

## ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБРАБОТКИ ГЕОФИЗИЧЕСКИХ ДАННЫХ

**Шахатова Алия Талгатовна**

*shakhatovaa@list.ru*

докторант Евразийского Национального Университета им Л.Н.Гумилева,

Нур-Султан, Казахстан

Научный руководитель – Мирғалиқызы Толқын

### Введение

Методы искусственного интеллекта и интеллектуального анализа данных широко используются во многих областях для решения задач классификации, планирования, диагностики, вычислений, прогнозирования и оптимизации.

Одним из таких методов является машинное обучение.

Машинное обучение (ML) - это область компьютерных наук, которая дает «компьютерам возможность учиться без явного программирования» [1]. Статистические или математические методы используются для того, чтобы построить модель из наблюдаемых данных, а не кодифицировать конкретный набор инструкций, которые определяют модель для этих данных [2].

Если рассмотреть использование метода машинного обучения:

Вводят конкретные примеры и создают модель, обобщающую за пределами этих примеров. ML - быстро растущая дисциплина как в теоретическом, так и в прикладном аспектах. SEG недавно провела конкурс ML, и его ежегодное собрание включало многочисленные применения ML и связанных с ними методов. Была целая сессия «Машинное обучение и анализ образов». Байесовское изучение особенностей и изучение словаря в некоторой степени похоже на фазовое слежение и применяются к шумоподавлению сейсмических данных. Supervised ML автоматически выполняет анализ сегментов AVO, фаций, разломов, геологических объектов и даже пород. ML без присмотра был разработан для автоматического выбора скорости. Машина опорных векторов использовалась для удаления ложных данных из подлинных микросейсмических событий. Глубокое обучение было применено к построению предшествующей модели FWI, и было показано, что оно работает лучше, чем традиционные методы расчета временных интервалов несоответствия и идентификации отражателей и дифракторов. В данной статье кратко рассмотрим некоторые из перечисленных выше методов, их успехи и открытые вопросы. Классический анализ ценности информации также может быть использован, при этом полезность определяется взаимно с неопределенностью.

Методы машинного обучения используются для получения навыков на модельных примерах. Ранее невидимые записи могут быть классифицированы с использованием правил, сгенерированных во время обучения. Вместо этого неконтролируемые алгоритмы обучения группируют данные на основе сходства, а не категорий моделей. Кроме того, другой тип подхода основан на концепции «обучения с подкреплением», когда машина изучает свое поведение на основе обратной связи от «внешней среды». В настоящее время все эти подходы находят применение во многих секторах, включая медицинские, социальные и финансовые дисциплины. Число областей применения ML также впечатляюще растет в геонауках, включая геофизику [3]. Примерами приложений являются распознавание и классификация сейсмических фаций, автоматическая интерпретация геофизических данных, каротаж скважин, и в частности электрическая разведка и анализ. Еще одна область, где ML может найти полезные приложения, - это интеграция междисциплинарной информации, такой как сейсмические, электромагнитные, гравитационные и магнитные данные [4]. Процесс интеграции обычно выполняется путем применения взаимосвязанных методов, которые

преобразуют междисциплинарные и гетерогенные измерения в многопараметрическую модель Земли.

Этот процесс может быть реализован с помощью совместного моделирования, одновременной совместной инверсии или, чаще, с помощью комбинации подходов. В принципе, такая интегрированная модель Земли может значительно снизить риск разведки.

В этой статье мы представляем подход, направленный на максимальное использование преимуществ интегрированных геофизических методов и методов машинного обучения. Идея заключается в объединении огромной и разнородной информации как в пространстве данных, так и в пространстве моделей с использованием методов, обычно применяемых в области науки о данных. Целью является снижение риска разведки и оптимизация процесса оценки месторождений.

1. Методология применения машинного обучения для обработки геофизических данных

Геофизические данные могут быть охарактеризованы с точки зрения атрибутов, обычно связанных с соответствующими физическими свойствами. Например, сейсмические данные могут быть охарактеризованы с использованием таких свойств, как средняя амплитуда, средняя энергия, частота и тренды AVO, когерентность, наклон, кривизна и т. д. [5]. Подобные атрибуты обычно используются в электромагнитных методах. Например, в электромагнитных методах морского контролируемого источника (Marine CSEM), тенденция электрической и магнитной амплитуды, фазы может предоставить диагностическую информацию о возможном присутствии заполненных углеводородом резервуаров. Другие геофизические измерения обычно интерпретируются с использованием аналогичного подхода, основанного на атрибутах, как в гравитационных и магнитных методах. Объединение стольких атрибутов обычно требует больших вычислительных усилий, технических ресурсов и профессиональных навыков.

2. Описание рабочего процесса для объединения QUIS (система количественной интеграции) и ML (машинное обучение)

Наш подход направлен на поддержку процесса интеграции посредством применения алгоритмов и процедур, обычно используемых в расширенной аналитике «больших данных» и в области машинного обучения. Рабочий процесс суммирован на рисунке 1. Данный процесс разделен на две дополнительные ветви.



Рисунок 1 Рабочий процесс, объединяющий QUIS (система количественной интеграции) и ML (машинное обучение)

Левая ветвь включает в себя оптимизированную комбинацию итеративного моделирования, алгоритмов ограниченной, кооперативной и совместной инверсии (образуя «Систему количественной интеграции», кратко называемую «QUIS»). Целью этой части рабочего процесса является получение мультифизической модели Земли, учитывающей все данные и характеризующейся набором свойств, таких как сейсмическая скорость, удельное сопротивление и плотность. Пунктирные стрелки показывают, что эта часть рабочего процесса обычно является круглой. Этот цикл может быть вызван обратной связью от одного геофизического домена к другому. Например, результат электромагнитной инверсии (или гравитационного моделирования) может вызвать пересмотр интерпретации сейсмических данных.

«Правильный путь» рабочего процесса состоит в основном в применении статистических, аналитических и обучающих алгоритмов. Он начинается с применения передовых аналитических методов для статистического анализа различных наборов данных и их взаимных корреляций. Затем набор геофизических «признаков» извлекается из данных и параметрических моделей. Эти характеристики могут принадлежать сейсмическим, гравитационным, магнитным и электромагнитным областям как в пространстве данных, так и в пространстве моделей. Красная стрелка на рисунке 1 указывает, что геофизические параметры, полученные через левую ветвь рабочего процесса, используются в качестве геофизических признаков в процессе машинного обучения. Все они вместе с другими атрибутами, принадлежащими пространству данных, согласуются, чтобы сформировать «глобальную матрицу характеристик», включая сейсмические амплитуды, гравитационные и электромагнитные атрибуты, сейсмические скорости, удельное сопротивление, плотность и так далее. Следующим шагом является определение набора обучающих данных, состоящего из помеченных данных. Мультифизические измерения и параметры калибруются по информации из скважины, если она имеется. Эти помеченные наборы данных используются для обучения алгоритмов учащихся. Это могут быть искусственные нейронные сети, машина опорных векторов, случайный лес, байесовские сети или другие методы классификации. Наконец, мы выбираем подходы, которые дают наиболее надежные результаты, в зависимости от конкретной проблемы классификации, качества данных и геологической сложности. Эффективность подхода классификации должным образом проверена с использованием хорошо известных методов, таких как методы перекрестной проверки и матрицы путаницы. Они позволяют сравнивать достоверность различных методов путем построения графиков прогнозируемых и фактических значений в подмножестве помеченных данных, называемом набором тестовых данных. Результаты классификации в конечном итоге отображаются с точки зрения пространственного распределения классов. Например, мы можем создать вероятностные карты распределения нефти и рассола на заданной глубине. Наконец, обе ветви рабочего процесса способствуют снижению риска разведки и / или оптимизации процесса полевой оценки.

#### Заключение

Машинное обучение (ML) может поддерживать рабочий процесс интеграции разнородных геофизических наборов данных в процессе оценки риска разведки и / или полевой оценки. Мы обсудили рабочий процесс, сочетающий подходы статистической и автоматической классификации с геофизическим моделированием / инверсией. Мы применили этот «гибридный подход» к двум междисциплинарным наборам геофизических данных, записанных в разных геологических условиях, получив обнадеживающие результаты в обоих случаях. Конечно, тот же подход может быть применен во многих других секторах геологоразведочных наук, не обязательно ограниченных углеводородной промышленностью.

*Работа поддержана грантом МОН РК по договору № 132 от 12.03.18 «Разработка алгоритмов и встроенного программного обеспечения по определению геоэлектрического разреза для геоинформационной технологии GPR» (ИРН AP05133922).*

### Список использованных источников

1. Bishop, C. M., 2006. Pattern recognition and machine learning: Springer.
2. Castagna, J.P.; Backus, M.M., 1993. Offset Dependent Reflectivity - Theory and Practice of AVO Analysis. Society of Exploration Geophysicists. ISBN 1-56080-059-3
3. Aminzadeh, F. and de Groot, P., 2006. Neural Networks and Other Soft Computing Techniques with Applications in the Oil Industry, EAGE Publications.
4. Samuel, A., L., 1959. Some studies in machine learning using the game of checkers, in IBM Journal of research and development.
5. Dell'Aversana, P., and M. Vivier, 2009. Expanding the frequency spectrum in Marine CSEM applications: Geophysical Prospecting, Volume 57, Issue 4, pages 573-590.

УДК 004.43, 004.056.57

### РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ДЛЯ ПАРСИНГА И ВАЛИДАЦИИ БИНАРНЫХ ОБЪЕКТОВ С ПОСЛЕДОВАТЕЛЬНОЙ СТРУКТУРОЙ.

**Штейнбрехт Евгений Вадимович**

*[shtjen.kz@mail.ru](mailto:shtjen.kz@mail.ru)*

Магистрант специальности «6М070400 – Вычислительная техника и программное обеспечение», экспериментальная образовательная программа «Администратор по управлению и защите компьютерных систем и сетей на предприятиях»

факультета информационных технологий,  
ЕНУ им. Л.Н.Гумилева, Нур-Султан, Казахстан  
Научный руководитель – Д.Ж. Сатыбалдина

Валидация файлов согласно официальным спецификациям, выпущенным разработчиками форматов, является важным аспектом обеспечения информационной безопасности. Зачастую вредоносные инъекции происходят в обход официальных стандартов файлов. В связи с этим возможность статически проанализировать файл и выделить в нём расходящиеся с правилами элементы позволяет обнаружить значительное количество вредоносных объектов [1].

В настоящее время активно используются сотни различных форматов, для многих из которых выпущены различные образцы спецификаций. При этом каждый из форматов может иметь десятки версий, отличающихся различными элементами [2]. При этом существующих решений для их валидации существует крайне небольшое количество. Среди них к наиболее известным относятся YARA и Kaitai.

Основное назначение языка YARA – поиск сигнатур внутри файлов. По структуре он находится ближе к классу языков регулярных выражений [3]. Таким образом, программы, написанные на нём, предназначены не для последовательного сканирования и валидации объектов, состоящих из байтовых полей, а для поиска и анализа структур с определёнными характеристиками.

Функционал Kaitai включает в себя парсинг и запись обработанных данных в форматированные структуры. В итоговом результате генерируются объекты, хранящие в себе наборы извлечённых данных [4]. Для непосредственно анализа и валидации бинарных объектов данный язык пригоден относительно слабо.

Таким образом, на данный момент времени отсутствуют продукты для удобного и быстрого создания алгоритмов валидации файлов непосредственно на основе спецификаций. В связи с этим целью исследований, результаты которых представлены в данной работе, является разработка оригинального языка и программного средства для парсинга и валидации бинарных файлов на основе их спецификаций.

На начальном этапе разработки был проанализирован набор файлов различных форматов и их составляющих, которые описаны далее.