Применение машинного обучения по ансамблю решающих правил для вычисления прогноза дополнительного коэффициента извлечения нефти

Ф.В.Краснов, Н.Г.Главнов, А.Н.Ситников

***Аннотация*— Суррогатное моделирование имеет широкий спектр применения в различных индустриях. Хорошо изучены вопросы выбора математической модели для построения суррогатной модели. Но для большинства математических моделей не существует эффективных универсальных процедур для их автоматического применения. Авторы рассмотрели применение двух математических моделей - многомерной линейной интерполяция (МЛИ) и случайный лес (RF) - для построения прогноза дополнительного коэффициента извлечения нефти (КИН) на основе частной суррогатной модели.**

***Ключевые слова*— КИН, коэффициент извлечения нефти, EOR, случайный лес, random forest, многомерная линейная интерполяция, regular grid interpolation, суррогатные модели.**

I. Введение

Согласно [1] концепция создания суррогатных моделей состоит из следующих этапов:

1. Характеристика объекта Z, определяющая свойства объекта в некоторых условиях, может быть описана в виде функциональной зависимости $Z = \Phi(X, Y)$, где переменная X описывает сам объект, а переменная Y задает условия функционирования.
2. Функция Ф является неизвестной, и для ее вычисления проводятся вычислительные эксперименты.
3. Имеется некоторое количество измерений

$$\Xi = \{X_i, Y_i, Z_i = \Phi_i(X_i, Y_i),\ i \in \Re\}, \qquad (1)$$

где значение $Z_i = \Phi_i(X_i, Y_i)$ характеристики Z получено методом M_i для объекта, имеющего описания X_i, в условиях функционирования Y_i.

4. По известному множеству Ξ с помощью тех или иных математических методов анализа и обработки данных строится функция $\Phi^s(X, Y)$, значение которой принимаются в качестве приближенного значения характеристики Z для объекта с описанием X в условиях функционирования Y.

Если все значения в множестве Ξ получены при помощи одной и той же модели M и $\Phi^s(X, Y) \approx \Phi^m(X, Y)$, то построенная функция Φ^s может рассматриваться как «заменитель» (суррогат) функции Φ^m.

В ряде вычислительных экспериментов для решения задач нефтегазовой отрасли используется вышеописанный мета-алгоритм. Например, сначала в гидродинамическом симуляторе производятся расчеты значения функции для определённых «узловых» значений параметров X_i на основании физических законов движения жидкостей в пористой среде M_i, а потом, заданная таким числовым образом функция Ф используется для получения значений функции Y_i либо на более детализированном множестве значений параметров, либо для значений параметров, выходящих за рамки «узловых» значений X_i.

Такой подход, например, используется для расчета дополнительной добычи нефти, полученной в результате применения различных методов увеличения нефтеотдачи (МУН) [15].

Одна из основных причин возникновения описанного выше мета-алгоритма на взгляд авторов заключаются в ограничениях на скорость гидродинамической моделирования. В будущем, когда в любое время любой специалист организации сможет варьировать значения параметров в широком диапазоне и в режиме близком к реальному времени получать искомые значения функции потребность в суррогатных моделях скорее всего отпадет. А пока моделирование производится на дорогостоящих высокопроизводительных кластерах, специалистами за времена, измеряемые в часах, а иногда и днях для одного набора параметров существует потребность в прозорливой подготовке данных которые могут понадобиться в дальнейшем. Так как, потребность в изменении параметров может возникать по несколько раз в день и у самых разных специалистов различных подразделений организации, то применение суррогатного моделирования является

Статья получена 1 сентября 2017.
Ф.В.Краснов, к.т.н., эксперт, ООО «Газпромнефть НТЦ», г. Санкт-Петербург, набережная реки Мойки д.75-79, 190000. krasnov.fv@gazprom-neft.ru, orcid.org/0000-0002-9881-7371, РИНЦ 8650-1127
Н.Г.Главнов, эксперт, ООО «Газпромнефть НТЦ», г. Санкт-Петербург, набережная реки Мойки д.75-79, 190000. (email: Glavnov.NG@gazprom-neft.ru),
А.Н.Ситников, ЗГД по ГИРМ, ООО «Газпромнефть НТЦ», г. Санкт-Петербург, набережная реки Мойки д.75-79, 190000. (email: Sitnikov.AN@gazprom-neft.ru)

насущной необходимостью.

Получаемая суррогатная модель Φ^s, иногда ее называют прокси-моделью [2], [3], превосходит изначальную модель Φ^m по вычислительной силе во много раз, то есть, не требует большого объема вычислительных ресурсов и работает в режиме близком к реальному времени.

Рассмотрим две математические модели, используемые для построения прокси-модели, на примере вычислительного эксперимента по исследованию прироста КИН при организации смешивающего вытеснения путем закачки углекислого газа.).

II. Вычислительный эксперимент

Для получения первой оценки дополнительной добычи нефти от третичных методов увеличения нефтеотдачи часто используются типовые кривые в координатах прокаченный поровый объем и дополнительный КИН (Рис.1 и 2).

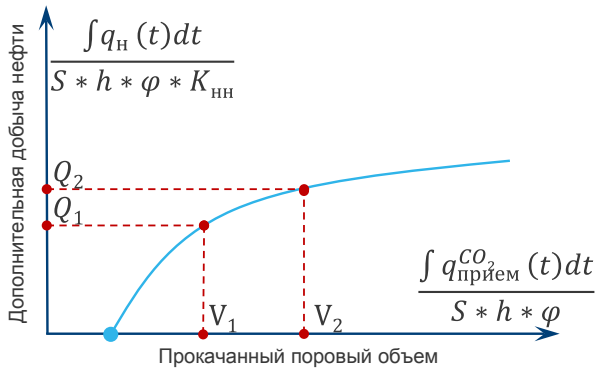

Рисунок 1 Зависимость дополнительной добычи нефти от прокачанного порового объема.

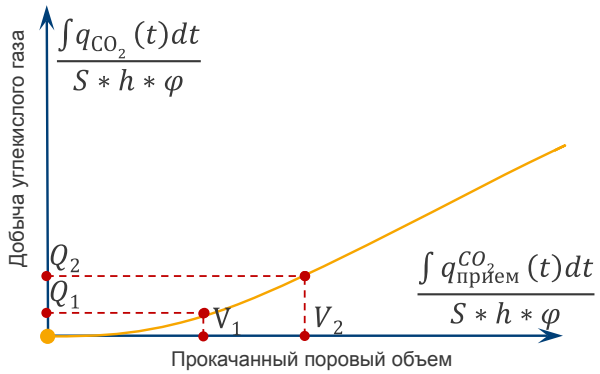

Рисунок 2 Зависимость добычи углекислого газа от прокачанного порового объема.

Данные кривые чаще всего получаются в результате статистического анализа фактически реализованных проектов, либо по упрощенным аналитическим зависимостям и реже по результатам многовариантных расчетов на синтетических гидродинамических моделях. Последний способ получения типовых кривых для технологии попеременной закачки углекислого газа и воды в смешивающемся режиме в пласт был применен авторами в данной статье. Моделирование осуществлялось на базе композиционного симулятора Eclipse 300 (Schlumberger), позволяющего воспроизводить процесс смешивающегося вытеснения. Модель представляет собой сегмент пятиточечного элемента системы разработки, с вертикальными скважинами в углах.

В нашем эксперименте дополнительный КИН рассчитывается при варьировании следующих параметров:

- Свойства нефти (плотность, вязкость, давление насыщения). С целью учета влияния свойств пластовой системы на эффективность вытеснения созданы три модели пластовой нефти с характеристиками, охватывающими весь диапазон (223 объекта) изменения свойств пластовой нефти имеющейся выборки месторождений.
- Фактическое значение остаточной нефтенасыщенности, полученное по результатам потоковых экспериментов в системе нефть-вода, определяет неизвлекаемый при заводнении объем нефти. При вытеснении диоксидом углерода фиксируется снижение остаточной нефтенасыщенности вследствие уменьшения межфазного натяжения между вытесняющим агентом и нефтью, сопровождающееся процессом растворения.
- Неоднородность проницаемости: на основе результатов интерпретации геологофизических исследований разведочных скважин через формулу определения коэффициента Дикстра-Парсонса рассчитаны значения для всех рассматриваемых объектов, для вариации выбрано 4 значения с почти равномерным покрытием всего интервала.
- Относительные фазовые проницаемости (ОФП): для определения узловых значений концевых точек ОФП, величин остаточной водо- и нефтенасыщенности проведено обобщение результатов лабораторных исследований керна. Для охвата моделированием всего диапазона фазовых проницаемостей значения максимальных ОФП по газу и воде были выбраны соответствующими среднему значению, и двум близким к максимуму и минимуму значениям.
- Текущая нефтенасыщенность: учет степени выработки запасов или другими словами текущей обводненности в расчетах производился изменением начальной нефтенасыщенности куба при инициализации модели. Было выделено три значения первый безводный, средняя выработка, и выработанный объект.

Всего сгенерировано 324 гидродинамические модели, на базе которых выполнено 972 расчета (по 3 на каждой). Результаты расчетов группировались в один сводный файл, в котором обрабатывались до 486 типовых кривых.

Ниже приведена статистика затраченного времени на расчет одного варианта, среднее значение которого составило – 90 минут.

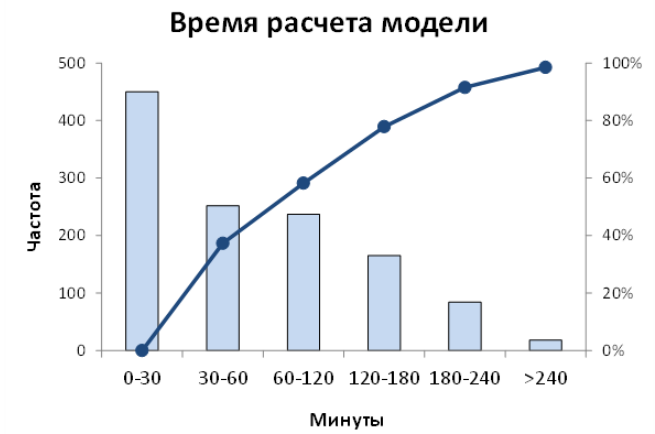

Рисунок 3 Временные затраты на моделирование.

III. Создание прокси-модели на основании многомерной линейной интерполяции

Одним из подходов к созданию прокси-модели является многомерная линейная интерполяция, описанная в работе [4]. Для ее понимания рассмотрим пространство параметров X как многомерный куб, в котором каждое измерение образовано вектором значений одного параметра X_i. Тогда результирующую функцию также можно представить в виде вектора Z_i. Процесс создания прокси-модели $\Phi^s(X,Y)$ будет содержать следующие шаги:

1. Чтение параметров и значений функции из результатов гидродинамического моделирования
2. Векторизация параметров и результирующей функции
3. Построения многомерного куба параметров
4. Построение интерполяционной функции
5. Определение размерностей новых векторов параметров
6. Создание новых векторов параметров
7. Интерполяция новых векторов параметров с помощью интерполяционной функции
8. Запись полученной прокси-модели в формате удобном для использования.

Суть моделирования на основании многомерной линейной интерполяции состоит в том, чтобы подобрать шаг изменения новых параметров так, чтобы полученные вектора параметров имели в своем составе «узловые значения» и покрывали необходимый для модели диапазон с достаточной частотой.

Другими словами, если у вас есть параметр X_i с размерностью $i \in [0,2]$ для которого произведены расчеты в значениях 0.1, 0.5., 0.9, а есть потребность в значениях функции при 0.4 и 0.8, то для нового вектора будет достаточно выбрать шаг 0.1 и размерность $i \in [0,8]$. Таким образом, параметры разбиваются по регулярной сетке.

В случае, когда размерность пространства параметров больше 2 мы уже не можем применять Spline-методы, описанные, например, в [5]. В рассматриваемом эксперименте размерность пространства параметров равна 6 (с учетом параметра прокаченный поровый объем).

Кроме того, стоит отметить, что выбор шага нужно делать с учетом возможностей вычислительных ресурсов. Векторизованное пространство параметров, представленное в виде многомерного массива рациональных чисел должно соотноситься с размерами доступной оперативной памяти сервера для вычислений.

Для простоты под прокси-моделью в нашем случае можно понимать таблицу MS Excel с семью колонками: 6 для входных параметров и одна для результирующей функции. Такое представление максимально понятно широкому кругу специалистов и позволяет им вести дальнейшие исследования на основе данных.

На рисунке 4 отображена зависимость дополнительного КИН от остаточной нефтенасыщенности и неоднородности проницаемости при фиксированных остальных параметрах.

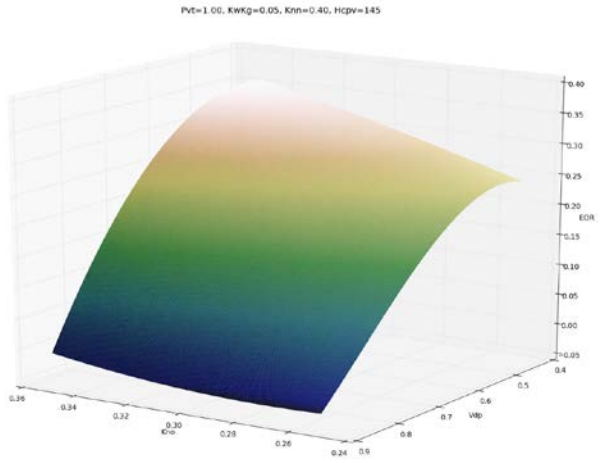

Рисунок 4 Дополнительный КИН по модели многомерной линейной интерполяции при зафиксированных параметрах.

IV. Создание прокси-модели на основании методов машинного обучения

С точки зрения методов машинного обучения наша задача относится к классу задач построению регрессии. Одним из распространенных и универсальных регрессоров является Random Forest придуманный Лео Брейманом [11].

Random forest — это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме [10]:

- Выбирается подвыборка обучающей выборки размера samplesize по ней строится дерево. Для каждого дерева — своя подвыборка.
- Для построения каждого расщепления в дереве просматривается max_features случайных признаков. Для каждого нового расщепления — свои случайные признаки.
- По заранее заданному критерию выбираются наилучший признак и производится расщепление по нему. В оригинальном алгоритме дерево строится до исчерпания выборки, то есть, пока в листьях не

останутся представители только одного класса. Но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Данная схема построения соответствует главному принципу ансамблирования [13] - построению алгоритма машинного обучения на базе нескольких, в данном случае, решающих деревьев: базовые алгоритмы должны быть хорошими и разнообразными.

В описанной выше постановке задачи мы производим обучение регрессора на имеющихся 6 параметрах и значениях дополнительного КИН. А затем используем полученную регрессионную модель для расчётов значений функции дополнительного КИН от новых значений параметров.

При проверке точности модели методом предсказания значений уже известных параметров Коэффициент детерминации (скоринг R^2) получается 0.99 для тестовой выборки из 100 наборов параметров.

Мы так же можем сразу распределить параметры по важности (Таблица 1).

Название параметра X_i	**Важность**
Свойства нефти (плотность, вязкость, давление насыщения)	0.016
Фактическое значение остаточной нефтенасыщенности, полученное по результатам потоковых экспериментов в системе нефть-вода	0.032
Неоднородность проницаемости	0.488
Относительные фазовые проницаемости	0.041
Текущая нефтенасыщенность	0.026
Прокачанный поровой объем	0.397

Таблица 1 Важность параметров.

Точность результата будет завесить от Неоднородности проницаемости (Vdp) и Прокачанного порового объема (Time) в большей степени чем от остальных параметров, что видно на рисунке 4.

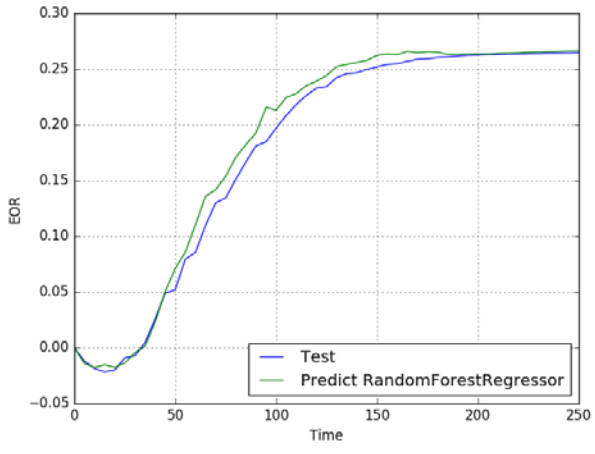

Рисунок 5 Дополнительный КИН по модели Random Forest.

Так же можно оценить влияние количества рассчитанных на гидродинамическом симуляторе экспериментов на точность предсказываемого результата. Таким образом, есть возможность найти оптимальное количество требуемых расчетов на гидродинамическом симуляторе.

V. Вычислительные методы и алгоритмы

Для расчетов была выбрана среда Python. Выбор Python обусловлен наличием широких возможностей по работе с массивами данных как с матрицами, предоставляемых библиотекой NumPy [8]. Для работы по загрузке и выгрузке данных в формат MS Excel была применена библиотека Pandas [9]. Для интерполяции многомерных поверхностей был использованы классы библиотеки SciPy. Отображение 3D поверхностей сделано с помощью библиотеки Matplotlib. В качестве программной реализации многомерной линейной интерполяции был выбран метод Regular Grid Interpolator из библиотеки SciPy [6], [7]. Одним из преимуществ этого метода является то, что он использует возможности регулярной сетки вместо ресурсоемкой триангуляции пространства параметров. Для Random Forest использована реализация scikit-learn [14].

Производительность примененных библиотек была в рамках требований к времени вычислений «on demand». Вычисления проводились на 64-ядрах под управлением ОС Linux (CentOS 7). Для наиболее полной утилизации Multi-Core (многоядерности) авторы использовали библиотеку Math Kernel Library.

VI. Выводы и дальнейшие направления исследования

Важно отметить, что данный подход целесообразно применять в региональном масштабе. Авторы производили расчет для всего спектра значений параметров для месторождений ПАО «Газпром Нефть». Другими словами, единожды рассчитав дополнительный КИН можно в дальнейшем работать по таблицам (суррогатной модели) не прибегая к ресурсоемким расчетам, а только отыскивая нужный набор параметров и соответствующий дополнительный КИН.

Многомерная регрессия по методу Random Forest является современным и высокопроизводительным средством и удовлетворяет требованиям задачи для построения прокси-моделей расчетов дополнительного КИН в диапазоне свойств покрывающим все объекты ПАО «Газпром нефть». Важно отметить полную преемственность по отношению к многомерной линейной интерполяции.

У многомерной регрессии по методу Random Forest существует ряд дополнительных преимуществ по сравнению с методом основании многомерной линейной интерполяции. А именно:

- Возможность оценки важности параметров.
- Возможность определения достаточного количества расчетов гидродинамических моделей, исходя из требуемой точности.

Основным выводом данной статьи авторы считают существенное упрощение вычислений, значительное уменьшение требований к вычислительным ресурсам и достижение лучшей прогнозируемости моделируемой функции при применении методов машинного обучения.

Библиография

[1] А.П.Кулешов, «Когнитивные технологии в адаптивных моделях сложных объектов», ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ 1/2008

[2] Guo, Z., Reynolds, A. C., & Zhao, H. (2017, February 20). A Physics-Based Data-Driven Model for History-Matching, Prediction and Characterization of Waterflooding Performance. Society of Petroleum Engineers. doi:10.2118/182660-MS

[3] Shehata, A. M., El-banbi, A. H., & Sayyouh, H. (2012, January 1). Guidelines to Optimize CO2 EOR in Heterogeneous Reservoirs. Society of Petroleum Engineers. doi:10.2118/151871-MS

[4] Weiser, Alan, and Sergio E. Zarantonello. “A note on piecewise linear and multilinear table interpolation in many dimensions.” MATH. COMPUT. 50.181 (1988): 189-196.

[5] Dierckx, Paul. Curve and surface fitting with splines, Monographs on Numerical Analysis, Oxford University Press, 1993.

[6] Travis E. Oliphant. Python for Scientific Computing, Computing in Science & Engineering, 9, 10-20 (2007), DOI:10.1109/MCSE.2007.58

[7] K. Jarrod Millman and Michael Aivazis. Python for Scientists and Engineers, Computing in Science & Engineering, 13, 9-12 (2011), DOI:10.1109/MCSE.2011.36

[8] Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22-30 (2011), DOI:10.1109/MCSE.2011.37

[9] Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010)

[10] Блог Александра Дьяконова «Случайный лес (Random Forest)», 2016/11/14, https://alexanderdyakonov.wordpress.com

[11] Breiman, Leo (2001). «Random Forests». Machine Learning 45 (1): 5–32. DOI:10.1023/A:1010933404324.

[12] Gashler, M. and Giraud-Carrier, C. and Martinez, T., Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous, The Seventh International Conference on Machine Learning and Applications, 2008, pp. 900-905., DOI 10.1109/ICMLA.2008.154

[13] Opitz, D.; Maclin, R. (1999). "Popular ensemble methods: An empirical study". Journal of Artificial Intelligence Research. 11: 169–198. doi:10.1613/jair.614

[14] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[15] Fadili, A., Kristensen, M. R., & Moreno, J. (2009, January 1). Smart Integrated Chemical EOR Simulation. International Petroleum Technology Conference. doi:10.2523/IPTC-13762-MS

A review of two algorithms for proxy model of enhanced oil recovery

Fedor Krasnov, Nikolay Glavnov, Alexander Sitnikov

Abstract **— In a number of computational experiments, a meta-algorithm is used to solve the problems of the oil and gas industry. Such experiments begin in the hydrodynamic simulator, where the value of the function is calculated for specific nodal values of the parameters based on the physical laws of fluid flow through porous media. Then, the values of the function are calculated, either on a more detailed set of parameter values, or for parameter values that go beyond the nodal values. Among other purposes, such an approach is used to calculate incremental oil production resulting from the application of various methods of enhanced oil recovery (EOR). The authors found out that in comparison with the traditional computational experiments on a regular grid, computation using machine learning algorithms could prove more productive.**

***Keywords*— Enhanced Oil Recovery, EOR, random forest, regular grid interpolation, proxy-model.**