PHÂN TÍCH CƠ SỞ DỮ LIỆU TRONG DẦU KHÍ SỬ DỤNG DATA ANALYTICS

Nguyễn Thị Thí Linh, Đào Thị Minh Huyền, Trần Thái Triều, Thái Bá Ngọc
Khoa KT Địa chất & Dầu khí, Đại học Bách Khoa, ĐHQG-HCM
tbngoc@hcmut.edu.vn

Tóm tắt

Xu hướng gần đây của Big Data đã phát triển các phương pháp phân tích mới để xử lý hiệu quả dữ liệu đa chiều và trực quan hóa chúng để khám phá các mỏ mới. Mục tiêu chính của nghiên cứu này là áp dụng ba phương pháp được sử dụng trong phân tích dữ liệu (Data Analytics): hồi quy tuyến tính, hồi quy tuyến tính với lựa chọn tính năng và mạng Bayesian, cho các tập dữ liệu tầng chứa. Sử dụng cơ sở dữ liệu tầng chứa thương mại, nghiên cứu sẽ tạo và thử nghiệm ba mô hình này để dự đoán khả năng thu hồi dầu và khí cuối cùng. Trong số các mô hình được thiết kế để ước tính các hệ số thu hồi, các mô hình hồi quy tuyến tính được tạo bằng các biến được chọn với phương pháp lựa chọn đặc trưng tuần tự (sequential feature selection) là tốt nhất, cho thấy mối tương quan mạnh mẽ giữa giá trị thực tế và dự đoán khả năng thu hồi của tầng chứa. So với mô hình này, mô hình mạng Bayesian và mô hình hồi quy tuyến tính đơn giản cho kết quả khá kém.

Từ khóa: data analytic, big data, hồi quy tuyến tính, hồi quy tuyến tính với lựa chọn tính năng và mạng Bayesian.

ANALYZING DATABASES IN PETROLEUM USING DATA ANALYTICS

Nguyễn Thị Thí Linh, Đào Thị Minh Huyền, Trần Thái Triều, Thái Bá Ngọc
Faculty of Geology&Petroleum Engineering, HCM University of Technology, VNU-HCM
tbngoc@hcmut.edu.vn

Abstract

The recent trend of Big Data has given rise to novel analytic methods to effectively handle multidimensional data, and to visualize them to discover new patterns. The main objective of this research is to apply three methods used in data analytics: linear regression, linear regression with feature selection, and Bayesian network to datasets with reservoir data. Using a commercial reservoir properties database, we created and tested three data analytic models to predict ultimate oil and gas recovery efficiencies. Among the models designed to estimate recovery factors, the linear regression models created with variables selected with sequential feature selection method performed the best, showing strong positive correlations between actual and predicted values of reservoir recovery efficiencies. Compared to this model, Bayesian network model, and simple linear regression model performed poorly.

Key words: data analytic, big data, linear regression, linear regression with feature selection, and Bayesian network.