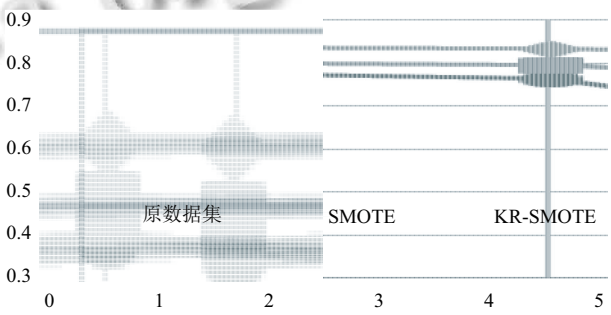

图 5 不同数据集下决策树 F1 值比较

表 4 是 KR-SMOTE 测试集的决策树混淆矩阵, 整个测试集的准确率达到 83.20%, 相比回归分类模型准确率 76.48% 大大提高, 而且决策树是基于划分规则实现数据集的, 更容易解释和理解.

表 4 混淆矩阵

类别	1	2	3	4	召回率
1	360	38	26	14	0.8219
2	37	394	18	17	0.8455
3	24	30	322	16	0.8214
4	7	13	15	187	0.8423
准确率	0.8411	0.8295	0.8451	0.7991	0.8320

所以, 综上所述, 不管是逻辑回归分类的代价函数优化和还是决策树分类的 KR-SMOTE 过采样都可以有效地解决油藏数据中数据不均衡问题, 提高底层分类的性能, 可根据实际情况使用不同的解决策略.

5 结论

随着油田信息化的飞速发展, 传统的数据分析方法和工具已不在符合当前的数据处理要求. 为了方便油藏属性的分析和石油的勘探开发过程, 本文搭建了 Spark 分布式集群并利用数据挖掘算法中的分类算法:

逻辑回归和决策树来分析油藏数据, 进行知识挖掘. 根据油藏数据的特点建立多维异常检测函数, 并在解决油藏不平衡数据时, 针对逻辑回归分类和决策树的算法不同, 分别提出了代价函数优化和 KR-SMOTE 过采样样本扩充两种策略, 从实验数据分析可知, 这两种方法都可以有效处理数据不平衡问题, 提高分类精度. 本文只是对油藏数据的地层分类进行了研究, 分类方法和数据处理比较简单, 对于数据挖掘中其他预测算法还未进行分析和研究, 尤其是各油藏属性对油田产量的影响将是本文以后的研究问题之一.

参考文献

1 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. 中国石油和化工, 2015, (1): 49–51.

2 段泽英, 蔡贤明, 滕卫卫, 等. 大数据分析技术在油田生产中的研究与应用. 中国管理信息化, 2015, 18(18): 64–65. [doi: 10.3969/j.issn.1673-0194.2015.18.046]

3 Brulé MR. The data reservoir: How big data technologies advance data management and analytics in E&P. SPE Digital Energy Conference and Exhibition. The Woodlands, Texas, USA. 2015.

4 邢明海, 陈祥光, 王渝. 应用模糊神经网络预测油田产量. 计算机仿真, 2005, 22(2): 187–190.

5 张方舟, 严胡勇, 杨立全, 等. 改进型灰色神经网络模型在油田产量中的应用. 计算机技术与发展, 2013, 23(6): 241–244, 248.

6 杨婷婷. 基于人工神经网络的油田开发指标预测模型及算法研究[硕士学位论文]. 大庆: 东北石油大学, 2013.

7 Zhou QM, Dilmore R, Kleit A, *et al*. Evaluating gas production performances in marcellus using data mining technologies. Journal of Natural Gas Science and Engineering, 2014, (20): 109–120.

8 Lee BB, Lake LW. Using data analytics to analyze reservoir databases. SPE Annual Technical Conference and Exhibition. Houston, Texas, USA. 2015.

9 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Proc. of the 6th Conference on Symposium on Opearting Systems Design & Implementation. San Francisco, CA, USA. 2004. 137–150.

10 White T. Hadoop: The definitive guide. 3rd ed. O'Reilly Media, Yahoo Press, 2012.

11 田景文. 地下油藏的仿真与预测[博士学位论文]. 哈尔滨: 哈尔滨工程大学, 2001.

12 赵云胜. 灰色系统理论在地学中的应用研究. 武汉: 华中理工大学出版社, 1997.

13 李武广, 邵先杰, 康园园, 等. 油藏分类体系与方法研究. 岩性油气藏, 2010, 22(2): 123–127.

14 Chugh S, Baker R, Telesford A, *et al*. Mainstream options for heavy oil: Part I-cold production. Journal of Canadian Petroleum Technology, 2000, 39(4): 31–39.

15 Cai YX, Wang X, Hu KZ, *et al*. A data mining approach to finding relationships between reservoir properties and oil production for CHOPS. Computers & Geosciences, 2014, (73): 37–47.

16 Istchenko C, Gates ID. The well-wormhole model of CHOPS: History match and validation. Proc. of the SPE Heavy Oil Conference. Calgary, Alberta, Canada. 2012. 1–9.

17 Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical Programming, 1989, 45(1-3): 503–528. [doi: 10.1007/BF01589116]

18 周志华. 机器学习. 北京: 清华大学出版社, 2016.

19 姚秀云, 张凤莲, 赵鸿儒. 岩石物性综合测定——砂、泥岩孔隙度与深度及渗透率关系的定量研究. 石油地球物理勘探, 1989, 24(5): 533–541.