PetroleumPlanninganddesign, 石油规划与设计 (1)2020,2
ISSN:2661-3794(Print)

基于机器学习的测井大数据一体化应用平台开发研究

赵巍巍
中石化华北石油工程有限公司测井分公司　河南　郑州　450001

【摘　要】近些年来，随着机器学习与其分支深度学习的快速发展，越来越多的模型算法应用到了测井数据的分析预测等领域，但这些应用大多场景单一，且效率低下，缺乏一个集成化的智能平台去满足生产科研中的实际需求。本文采用集成优选机器学习算法为水淹层级别识别、岩性识别、储层评价、测井曲线生成等应用建立统一的初始学习模型集，训练后去除模型集中错误率较高的子模型，再通过参数自动化搜索调优以及投票机制，最终生成集成化最优模型。且从数据清洗预处理、特征选择与降维、模型训练等各个步骤均实现了智能自动化，用户只需设置很少的目标参数即可进行相关数据分析与预测应用。

【关键词】机器学习；测井资料；算法智能集成；一体化平台

测井大数据是石油地质中很重要的共享资源，有着广泛的地质应用，而在实际科研生产中使用机器学习或深度学习方法对测井数据进行分析、预测以及潜在特征的提取已经凸显出越来越重要的作用。但这些应用大多被限制于特定的应用场景，且算法应用支离破碎，流程十分繁琐，数据清洗与预处理、特征选择与数据降维模块严重缺失，导致模型精度偏低，训练过程繁琐复杂，基本无法应用于实际生产与科研。

1 总体设计

1.1 方法研究

1.1.1 系统功能设计

结合一般数据处理分析与机器学习流程，以及测井数据应用处理时的各种实际需求，设计系统功能主要包括用户目标设置、数据清洗与预处理、特征优选与降维、算法优选集成、可视化数据分析等。

1.1.2 开发语言与框架

传统机器学习与深度学习中有很多框架，比如 sk-learn、Tensorflow、Caffe、keras、CNTK 等，采用的开发语言也不尽相同，有 Python、Java、C++ 等。经过综合研究对比，考虑到运行效率与后期可维护性，系统在传统机器学习模块采用 sk-learn 框架，深度学习采用 Tensorflow 框架，开发语言采用 python。前台的用户界面考虑到与 python 的兼容性，最终采用 Pyqt5 作为界面开发框架。

1.1.3 算法择优集成

系统首先将常用的机器学习模型算法（svm、rf、logstic、lr 等约 10 种）与深度学习模型算法（cnn、lstm 等）放入算法池中，依据用户的应用场景自动挑选相应分类、回归或聚类算法，根据用户设置的评估指标从中择优选择若干模型算法，对各个初选算法做参数最优搜索，再对初选算法分别作 bagging 和 boosting 集成，从而选择最优的集成算法。

1.2 软件开发

本节重点介绍测井数据清洗与预处理、特征优选与降维、多模型算法择优集成、数据可视化以及用户 UI 界面开发做介绍，而对各个算法详细原理不做详细介绍。

1.2.1 数据清洗与预处理

由于测井仪器的故障与人为操作不当等因素，测井数据经常会出现缺失或异常值，且各个数据量纲也不统一，这些因素对模型精确度有很大影响。所以首先系统将会利用 sk-learn 对数据做异常点检测、归一化、标准化，也会对缺失值做相应处理，会根据该特征缺失值比例，选择删除该特征或者使用指定数据进行填充。若数据样本有较为严重的样本不均衡情况，则系统会进行上采样或下采样对数据样本均衡化处理。

1.2.2 特征选择与降维

研究发现，机器学习中特征的优化选择与数据维度对模型训练有着很大的影响。系统采用派尔森系数对特征进行过滤选择，同时消除递归特征，再根据得出的特征选择结果确定是否需要降维，若需要，则使用 PCA 算法对特征进行降维，最终得出模型所需要的数据特征。

1.2.3 算法优选集成

系统首先根据应用类别自动从算法池中选出相关算法，使用各个算法默认的参数组开始训练，得出结果后根据用户设置的应用评估指标从中择优选择若干模型算法，同时对各个初选算法做参数最优搜索，再对参数最后后的初选算法分别作 bagging 和 boosting 集成，再对集成模型作参数最优搜索，最终根据测试数据的评估指标选出最优的集成模型，并将其序列化保存。

1.2.4 可视化数据分析

系统中的可视化数据分析模块可以帮助用户加强对数据的理解，以更直观的视角分析数据。系统在数据描述统计、特征选择、算法性能比较评估等许多数据描述场景应用了

matplotlib 与 seaborn 实现多种数据图的展示，为用户做决策提供直观的支持。

1.2.5 用户界面开发

前台的用户界面考虑到与 python 的兼容性，最终采用 Pyqt5 作为开发框架，系统界面设计美观，用户逻辑操作合理易用（图 1）。

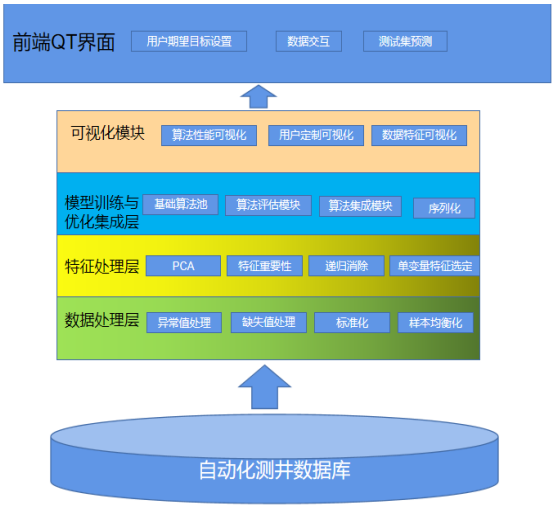

图 1 系统架构

2 系统特点

在深入研究目前目前测井大数据在实际生产科研中的需求与应用后，结合现有机器学习与深度学习模型算法的最新进展，开发了一套基于机器学习的测井大数据一体化应用平台。系统具有以下特点。

（1）具有高度自动智能化与用户可定制化。用户只需设置基本的机器学习评估参数即可进行相关模型训练而后进行数据预测或聚类等实际应用，具体的数据处理、特征选择、算法择优、参数选择等繁琐操作均由系统智能自动识别处理完成，训练过程不需用户干预。

（2）智能选择最优算法与最优参数完成集成，各个算法模块低耦合，方便改进或扩展。利用相关设计模式，使得各个算法模块扩展维护方便，提高后期系统优化效率。

（3）基于 pyqt5 开发，可在多种平台下进行应用。

3 应用实例

3.1 聚类应用 – 水淹层 DBSCAN 聚类分析

数据聚类分析模块优选 DBSCAN 算法对水淹层数据完成了聚类分析，选取的主要特征数据有该层测井曲线中的 gr、cnl、ac、rxo、rls、sp 泥、sp 砂等值以及该层含水率；在对数据进行聚类分析前做了标准化处理，而后系统根据用户设置的分类期望范围与轮廓系数自动选取合适的 eps 与 minsample 参数，最终将数据分为 4 类，在做 PCA 降维处理后在散点图上表示如图 2 所示。

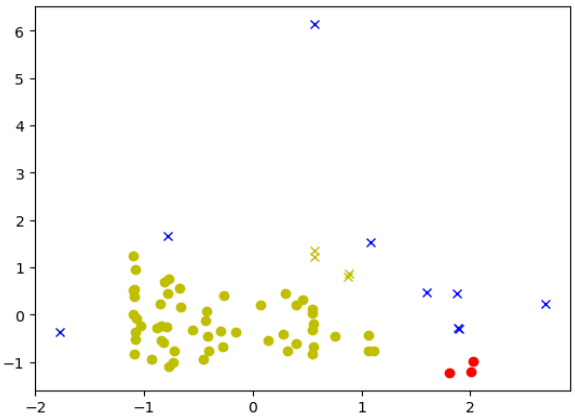

图 2 水淹层样本 DBSCAN 聚类图

3.2 回归应用——水淹层含水率预测

在根据用户设置与需求对数据进行自动清洗与预处理后，回归预测模块从算法池中选出相应回归算法，通过准确率做初选，最终选择 LR、SVM 以及 RFR、ETR、GBR 作为集成算法，随后系统依据验证集误差数据自动从中选出 ETR 与 GBR 作为最终模型，随后自动从相应参数 grid 中选出相关最优参数，如基回归器个数等（图 3）。

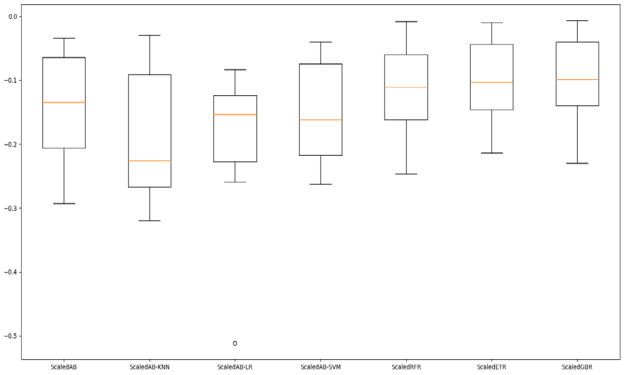

图 3 各个集成算法在水淹层测试集上表现对比

3.3 特征处理应用—水淹层特征影响因素排序

系统根据用户设置使用特征处理模块中的 ETR 算法对数据进行迭代训练，从而对输入数据中的特征重要性做相应的分析处理，得出各个特征对标签值含水率的影响因子大小，如图 4 所示。

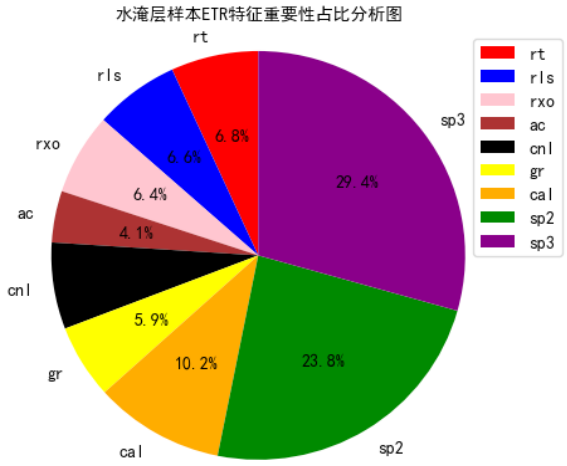

图 4 水淹层各个特征重要性分析

4 结束语

基于机器学习的测井大数据一体化应用平台相较于以前的零散机器学习测井数据应用有着明显的优势，如数据自动预处理、算法自动择优集成等。该系统使得一般的测井科研人员能够真正的将机器学习作为数据分析、预测、处理的有力应用工具，摒弃了之前零碎的、复杂繁琐的应用过程，且在最终模型算法准确率上也有了一定的提高。

【参考文献】

[1] 刘洪，马力宁，黄桢 . 集成化人工智能技术及其在石油工程中的应用 [M]. 北京：石油工业出版社，2008.
[2]（美）BruceEckel.Java 编程思想 第 4 版 [M]. 北京：机械工业出版社，2013.
[3] 裔隽，张怿檬，张目清 .Python 机器学习实战 [M]. 北京：科学技术文献出版社，2018.
[4]（美）Adam Freeman.HTML5 权威指南 [M]. 北京：人民邮电出版社，2018.
[5] 雍世和，张超谟 . 测井数据处理与综合解释 [M]. 东营：中国石油大学出版社，2007.
[6] 郑泽宇，顾思宇 .TensorFlow 实战 Google 深度学习框架 [M]. 北京：电子工业出版社，2017.
[7] 李光军，王卫，王慧萍 .Logik 测井微机解释系统开发技术分析 [J]. 石油天然气学报，2011（8）：91-95，3.