УДК 681.518:622.276

Региональный прогноз областей притока нефти из баженовско-абалакского комплекса на территории ХМАО-Югры методом машинного обучения

Д.А. Ивлев, к.г.-м.н.
(АО «Зарубежнефть»)

Адрес для связи: dm.ivlev@gmail.com

Ключевые слова: баженовско-абалакский комплекс, машинное обучение, генетический алгоритм, решающие деревья, извлечение правил, региональный прогноз областей притока нефти

Regional forecast for zones of oil inflow from Bazhen-Abalak formation in KhMAO-Yugra region of Russia by machine learning method

D.A. Ivlev (Zarubezhneft JSC, RF, Moscow)

E-mail: dm.ivlev@gmail.com

Key words: Bazhen-Abalak formation, machine learning, genetic algorithm, decision tree, rules retrieval, regional forecast for zones of oil inflow

An approach to the regional forecast for zones of oil inflow from Bazhen-Abalak formation has been formalized and tested. The task was to classify the spatial attributes by machine learning through precedents by algorithm of single decision tree with the genetic selection of combination of such attributes. The rules have been retrieved and the factors have been identified which influence the forecast for zones of oil inflow from Bazhen-Abalak formation intervals. The results are shown in the regional forecast scheme with identification of Bazhen-Abalak formation sweet spots in KhMAO-Yugra region of Russia. Such sweet spots can be correlated with perspective zones to get the inflow from the Bazhen-Abalak formation.

В утвержденной Генеральной схеме развития нефтяной отрасли до 2020 г. баженовская свита Западной Сибири определена в качестве одного из приоритетов в инновационном развитии нефтяного комплекса России. Первый промышленный приток нефти дебитом 700 м^3/сут из отложений баженовско-абалакского комплекса получен в 1968 г. из скв. 12Р Салымская. Несмотря на успехи в изучении и опыт эксплуатации месторождений баженовской свиты, вопросы ее строения, генезиса природного резервуара, типа коллектора, морфологии залежей, величины запасов и методов их рациональной разработки до настоящего времени носят дискуссионный характер [1].

За последние годы кратно повысились объем и качество геолого-физических информационных ресурсов. Значительно усовершенствованы математический аппарат и алгоритмы анализа и обработки данных, например, методы построения алгоритмов, способных обучаться. Такое машинное обучение находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но имеет также собственную специфику, связанную с проблемами вычислительной эффективности и переобучения. Методы разрабатывались как альтернатива классическим статистическим подходам, тесно связаны с извлечением информации и интеллектуальным анализом данных.

Целью работы является региональный прогноз получения притока нефти из баженовско-абалакского комплекса на территории ХМАО-Югры на основе ограниченного набора региональных геолого-физических данных и информации о дистанционном зондировании Земли. Региональный прогноз рассматривается как задача классификации с обучением по прецедентам методом машинного обучения с выявлением закономерностей в эмпирических данных (атрибутах пространственных данных) и распространением закономерностей на всю изучаемую территорию (рис. 1). Для ее решения применялось свободное программное обеспечение QGIS, SAGA, GRASS, Python (Scikit-Learn, Pandas).

Для исследования использованы результаты испытания 438 скважин в интервале баженовско-абалакского комплекса и следующий набор атрибутов пространственных данных, описывающих рассматриваемую территорию в пределах района с индексом Р-42 ХМАО-Югра: карта гравитационных аномалий редукции Буге; карта магнитных аномалий; структурные карты по кровле и подошве баженовско-абалакского горизонта, верхней, средней и нижней юры, лайдинского горизонта; карты изопахит баженовско-абалакского горизонта, верхней, средней и нижней юры, лайдинского горизонта и их комбинации; карты расстояний до ближайших линеаментов и крупных тектонических элементов. Каждой скважине с координатами пластопересечения с баженовско-абалакским комплексом присвоено 26 атрибутов из набора пространственных данных.

Выборка из испытаний скважин разделена на три кластера по критериям методики проведения испытания на приток и результатам апробации интервала. К первому и второму кластерам отнесены скважины с раздельным испытанием интервала изучаемого комплекса в обсаженной колонне. Первый кластер включает скважины, в которых получен приток нефти или наблюдались признаки углеводородов (класс «приток»). Во втором кластере находятся скважины, испытания которых показали отсутствие притока и признаков углеводородов (класс «сухо»). Третий кластер сформирован по остаточному принципу и в дальнейшей работе не рассматривался. В него вошли скважины с совместным испытанием нескольких горизонтов в обсаженном и необсаженном стволе скважины.

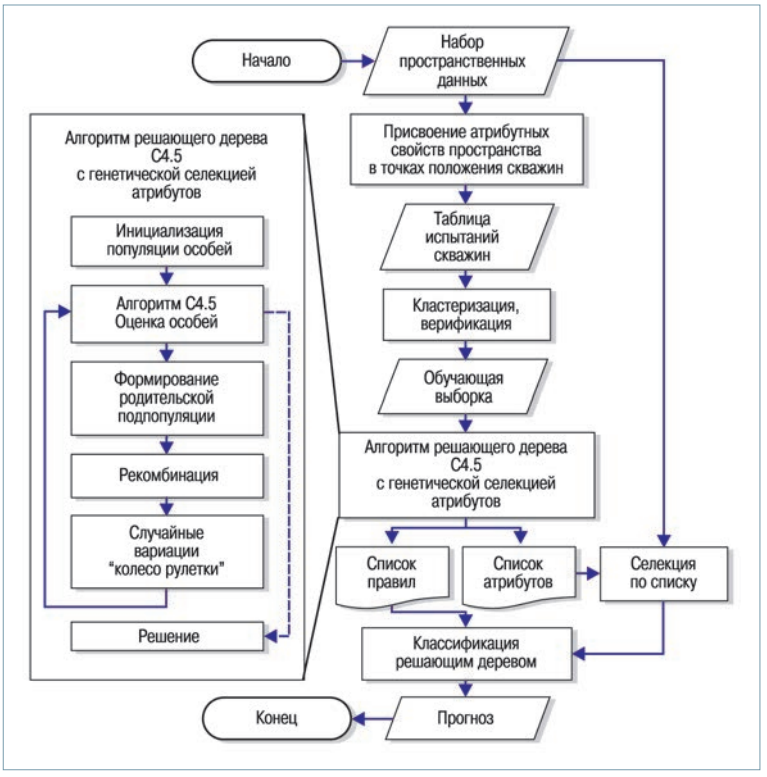

Рис. 1. Алгоритм регионального прогноза притока нефти из баженовско-абалакского комплекса

В итоговую обучающую выборку вошли 235 скважин: в первый класс «приток» – 127, во второй класс – «сухо» 108. Для анализа полученной выборки из многообразия алгоритмов машинного обучения выбран метод решающих деревьев, отвечающий следующим основным требованиям: 1) извлечение правил на естественном языке, возможность анализа полученных решений; 2) иерархическая классификационная модель. Решающие деревья воспроизводят логические схемы, позволяющие получить окончательное решение о классификации объекта с помощью ответов на иерархически организованную систему вопросов. Вопрос, задаваемый на последующем иерархическом уровне, зависит от ответа, полученного на предыдущем уровне. Таким образом, обучение решающих деревьев является способом формализованного анализа по прецедентам с целью извлечения новых знаний, правил и критериев [2, 3].

Построение дерева выполнялось по алгоритму C4.5 [4]. Обучение предусматривало перебор комбинаций атрибутов для снижения «шума данных» и влияния «проклятия размерности» с поиском оптимального набора признаков по критерию качества прогноза с помощью генетического алгоритма [5] методом селекции – «колесо рулетки» (см. рис. 1). В процессе формирования обучающей выборки и обучения алгоритма выявлены следующие проблемы проведения исследования, влияющие на качество прогноза.

Репрезентативность выборки скважин как объектов, описывающих генеральную совокупность свойств изучаемого пространства. Поиск традиционных месторождений осуществляется исходя из предпосылок осадочно-миграционной теории образования углеводородов. Наблюдается скученность скважинных данных в отдельных областях (антиклинальных структурах), остальные территории практически не охарактеризованы результатами бурения.

Репрезентативность выборки скважин как объектов, описывающих исход испытания. Обучающая выборка описывает два исхода испытания объекта: получен или не получен приток углеводородов. Однако для большинства исследований существует следующая вероятность:

а) не получен приток углеводородов – возможно, некачественно было проведено первичное или вторичное вскрытие, также возможно некачественно выполнены вызов и интенсификация притока;

б) получен приток углеводородов, существует вероятность, что при интенсификации притока произошло приобщение выше- или нижележащих продуктивных интервалов в результате перетока за обсадной колонной через цементный камень или горную породу.

Малый объем обучающей выборки. Для данной выборки скважин установлена невозможность однозначной классификации результатов испытаний скважин при апробировании одновременно нескольких перспективных объектов, включая исследуемый комплекс. Это снизило число прецедентов, вошедших в обучающую выборку: из 438 испытаний 181 было признано некондиционным.

Недостаточное разрешение пространственных данных. В некоторых случаях разрешения пространственных данных с регулярной сеткой 250×250 м, описывающих исследуемую территорию, недостаточно. Были выделены 44 случая с противоположным исходом испытания и идентичными атрибутами пространственных признаков за счет близкого расположения испытанных скважин. При этом территория признавалась перспективной, 22 скважины из кластера «сухо» были исключены из выборки.

Ограничения алгоритма одиночного решающего дерева. Данный алгоритм не может охватить всего многообразия причинно-следственных связей. Для повышения качества прогноза необходимо использовать ансамбли решающих деревьев, каждое из которых осведомлено об ошибках предыдущих, и таким образом комбинировать «слабые» классификаторы, чтобы получить «сильный». Примером данного алгоритма служит Gradient Boosted Decision Trees (GBDT) в его частной реализации XGBoost.

В результате 2000 итераций генетического алгоритма путем отбора из 26 атрибутов выделены 6, комбинация которых дала наименьшую ошибку классификации. Обучаемый алгоритм построения решающего дерева показал наилучшее качество классификации при кросс-валидации (табл. 1) со следующими пространственными атрибутами: толщина верхнеюрского горизонта; гравитационные аномалии редукции Буге; толщина баженовско-абалакского комплекса; толщина среднеюрского го-

Таблица 1

Точность прогноза	Чувствительность	Специфичность	Показатель AUC	Показатель Бриера
0,762	0,811	0,703	0,7523	0,400

Примечание. Показатель AUC – площадь под ROC-кривой (графиком, позволяющим оценить качество бинарной классификации).

ризонта; расстояние до ближайшего линеамента; толщина лайдинского горизонта.

Анализ параметров качества построенного решающего дерева [5] показал, что алгоритм дает приемлемое соотношение ложных срабатываний к правильной классификации (см. табл. 1). Так, AUC = 0,75 соответствует критерию полезной прогностической системы. При прогнозе класса «сухо» из 108 случаев в 32 алгоритм ошибочно распознал класс «приток», а из выборки класса «приток» из 127 случаев в 24 сделана неверная классификация (табл. 2). В целом точность правильного определения класса составила 76,2 %. Примененный метод машинного обучения путем построения одиночного бинарного решающего дерева, несмотря на то, что неспособен охватить все многомерное пространство причинно-следственных связей между атрибутами, обладает приемлемым прогностическим потенциалом.

Таблица 2

Факт	Прогноз		
	«Сухо»	«Приток»	Сумма
«Сухо»	76	32	108
«Приток»	24	103	127
Сумма	100	135	235

В результате машинного обучения алгоритмом одиночного решающего дерева на обучающей выборке из 235 результатов испытаний баженовско-абалакского комплекса получен набор формализованных и разделенных по иерархическим уровням правил классификации (табл. 3). Итоговая реализация бинарного решающего дерева позволила сформировать правила с четырьмя иерархическими уровнями (табл. 4). Первый уровень в иерархии правил с корневым признаком – толщина верхнеюрского горизонта, второй уровень образуют два атрибута (см. табл. 4). Для ветви 1 третий уровень иерархии формирует правило с селективным значением толщины лайдинского горизонта, которое классифицирует выборку в 3 случаях из 4. Самая длинная ветвь образована атрибутом – толщина баженовского горизонта с четвертым уровнем в иерархии правил.

Для удобства представления дерево разделено по второму уровню иерархии правил на четыре ветви 1а, 1б, 2а, 2б (см. табл. 3). Ветвь 1а сформирована следующим правилом: уменьшение толщины верхней юры (⩽ 28,5 м) с одновременным снижением толщины средней юры (⩽ 241,5 м) и толщины лайдинского горизонта (⩽ 27,75 м) определяет класс «приток». При тех же начальных условиях увеличение толщины лайдинского горизонта более 27,75 м формирует правило для класса «сухо».

Ветвь 1б описывается следующим правилом: уменьшение толщины верхней юры с одновременным увеличением толщин средней юры и лайдинского горизонта – класс «приток». При толщине лайдинского горизонта менее 32,25 м вводится дополнительно селективное правило: толщина баженовского горизонта 24,882 м. При превышении этого значения правило определяет класс «приток», при равном или меньшем значении – класс «сухо».

Ветвь 2а: при увеличении толщины верхней юры (> 28,5 м) с одновременным снижением значений гравитационной аномалии менее -16,225 мГал и уменьшением толщины средней юры (⩽ 213 м) ветвь классифицируется как «сухо», при тех же условиях увеличение толщины средней юры (> 213 м) правило определяет класс «приток».

Ветвь 2б: при увеличении толщины верхней юры (> 28,5 м) с одновременным ростом гравитационной аномалии более -16,225 мГал при уменьшении толщины средней юры менее 265,5 м ветвь классифицируется как «сухо». При толщине средней юры более 265,5 м добавляется ветвь по селективному признаку «расстояние до ближайшего линеамента», при уменьшении которого менее 4,5 км правило определяет класс «сухо», при увеличении расстояния выше граничного значения (> 4,5 км) – класс «приток».

Согласно полученным правилам была классифицирована исследуемая территория ХМАО-Югры. Результатом является региональная прогнозная схема с выделенными для территории классами, которые можно соотнести с перспективными и неперспективными областями для получения притока из баженовско-абалакского комплекса (рис. 2).

Таблица 3

Правило классификации	Прогноз класса	Точность правила	Выборка	Уровень	Ветвь
Толщина верхней юры ⩽28,5 м	«Приток»	0,709	127	1	1
Толщина средней юры ⩽241,5 м	«Приток»	0,679	56	2	1а
Толщина лайдинского горизонта ⩽27,75 м	«Приток»	0,783	46	3	
Толщина лайдинского горизонта >27,75 м	«Сухо»	0,800	10	3	
Толщина средней юры >241,5 м	«Приток»	0,732	71	2	1б
Толщина лайдинского горизонта ⩽32,25 м	«Приток»	0,516	31	3	
Толщина БАК ⩽24,882 м	«Сухо»	0,714	14	4	
Толщина БАК >24,882 м	«Приток»	0,706	17	4	
Толщина лайдинского горизонта >32,25 м	«Приток»	0,900	40	3	
Толщина верхней юры >28,500	«Сухо»	0,657	108	1	2
Значение гравитационной аномалии ⩽-16,225 мГал	«Приток»	0,520	25	2	2а
Толщина средней юры > 213 м	«Приток»	0,929	14	3	
Толщина средней юры ⩽213 м	«Сухо»	1,000	11	3	
Значение гравитационной аномалии > -16,225 мГал	«Сухо»	0,711	83	2	2б
Толщина средней юры ⩽265,500	«Сухо»	0,842	57	3	
Толщина средней юры >265,500	«Приток»	0,577	26	3	
Расстояние до линеамента ⩽4575,5 м	«Сухо»	0,750	12	4	
Расстояние до линеамента > 4575,5 м	«Приток»	0,857	14	4	

Таблица 4

Уровень	Ветвь 1	Ветвь 2
Первый	Толщина верхнеюрского горизонта	Толщина верхнеюрского горизонта
Второй	Толщина средней юры	Значение гравитационной аномалии
Третий	Толщина лайдинского горизонта	Толщина средней юры
Четвертый	Толщина баженовско - абалакского комплекса	Расстояние до ближайшего линеамента

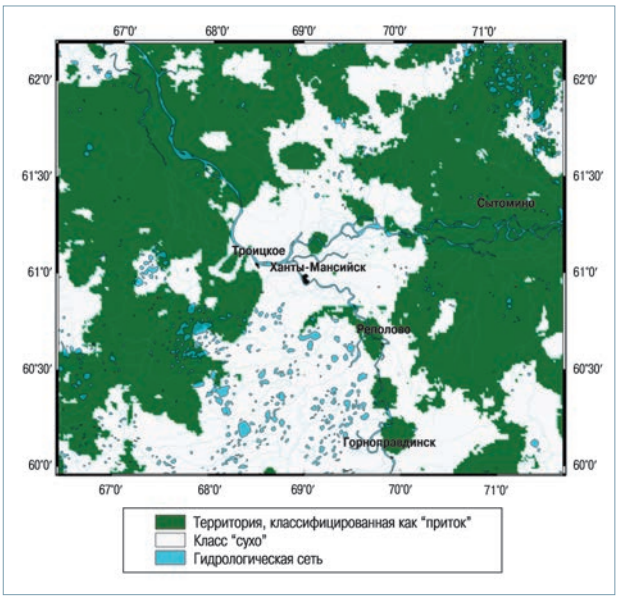

Рис. 2. Схема прогноза притока нефти из баженовско-абалакского комплекса

Таким образом, в результате машинного обучения созданы правила и выделены факторы, влияющие на прогноз получения притока из интервалов баженовско-абалакского комплекса. К благоприятным факторам можно отнести снижение толщины верхнеюрских отложений (< 28,5 м), только этот селективный фактор классифицирует 127 случаев как «приток» с точностью 0,709. При уменьшении толщины верхнеюрского разреза одновременно увеличенные или сокращенные толщины среднеюрского и лайдинского горизонтов повышают вероятность получения притока. Разнонаправленные тенденции изменения толщин являются неблагоприятным фактором, кроме случаев аномально увеличенных толщин баженовско-абалакского комплекса. К неблагоприятным факторам относится увеличение толщины верхнеюрских отложений (> 28,5 м), данное правило разделяет выборку с точностью прогноза класса «сухо» 0,657 для 108 случаев. Если при этом в разрезе уменьшается толщина средней юры, то точность прогноза класса «сухо» повышается до 1 при значениях гравитационной аномалии редукции Буге менее -16,225 мГал и до 0,842 при значениях более -16,225 мГал. При прочих равных условиях увеличение в разрезе толщин средней юры повышает вероятность получения притока, кроме случаев, когда значения гравитационной аномалии увеличены и скважина находится на расстоянии менее 4,5 км от ближайшего линеамента.

Выводы

1. Алгоритм построения бинарного решающего дерева с процедурой генетической селекции предикторов (пространственных атрибутов) обладает приемлемым прогностическим потенциалом со значением качества прогноза: точность прогноза (CA) составляет 0,762, AUC – 0,7523.

2. В процессе генетической селекции из 26 атрибутов пространственных признаков выбраны шесть, комбинация которых дала наименьшую ошибку классификации.

3. На точность прогноза влияют следующие пять проблем реализации исследования : 1) репрезентативность выборки скважин как объектов, описывающих генеральную совокупность свойств изучаемого пространства; 2) репрезентативность выборки скважин как объектов, описывающих исход испытания; 3) малый объем обучающей выборки; 4) недостаточное разрешение пространственных данных; 5) ограничения алгоритма одиночного решающего дерева.

4. Построенное бинарное решающее дерево содержит четыре уровня иерархии правил.

5. С учетом ограничений в исходных данных, проблем, связанных с обучающей выборкой, и особенностей метода машинного обучения с применением алгоритма одиночного решающего дерева к количественной оценке правил и результатам классификации необходимо относиться с осторожностью. Однако на качественном уровне выделенные при таком подходе иерархические уровни селективных признаков и тенденции их изменений при создании правил могут быть использованы для формирования представлений об условиях нефтеносности баженовско-абалакского комплекса.

Список литературы

1. *Брехунцов А.М., Нестеров И.И. (мл.), Нечипорук Л.А.* Битуминозные глинистые отложения баженовского горизонта – приоритетный стратегический объект нефтедобычи в Западной Сибири/ http://oilgasjournal.ru/vol_10/brekhuntsov.pdf

2. *Classification* and Regression Trees/L. Breiman, J. Friedman, R. Olshen, C. Stone // Wadsworth & Brooks, Pacific Grove, CA. – 1984. – 368 p.

3. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd. s.1. – 2009. – 739 p.

4. *Quinlan R.* C4.5 Programs for Machine Learning. Morgan Kaufmann, San Mateo, California. 1993. – 302 p.

5. *Poli R., Langdon W.B., McPhee N.F.* A Field Guide to Genetic Programming, Lulu.com, freely available from the internet. – 2008. – 250 p.

6. *Standardized* Verification System (SVS) for Long-Range Forecasts (LRF), New Attachment II-9 to the Manual on the GDPS (WMO-No. 485). – 2002. – V. I. – P. 14–20.

References

1. Brekhuntsov A.M., Nesterov I.I. Jr., Nechiporuk L.A., *Bituminoznye glinistye otlozheniya bazhenovskogo gorizonta – prioritetnyy strategicheskiy ob"ekt neftedobychi v Zapadnoy Sibiri* (Bituminous clay deposits of Bazhenov horizon – priority strategic facility of oil production in Western Siberia), URL: http://oilgasjournal.ru/vol_10/brekhuntsov.pdf

2. Breiman L., Friedman J., Olshen R., Stone C., *Classification and regression trees*, Wadsworth & Brooks, Pacific Grove, CA, 1984, 368 p.

3. Hastie T., Tibshirani R., Friedman J., *The elements of statistical learning: data mining*, inference, and prediction, 2009, 739 p.

4. Quinlan R., *Programs for machine learning*, San Mateo : Morgan Kaufmann, 1993, 302 p.

5. Poli R., Langdon W.B., McPhee N.F., *A field guide to genetic programming*, 2008, 250 p.

6. *Standardized Verification System (SVS) for Long-Range Forecasts (LRF)*, New attachment II-9 to the Manual on the GDPS (WMO-No. 485), 2002, V. I, pp. 14–20.