ГЕОФИЗИЧЕСКИЕ ИССЛЕДОВАНИЯ, 2012, том 13, № 1, с.23-28

УДК 550.34.012

КЛАССИФИКАТОР БАЙЕСА В РЕШЕНИИ ЗАДАЧИ ВЕРОЯТНОСТНОГО ПРОГНОЗА ВЕЩЕСТВЕННОГО СОСТАВА ГЛУБОКИХ ГОРИЗОНТОВ ЗЕМНОЙ КОРЫ ПО ГЕОФИЗИЧЕСКИМ ДАННЫМ

© 2012 г. П.А. Леляев[1], А.Я. Салтыковский[1], М.Е. Семенов[2], В.В. Мацковский[3]

[1] *Институт физики Земли им. О.Ю. Шмидта РАН, г. Москва, Россия*
[2] *Воронежский государственный университет, г. Воронеж, Россия*
[3] *Институт географии РАН, г. Москва, Россия*

Предлагается модель классификации глубинных пород по их плотности и скорости, основанная на классическом классификаторе Байеса. Объектом представляемых исследований является Воронежский кристаллический массив, хорошо изученный геофизическими методами.

Ключевые слова: классификация, алгоритм, порода, скорость, плотность, вероятность.

Одной из наиболее практически значимых задач, решаемых современной геофизикой, является вероятностный прогноз вещественного состава глубоких горизонтов земной коры, который напрямую связан с глубинными исследованиями, в том числе и геофизическими. Прогноз основан на сопоставлении характеристик глубокозалегающих горизонтов, получаемых косвенными методами (сейсмическими, гравиметрическими, электрокаротажными и т.д.), с петрофизическими данными – известными свойствами образцов горных пород, наиболее представительных для глубоких горизонтов земной коры исследуемого региона. При этом сопоставлении важное место отводится использованию различных математических методов. Осуществление подобного прогноза связано с рядом трудностей и имеет свои особенности.

Во-первых, образцы, включаемые в петрофизическую базу данных, должны быть представительными именно для исследуемого региона, поскольку образцы одних и тех же горных пород, отобранные в разных регионах, могут различаться по физическим характеристикам.

Во-вторых, существует множество способов классификации объектов (в данном исследовании – типов горных пород) по совокупности информативных признаков, в связи с чем в каждом конкретном случае следует выбирать метод, обеспечивающий наибольшую достоверность благодаря использованию объективных критериев.

В-третьих, существуют физические характеристики горных пород, которые могут быть установлены как на образцах в лабораторных условиях, так и для глубоких горизонтов с помощью геофизических методов (например, остаточная намагниченность, электрическая проницаемость и др.). Однако при исследованиях до 50–60 км доступными для измерения остаются две – скорость распространения продольных волн в среде и плотность пород. При использовании только этих двух параметров вероятность успешной классификации снижается; при включении других физических величин во множество наблюдаемых параметров вероятностный прогноз вещественного состава горизонта, залегающего на больших глубинах, становится неосуществимой задачей.

Объектом исследований, представляемых в настоящей статье, является Воронежский кристаллический массив – один из крупнейших сегментов Восточно-Европейской платформы, представляющий собой выступ докембрийского кристаллического фундамента, в состав которого входят сложно дислоцированные метаморфические и магматические породы архея и протерозоя. Сравнительно небольшая глубина залегания поверхности фундамента (от 0 до 600 м) позволила хорошо изучить свойства наиболее характерных для него пород [*Афанасьев*, 2001]. В результате анализа образцов, отобранных из верхней части Воронежского массива, и сопоставления их с образцами из других регионов были определены наиболее характерные для исследуемого массива десять типов пород: габбро, граниты, диориты, пироксениты, принадлежащие к магматическим породам, и амфиболиты, гнейсы, мигматиты, перидотиты, серпентиниты, сланцы, относимые к метаморфическим. Исходная база данных включала более 2 тыс. образцов с известными физическими характеристиками.

Статистические исследования предполагают анализ свойств образцов, включенных в базу данных, с целью выявления в ней закономерностей, на основе которых затем осуществляется классификация пород по совокупности информативных признаков. Такая классификация может быть осуществлена несколькими способами. Основными требованиями к ней являются достоверность (а, следовательно, и возможность ее расчета), универсальность методов исследования (для возможного применения в аналогичных закрытых регионах), наглядность полученных результатов.

Перед началом любого статистического анализа из базы данных необходимо удалить выбросы – элементы, по одному или нескольким признакам имеющие значения, сильно отличающиеся от остальных [Электронный…, 2001]. Поскольку петрофизические характеристики всех исследуемых в работе пород подчиняются нормальному закону распределения, удаление выбросов можно провести с помощью стандартного статистического правила “трех сигма”. Это означает, что с вероятностью 0.9973 (т.е. практически с единичной) нормально распределенная случайная величина окажется в пределах $\pm 3\sigma$ от среднего значения [*Боровков*, 1984]. Иначе говоря, отклонения от среднего больше чем на 3σ можно ожидать примерно в одном из 370 испытаний. Таким образом, все образцы, петрофизические характеристики которых выходят за границы указанного интервала хотя бы по одному из параметров, были удалены из базы.

Алгоритмы математической статистики, искусственного интеллекта и глубокого анализа данных *Data Mining* (метод линейной регрессии, нейросети, самообучающаяся карта Кохонена, построение дерева решений), реализованные в существующем программном обеспечении (*Deductor Studio, STATISTICA* и др.), дают крайне низкий (до 60%) уровень достоверности получаемых результатов, в связи с чем не могут быть использованы на практике [*Леляев и др*., 2010]. Поэтому возникла необходимость в алгоритме классификации, позволяющем не только с наибольшей вероятностью отнести объект к одному из рассматриваемых классов, но и оценить эту вероятность. Авторы предлагают рассмотреть классификатор Байеса, использующий теорему Байеса:

$$P(H_k \mid A_i) = \frac{P(H_k)P(A_i \mid H_k)}{P(A_i)}, \qquad (1)$$

где $P(H_k)$ – априорная вероятность принадлежности наблюдения *k*-ому классу; $P(A_i)$ – вероятность попадания параметров наблюдения в *i*-й (строго говоря, не одномерный) интервал; $P(H_k \mid A_i)$ – условная вероятность принадлежности наблюдения *k*-му классу при условии попадания параметров наблюдения в *i*-й интервал (именно ее и требуется найти для каждого интервала значений); $P(A_i \mid H_k)$ – условная вероятность того, что

порода *k*-го типа принадлежит *i*-му интервалу значений параметра наблюдений (частота). В случаях, когда наблюдение может с разной вероятностью принадлежать к различным классам, результатом работы классификатора будет вектор, компоненты которого являются вероятностями принадлежности к определенному классу.

Байесовский классификатор в каком-то смысле является оптимальным. Его результат не может быть улучшен, так как во всех случаях, когда возможен однозначный ответ, он его даст, а в тех случаях, когда ответ неоднозначен, результат количественно характеризует меру этой неоднозначности. Недостатком байесовского классификатора является так называемое “проклятие размерности” – для его построения требуется выборка, размер которой экспоненциально растет с ростом размерности параметра наблюдения [*Бочканов, Быстрицкий,* 1999–2012].

Для преодоления этой проблемы на практике используют так называемый наивный байесовский классификатор, построенный на предположении о независимости переменных, т.е. на том, что

$$\forall i,k : P(A_i \mid H_k) = P(a_{i1} \mid H_k) \cdot P(a_{i2} \mid H_k) \cdot \ldots \cdot P(a_{in} \mid H_k), \qquad (2)$$

где $A_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$; n – размерность параметра A; a_{ij} – событие, заключающееся в попадании j-й компоненты вектора A в i-й интервал наблюдений. Использование этого предположения позволяет не изучать взаимодействие всех возможных сочетаний переменных, ограничившись лишь влиянием каждой переменной по отдельности на принадлежность образа к одному из классов. Согласно теореме Байеса, в этом случае мы имеем

$$P(H_k \mid A_i) = \frac{P(H_k)}{P(A_i)} \cdot P(a_{i1} \mid H_k) \cdot \ldots \cdot P(a_{in} \mid H_k). \qquad (3)$$

Преимуществом такого подхода является то, что требования к размеру выборки сокращаются от экспоненциальных до линейных. Его недостаток связан с тем, что модель точна лишь тогда, когда выполняется предположение о независимости. В противном случае вычисленные вероятности могут не являться точными (и даже более того, их сумма может не равняться единице, вследствие чего потребуется нормирование результата). Однако на практике небольшие отклонения от независимости приводят лишь к незначительному снижению точности. Даже в случае существенной зависимости между переменными результат работы классификатора положительно коррелирует с истинной принадлежностью образа к классам. При этом достоинства классификатора – высокая скорость работы программы, реализованной на ЭВМ, простота и масштабируемость, умеренные требования к памяти – часто перевешивают его недостатки.

В многомерном случае, каким и является рассматриваемая задача, имея набор переменных $A = \{a_1, a_2\}$, где a_1 – плотность, a_2 – скорость, требуется определить апостериорную вероятность события H_k из множества возможных исходов $H = \{H_1, \ldots, H_{10}\}$, где $\forall k \;\; H_k$ – один из рассматриваемых типов пород. Преобразуем выражение (1):

$$P(H_k \mid a_{i1}, a_{i2}) = \frac{P(H_k)}{P(a_{i1}, a_{i2})} \cdot P(a_{i1}, a_{i2} \mid H_k), \qquad (4)$$

где $P(H_k \mid a_{i1}, a_{i2})$ – апостериорная вероятность классовой принадлежности, т.е. вероятность того, что объект с двумерным параметром A_i принадлежит H_k.

Пример визуализации такой апостериорной вероятности для амфиболитов приведен на рис. 1.

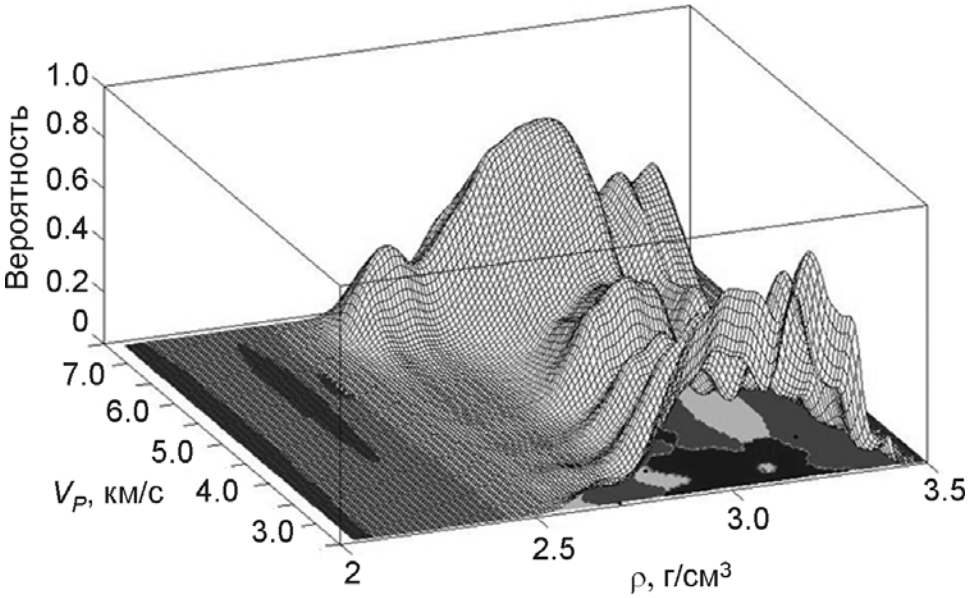

Рис. 1. Трехмерная визуализация плотности апостериорной вероятности для амфиболитов

Поскольку механизм работы классификатора основан на предположении о статистической независимости условных вероятностей независимых переменных, можно представить меру правдоподобия в виде произведения

$$P(A_i \mid H_k) = P(a_{i1} \mid H_k) \cdot P(a_{i2} \mid H_k) \tag{5}$$

и затем преобразовать выражение (4) для апостериорной вероятности

$$P(H_k \mid A_i) = \frac{P(H_k)}{P(a_{i1}, a_{i2})} \cdot P(a_{i1} \mid H_k) \cdot P(a_{i2} \mid H_k), \tag{6}$$

помечая наблюдение с параметром A_i меткой того класса H_k, апостериорная вероятность принадлежности которому наиболее высока.

Для реализации предложенного алгоритма необходимо оценить значения правой части формулы (6). Значения $P(a_{ij} \mid H_k)$ могут быть оценены по статистическим наблюдениям; то же самое можно сказать о $P(a_1, a_2)$, т.е. на практике плотности вероятности заменяются соответствующими частотами.

Априорные вероятности $P(H_k)$ могут быть получены путем анализа схожих геологических структур с известным вертикальным разрезом горных пород; в настоящей работе они полагались равными доле количества образцов каждой породы в общем числе исследуемых образцов.

Используя реализованный в программе *MATLAB* наивный байесовский классификатор, авторы получили трехмерные модели плотностей апостериорной вероятности для каждого из исследуемых типов пород. На рис. 1, иллюстрирующем такую плотность вероятности для амфиболитов, можно, например, видеть, что с наибольшей вероятностью амфиболиты встретятся в слое, имеющем плотность от 2.9 до 3.4 г/см^3 и скорости 6.0 – 6.5 км/с.

Результат работы программы-классификатора приведен на рис. 2, где показана плоскость в осях анализируемых параметров (плотность–скорость). Цвет соответствует типу породы, плотность апостериорной вероятности для которой в данной точке больше, чем у других типов пород.

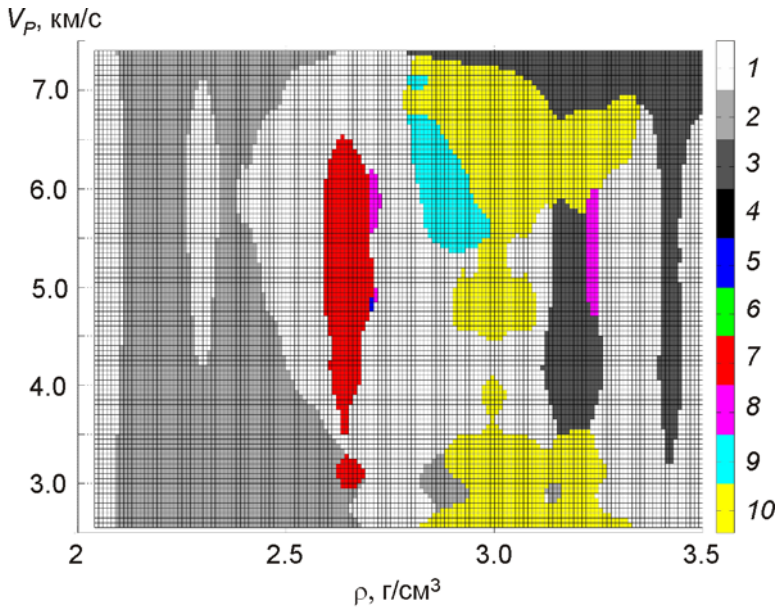

Рис. 2. Результат работы программы-классификатора: *1–10* – типы пород (см. таблицу)

В качестве примера приводится расчет плотности апостериорной вероятности с помощью байесовского классификатора для образца породы с ρ=2.8 г/см^3 и v =5 км/с (таблица).

k	$P(H_k \mid a_1 = 2.8,\ a_2 = 5)$	k	$P(H_k \mid a_1 = 2.8,\ a_2 = 5)$
1 (амфиболит)	0.014	6 (мигматит)	0.0101
2 (габбро)	0.0011	7 (перидотит)	0.036
3 (гнейс)	0.0085	8 (пироксенит)	0.0001
4 (гранит)	0.0575	9 (серпентинит)	0.1057
5 (диорит)	0.0001	10 (сланец)	0.7671

Так как $P(a_1, a_2)$ не зависит от k, решением задачи классификации для анализируемого образца с учетом формулы (6) будет

$$k = \arg\max_k P(H_k) \cdot P(a_1 = 2.8 \mid H_k) \cdot P(a_2 = 5 \mid H_k). \qquad (7)$$

В нашем примере k=10, т.е. рассматриваемый образец следует классифицировать как сланец.

Анализируя результаты классификации, можно сделать выводы о количественном соотношении пород в слое с заданными геофизическими характеристиками и на их основе построить модель конкретного геофизического профиля в любом закрытом в геологическом плане регионе.

Большей точности в расчетах можно достичь, объединяя для исследования однотипные породы в одну группу (например, граниты и мигматиты в группу кислых пород) и учитывая априорные вероятности содержания определенных типов пород в отдельных слоях или в коре исследуемого региона в целом.

Литература

Афанасьев Н.С. К вопросу петрофизической классификации кристаллических горных пород (на примере ВКМ) // Вестник Воронеж. ун-та. Сер. геология. 2001, №12. С. 159–172.

Боровков А.А. Математическая статистика. М.: Наука, 1984. 472 с.

Бочканов С., Быстрицкий В. Байесовский классификатор // ALGLIB® – numerical analysis library, 1999–2012. WEB: http://alglib.sources.ru/dataanalysis/bayes.php.

Леляев П.А., Салтыковский А.Я., Надежка Л.И., Семенов М.Е. Алгоритм распознавания типа пород в верхних горизонтах земной коры по плотности и скорости сейсмических волн (на примере Воронежского кристаллического массива) // Геофизические исследования. 2010. Т. 11, № 2. С.5–14.

Электронный учебник по промышленной статистике. М.: StatSoft, 2001. http://www.statsoft.ru/home/portal/textbook_ind/default.htm.

Сведения об авторах

ЛЕЛЯЕВ Петр Алексеевич – аспирант очной аспирантуры, Институт физики Земли им. О.Ю. Шмидта РАН. 123995, ГСП-5, Москва, Д-242, ул. Большая Грузинская, д. 10, стр. 1. Тел.: (495) 254-89-35. E-mail: norby76@yandex.ru

САЛТЫКОВСКИЙ Артур Яковлевич – доктор геолого-минералогических наук, главный научный сотрудник, Институт физики Земли им. О.Ю. Шмидта РАН. 123995, ГСП-5, Москва, Д-242, ул. Большая Грузинская, д. 10, стр. 1. Тел.: (495) 254-89-35. E-mail: saltyk@ifz.ru

СЕМЕНОВ Михаил Евгеньевич – доктор физико-математических наук, профессор, Воронежский государственный университет. 394006, г. Воронеж, Университетская пл., д. 1. Тел.: (4732) 207-522. E-mail: mkl150@mail.ru

МАЦКОВСКИЙ Владимир Владимирович – кандидат географических наук, Институт географии РАН. 119017, Москва, Старомонетный пер., д. 29. Тел.: (499) 125-90-11. E-mail: matskovskomu@gmail.com

BAYESSIAN CLASSIFIER IN SOLUTION OF THE PROBLEM OF PROBABILITY FORECASTING OF MATERIAL COMPOSITION OF DEEP HORIZONS OF THE EARTH'S CRUST FROM GEOPHYSICAL DATA

P.A. Lelyaev[1], A.Ya. Saltykovskiy[1], M.E. Semenov[2], V.V. Matskovskiy[3]

[1] *Schmidt Institute of Physics of the Earth, Russian Academy of Sciences, Moscow, Russia*
[2] *Voronezh State University, Voronezh, Russia*
[3] *Institute of Geography, Russian Academy of Sciences, Moscow, Russia*

Abstract. A model of rock type classification by its density and speed founded on the classical Bayesian classifier is proposed. As of the object of research in the article considered the Voronezh crystalline massif, well-studied by geophysical methods.

Keywords: classification, algorithm, rock, speed, density, probability.