Секция «Актуальные проблемы геологии нефти, газа и угля»

Использование классических методов машинного обучения для интерпретации результатов геохимических исследований

Научный руководитель – Большакова Мария Александровна

Новиков Евгений Владимирович
Студент (магистр)
Московский государственный университет имени М.В.Ломоносова, Геологический факультет, Кафедра геологии и геохимии горючих ископаемых, Москва, Россия
E-mail: enovickov@gmail.com

К классу машинного обучения (далее - МО) относятся алгоритмы и системы, качество работы которых увеличивается по мере накопления опыта [2].

Методы классического МО включают в себя два основных раздела - обучение с учителем и обучение без учителя. В первом случае используется обучающая выборка данных - в случае классификации органического вещества (далее - ОВ) это непосредственно его признаки, а также метки - его типы - I, II или III. Во втором случае никаких меток нет, а имеются только признаки - данные водородного индекса и температура максимального выхода газообразных углеводородов.

Для исследования были использованы пиролитические данные множества образцов Западной Сибири.

В первую очередь использовались методы обучения без учителя (кластеризация). Было использовано 3 подхода: K-средних, OPTICS, агломеративной кластеризации. Подробности об алгоритмах читатель может узнать в [1], [2] или на сайте с документацией библиотеки scikit-learn [3]. Наилучшим образом сработали методы K-средних и агломеративной кластеризации, но из-за отсутствия информации о направлении кривых на подложке, а также из-за не совсем репрезентативного набора данных, алгоритмы все-равно не могут обнаружить реальную закономерность распределения данных. Для борьбы с этим было решено перейти к другому подходу, а именно к использованию обучения с частичным привлечением учителя. Идея состоит в том, чтобы использовать методы кластеризации для первичного определения меток и использования их для дальнейшего обучения алгоритма обучения с учителем. Однако, здесь есть свои подводные камни - исходный набор данных размечается довольно плохо, поэтому был создан синтетический тренировочный набор данных с точками и соответствующими им метками вдоль направлений трансформации каждого типа ОВ. На новом наборе было проведено обучение различных алгоритмов, таких как полиномиальная регрессия, метод опорных векторов, алгоритм ближайших соседей, деревья решений, а также ансамблевых методов - градиентного бустинга и случайного леса. Лучший результат классификации показал случайный лес.

Впоследствии было написано приложение с графическим интерфейсом, которое способно загружать/сохранять/выгружать данные пиролиза, визуализировать модифицированную диаграмму Ван - Кревелена и несколько других геохимических индексов, а также производить интерпретацию каждого индекса и записывать результат в таблицу для каждого образца.

Источники и литература

1) Рашка С., Мирджалили В. Python и машинное обучение. Пер.,СПб., 2020.
2) Флах П. Машинное обучение. Пер., М., 2015.
3) scikit-learn.org/stable/user_guide.html – библиотека МО для языка Python (англ)