

Метод идентификации аномальных геологических зон¹

В.Г. Гитис*, А.Б. Дерендяев*, К.Н. Петров*

** Институт проблем передачи информации, Российская академия наук, Москва, Россия*

Поступила в редколлегию 07.06.2019

Аннотация—В статье предложен новый подход к идентификации зон с редкими аномальными проявлениями геологических процессов. Подход основан на двух методах машинного обучения одноклассовой классификации: на методе минимальной области тревоги и на методе предпочтений. Алгоритм минимальной области тревоги является непараметрическим. Он обучается по выборке примеров аномальных событий и вычисляет поле аномальных зон. Полученное этим методом знание является невербализованным. Метод предпочтений позволяет аппроксимировать полученное решение с помощью достаточно простого логического правила, которое определяет аномальную область в терминах анализируемых свойств геологической среды. Рассмотрены примеры применения данного подхода для выделения зон возможных очагов сильных землетрясений и для регионального прогноза месторождений.

КЛЮЧЕВЫЕ СЛОВА: машинное обучение, пространственный прогноз, сейсмическая опасность, рудные месторождения.

1. ВВЕДЕНИЕ

Задачи пространственного прогноза свойств геологической среды, которые не зависят от времени, возникают при экологическом районировании территории, при анализе и оценке природной опасности и при прогнозе месторождений полезных ископаемых. Основное предположение состоит в том, что в пределах изучаемой области значение прогноза связано однозначной зависимостью (правилом прогноза) с характеристиками геологической среды, которые известны для всего региона. Это означает, что исследуемый регион должен представлять собой однородную в определенной смысле геологическую зону. Например, маловероятно, что зависимость между максимальной энергией ожидаемых землетрясений для зон орогенных складчатых сооружений может совпадать с аналогичной зависимостью для платформенных зон [1]. Для каждого типа регионов следует решать отдельную задачу прогноза.

Суть задачи пространственного прогноза состоит в том, чтобы по выборке примеров целевых событий и комплексу геолого-геофизических данных вычислить поле числового или номинального свойства геологической среды и представить это свойство в виде карты.

В ряде случаев прогнозная карта составляется высоко квалифицированным экспертом в изучаемой проблемной области. При этом в некоторых зонах карты решения эксперта могут быть интуитивными и трудно интерпретируемыми. Поскольку эксперт обычно не формулирует количественного правила прогноза, то экстраполяция его решения на другой регион с аналогичными геолого-геофизическими свойствами невозможна.

Исходные данные более объективны, чем их комплексная интерпретация, определяющая прогноз эксперта. Поэтому разрабатываются методы машинного обучения для осуществления прогноза по выборке прецедентов. Формальное правило прогноза позволяет экстраполировать прогноз с изученных на менее изученные регионы с близким геологическим строением.

¹ Исследование выполнено в ИППИ РАН частично за счет гранта РФФИ 20-07-00445.

В некоторых случаях экспертное решение задачи прогноза может представлять собой только предварительный результат исследования. Тогда экспертную прогнозную карту целесообразно использовать как экспертные оценки прогнозируемой величины, чтобы затем найти прогнозирующую функцию и построить новую прогнозную карту. Полученные результаты помогают эксперту более аргументировано объяснить свое решение и позволяют откорректировать окончательную версию прогнозной карты [2].

Одна из задач пространственного прогноза состоит в выделении и картировании зон с редкими аномальными проявлениями природных процессов. В науках о земле к этим задачам относятся идентификация зон потенциальных опасностей геологических процессов и прогноз полезных ископаемых.

При анализе данных нам известны выборка примеров мест аномальных проявлений и ограниченный набор пространственных свойств геологической среды. Задача состоит в том, чтобы по имеющимся данным идентифицировать зоны, в которых возможен рассматриваемый тип аномальных явлений. Сложность задач заключается в необходимости обнаружения таких зон по малому количеству примеров аномальных явлений и по неполным данным о свойствах геологической среды, а социальные и экономические риски обусловлены возможными ошибками пропуска целевых и добавления ложных зон в карту анализируемой территории.

Задача идентификации аномальных геологических зон связана с обнаружением редких аномальных объектов. Такие задачи решаются с помощью методов машинного обучения, предназначенным одноклассовой классификации данных [3, 4, 5]. Предлагаемый подход основан на двух методах одноклассовой классификации: методе минимальной области тревоги [6] и методе предпочтений [7].

В методе минимальной области тревоги задача формулируется следующим образом. Пусть имеется множество объектов. Объект описывается набором своих свойств, выраженных в числовом виде (вектором признаков). Значения свойств объектов, близкие к максимальным или минимальным, имеют малую вероятность. Среди множества объектов имеются аномальные объекты. Они отличаются от остальных объектов тем, что значения некоторых из их свойств близки к максимальным или минимальным возможным. Для простоты объяснения будем считать, что свойства аномальных объектов близки только к максимальным значениям. Пусть имеется учебная выборка из аномальных объектов (прецеденты). Представляется естественным классифицировать объект как аномальный, если соответствующий ему вектор признаков больше или равен покомпонентно одному из векторов, соответствующих прецеденту (условие монотонности). Однако описание свойств объектов обычно является неполным. Поэтому у некоторых прецедентов отсутствуют свойства, которые близки к максимальным значениям. Для таких прецедентов число объектов, классифицированных по ним как аномальные, может оказаться весьма большим, а сами объекты с большой вероятностью будут ошибочно классифицированы как аномальные.

Задача состоит том, чтобы обнаружить наибольшее число прецедентов при условии, что число объектов, классифицированных по ним как аномальные, не больше заданного. Идея алгоритма состоит в следующем. Пусть каждому прецеденту соответствует функция, которая принимает значение 1 для векторов пространства признаков больше или равных покомпонентно вектору прецедента. По условию монотонности предполагается, что объекты, векторам которых присвоено значение 1, являются аномальными. Таким образом, получаем, что каждому прецеденту соответствует множество классифицированных по нему аномальных объектов. Далее алгоритм выбирает максимальное число прецедентов, для которых мощность объединения этих множеств не превосходит заранее заданное число. Решающее правило классифицирует объекты только по отобранным прецедентам. Для остальных прецедентов следует искать новые отличительные свойства и добавлять их в пространство признаков.

Алгоритм минимальной области тревоги является непараметрическим. Он анализирует данные и отбирает аномальные точки пространства признаков, которые затем отображаются в географических координатах в виде карты аномальных зон. Полученное этим методом знание является невербализованным. Метод предпочтений позволяет аппроксимировать полученное решение с помощью достаточно простого логического правила, которое определяет аномальную область в терминах анализируемых свойств геологической среды. Это дает возможность специалистам предметной области интерпретировать полученное решение и применить правило прогноза для нахождения подобных зон в других регионах с близким геологическим строением. Кроме того, интерпретируемые эмпирические количественные закономерности могут представлять самостоятельный теоретический интерес для построения причинно-следственных моделей прогноза.

Метод идентификации аномальных зон рассматривается в разделе 2. В разделе 3 изложены результаты применения данного метода для выделения зон ВОЗ с магнитудами $m \geq 6.0$ и $m \geq 5.5$ для Кавказского региона и для регионального прогноза рудных месторождений Болгарии и Камчатки.

2. МЕТОД ИДЕНТИФИКАЦИИ АНОМАЛЬНЫХ ГЕОЛОГИЧЕСКИХ ЗОН

Пусть для анализируемой области известна выборка мест аномальных проявлений $q = 1, \dots, Q$ (*целевые события*) и набор пространственных сеточных полей признаков $\mathbf{F}_i, i = 1, \dots, I$, представляющих свойства геологической среды. Поля $\mathbf{F}_i$ заданы в единой координатной сетке. Значения полей в узлах сетки $n = 1, \dots, N$ соответствуют точкам I -мерного пространства признаков $\mathbf{f}^{(n)} \in \mathbb{R}^I$. Точки пространства признаков $\mathbf{f}^{(q)} = \{f_i^{(q)}\}$, которые соответствуют узлам сетки полей признаков, ближайшим к целевым событиям, будем называть *прецедентами*. Задача состоит в том, чтобы по выборке прецедентов и полям признакам найти *поле тревоги* $\Phi(\mathbf{F}_i)$, которое определяет аномальные зоны, и вычислить логическое правило $\Psi(\mathbf{F}_i)$, которое эти зоны аппроксимирует.

Модель данных содержит два предположения.

1. *Условие аномальности*: места аномальных проявлений приурочены к зонам, в которых значения некоторых полей признаков маловероятны и близки к максимальным (или минимальным).
2. *Условие монотонности*: точки пространства признаков, которые покомпонентно больше или равны (или меньше или равны) прецедента $\mathbf{f}^{(q)}$ целевого события q , также могут быть прецедентами аналогичных целевых событий.

Для упрощения объяснения будем считать, что аномальные зоны характеризуются тем, что в них некоторые поля признаков принимают значения, близкие к максимальным. Тогда условие монотонности состоит в том, что, если $\mathbf{f}^{(n)} = \{f_i^{(n)}\} \geq \{f_i^{(q)}\}$, то $\mathbf{f}^{(n)}$ также является прецедентом аналогичного события.

Будем называть ортантом $O^{(q)}$ с вершиной в точке $\mathbf{f}^{(q)}$ область пространства признаков, точки которой покомпонентно больше или равны $\mathbf{f}^{(q)}$, т.е. $O^{(q)} = \{\mathbf{f}^{(n)} \in \mathbb{R}^I : f_i^{(n)} \geq f_i^{(q)}, i = 1, 2, \dots, I\}$. Точки $\mathbf{f}^{(n)} \in O^{(q)}$ будем называть *базовыми точками* целевого события q . Обозначим W_2 — множество всех узлов сетки географической области анализа, $|W_2| = S$; W_1 — множество всех прецедентов, $|W_1| = Q$; $W_2^{(q)}$ — подмножество узлов координатной сетки, соответствующих базовым точкам целевого события q , $|W_2^{(q)}| = S^{(q)}$; $W_1^{(q)}$ — подмножество базовых точек целевого события q , состоящее из прецедентов других целевых событий, $|W_1^{(q)}| = Q^{(q)}$; $P_2^{(q)} = S^{(q)}/S$ — вероятность базовых точек целевого события q (объем тревоги

события q); $P_1^{(q)} = Q^{(q)}/Q$ — вероятность прецедентов среди базовых точек целевого события q ; $L^{(q)} = P_1^{(q)}/P_2^{(q)}$ — отношение правдоподобия для целевого события q .

Качество решения определяют два показателя: (1) вероятность обнаружения U , равная доле обнаруженных целевых событий Q^* от всех целевых Q событий, $U = Q^*/Q$, и (2) объем тревоги V , равный доле числа всех узлов сетки в аномальных зонах S^* от числа всех узлов сетки S области анализа, $V = S^*/S$. Из определения можно видеть, что объем тревоги V равен вероятности обнаружения целевых событий случайными зонами, состоящими из $S^* = VS$ узлов сетки. Значения $\varphi(\mathbf{f})$ поля тревоги $\Phi(\mathbf{F}_i)$ в узлах сетки равны объемам тревоги. Часть узлов сетки поля тревоги S_0^* со значениями $\varphi(\mathbf{f}) = \{V: V \leq V_0 = S_0^*/S\}$ составляет область тревоги. Область тревоги определяет карту аномальных зон. Величину V_0 будем называть объемом области тревоги. Тестируемые целевые события обнаружены, если ближайшие к ним узлы сетки попадают в область тревоги. Наглядным представлением качества прогноза является зависимость $U(V)$, которая практически совпадает с ROC-кривой [8].

Алгоритм обучения является оптимальным, если он строит такое поле тревоги $\Phi(\mathbf{F}_i)$, которое на учебной выборке для любого значения V обеспечивает максимальное значение U . В нашем случае такое решение требует большого перебора из-за того, что множества базовых точек, относящихся к различным прецедентам, пересекаются. Поэтому мы рассматриваем решения, близкие к оптимальному.

Алгоритм обучения состоит из трех шагов:

1. Сгенерировать учебную выборку $\{\mathbf{f}^{(q)}, P_2^{(q)}, L^{(q)}\}$.
2. Вычислить поле тревоги $\Phi(\mathbf{F}_i)$.
3. Вычислить логическое правило, которое аппроксимирует область тревоги.

Имеются три наиболее существенные версии алгоритма обучения. Версии отличаются только на шаге 2. Первая версия алгоритма является наименее вычислительно-ёмкой. В ней поле тревоги определяется последовательностью, которая задается заранее. В алгоритме могут задаваться два типа последовательностей: (А) последовательность, в которой увеличиваются значения объемов тревоги объемов целевых событий $P_2^{(1)} \leq P_2^{(2)} \leq \dots \leq P_2^{(q)} \leq \dots \leq P_2^{(Q)}$, либо (Б) последовательность, в которой уменьшаются значения отношений правдоподобия $L^{(1)} \geq L^{(2)} \geq \dots \geq L^{(q)} \geq \dots \geq L^{(Q)}$. Вторая версия алгоритма позволяет оптимизировать поле тревоги так, чтобы при каждом увеличении значения вероятности обнаружения U на $1/Q$ для вычисления значений поля тревоги выбиралась такая очередная точка $\mathbf{f}^{(q)}$, для которой объем тревоги увеличивается минимально, либо отношение правдоподобия уменьшается минимально. Третья версия алгоритма позволяет оптимизировать прогнозное поле так, чтобы оно обнаруживало максимальное число событий при объеме тревоги не более заданного. Для тестирования выбраны первая и вторая версии алгоритма.

Рассмотрим первую версию алгоритма метода идентификации аномальных зон, в которой используется последовательность объемов тревоги целевых событий. Вариант алгоритма с использованием последовательности отношений правдоподобия аналогичен.

1. Сгенерировать учебную выборку $\{\mathbf{f}^{(q)}, P_2^{(q)}, L^{(q)}\}$. Упорядочить прецеденты $\mathbf{f}^{(q)}$, $q = 1, \dots, Q$, в соответствии с увеличением объемов тревоги целевых событий $P_2^{(1)} \leq P_2^{(2)} \leq \dots \leq P_2^{(q)} \leq \dots \leq P_2^{(Q)}$ (или для варианта «Б» в соответствии с уменьшением значений правдоподобия $L^{(1)} \geq L^{(2)} \geq \dots \geq L^{(q)} \geq \dots \geq L^{(Q)}$).
2. Вычислить поле тревоги $\Phi(\mathbf{F}_i)$.
 - 2.1. Присвоить всем узлам сетки поля тревоги Φ значение 1 (или для варианта «Б» — значение 0). Изменить значение 1 поля Φ на $V(1) = P_2^{(1)}$ в множестве $W_2^{(1)}$ узлов сетки, соответствующих базовым точкам прецедента $\mathbf{f}^{(1)}$; заменить значение 1 на $V(2) =$

$|W_2^{(1)} \cup W_2^{(2)}|/S$ в узлах сетки, относящихся базовым точкам прецедента $\mathbf{f}^{(2)}$; заменить значение 1 на $V(3) = |W_2^{(1)} \cup W_2^{(2)} \cup W_2^{(3)}|/S$ в узлах сетки, относящихся базовым точкам прецедента $\mathbf{f}^{(3)}$, и далее последовательно заменять значения 1 на соответствующие значения объемов тревоги. Полученное поле Φ принимает значения $V(1) \leq V(2) \leq \dots \leq V(q) \leq \dots \leq V(Q)$ или 1. Значения $V(1)$ относятся к узлам сетки из множества $W_2^{(1)}$, $V(2)$ относится к узлам сетки из множества $W_2^{(2)} \setminus W_2^{(1)}$, $V(3)$ относится к узлам сетки из множества $W_2^{(3)} \setminus (W_2^{(2)} \cup W_2^{(1)})$, и т.д. Таким образом, поле Φ определяет значения объема тревоги для всех узлов сетки, в которые попадают целевые события. Эта связь задает ROC-кривую $U(V)$.

- 2.2. Отобрать целевые события, аномальные свойства которых достаточно полно представлены значениями полей признаков. Для этого по ROC-кривой специалист назначает максимальную величину объема тревоги V_0 , начиная от которой целевые события считаются не обнаруженными. В географических координатах узлы сетки поля тревоги со значениями меньше или равными V_0 определяют область тревоги и множество успешно обнаруженных целевых событий $q = 1, 2, \dots, Q_0 \leq Q$.
3. Вычислить логическое правило, которое аппроксимирует область тревоги.
 - 3.1. Построить матрицу принадлежности $\mathbf{M} = \|m_{pq}\|$ для ортантов с объемами тревоги $V \leq V_0$. Строки матрицы обозначают ортанты прецедентов $O^{(p)}$ и дополнительные ортанты $O^{(r)}$, а столбцы обозначают прецеденты $\mathbf{f}^{(q)}$. Элемент матрицы $m_{pq} = 1$, если $\mathbf{f}^{(q)} \in W^{(p)}$, т.е., если ортант $O^{(p)}$, покрывает прецедент $\mathbf{f}^{(q)} = \{(f_i^{(q)}): f_i^{(p)} \geq f_i^{(q)}, i = 1, 2, \dots, I\}$ и $m_{pq} = 0$ в противном случае. Можно видеть, что строка матрицы p соответствует конъюнкции: $(f_1 \geq f_1^{(p)}) \cap (f_2 \geq f_2^{(p)}) \cap \dots \cap (f_I \geq f_I^{(p)})$. Совокупность строк матрицы соответствует дизъюнкциям конъюнкций.
 - 3.2. Определить подмножество из минимального числа ортантов, которые содержат все прецеденты $q = 1, 2, \dots, Q_0$. Для этого надо в матрице принадлежности $\mathbf{M}$ найти подмножество из минимального числа строк, для которого в каждом столбце матрицы оставалась хотя бы одна единица. Минимизация строк матрицы принадлежности выполняется с помощью метода минимизации дизъюнктивных нормальных форм булевых функций [9].

Обучение в задачах идентификации аномальных геологических зон обычно выполняется по весьма ограниченному числу примеров аномальных событий. При этом число строк матрицы, оставшихся после минимизации, может оказаться слишком большим для наглядной интерпретации области тревоги в виде дизъюнкции конъюнкций. Для увеличения надежности решения и для упрощения логического правила, для объяснения области тревоги выполняется коррекция. Рассмотрим два интерактивных метода коррекции.

1. *Коррекция области тревоги.* Область тревоги обнаруживает целевые события с разным уровнем надежности. Надежность того, что некоторая точка пространства признаков $\mathbf{g}$ является аномальной тем больше, чем большее число ортантов на этапе обучения покрывает эту точку. Наименьшее число таких покрытий имеют точки, близкие к границе аномальной области в пространстве признаков.

Для увеличения надежности решения выполняется коррекция вершин ортантов со значениями объемов тревоги, начиная $V_{00} < V_0$. Значение V_{00} выбирает специалист. При коррекции в выборку $\{\mathbf{f}^{(q)}, P_2^{(q)}, L^{(q)}\}$ добавляются точки $\mathbf{f}^{(r)} = \{f_i^{(q)} - c\sigma_i\}$, где $c \geq 0$ — коэффициент, $\sigma_i, i = 1, 2, \dots, I$ — среднеквадратичные отклонения значений полей признаков $\mathbf{F}_i$ в пределах области анализа. Далее для построения поля тревоги повторяются пункты алгоритма 2.1 и 2.2. Область тревоги может увеличиваться только, начиная от значений поля более

V_{00} . Значение V_{00} выбирается таким, чтобы дополнительные точки увеличивали число покрытий только для тех точек пространства признаков, которые находятся вблизи границы аномальной области.

2. *Коррекция логического правила.* Часто число строк матрицы, оставшихся после минимизации, оказывается слишком большим для наглядной интерпретации области тревоги в виде дизъюнкции конъюнкций. Алгоритм коррекции создает логическое правило, состоящее из заданного специалистом числа конъюнкций, которое покрывает все отобранные по объему тревоги V_0 прецеденты $q = 1, 2, \dots, Q_0$. Алгоритм работает с вершинами ортантов, отобранными после минимизации матрицы $\mathbf{M}$. Для всех пар ортантов с вершинами в точках $\mathbf{f}^{(q)}$ и $\mathbf{f}^{(r)}$ строятся новые ортанты $O^{(p)}$ с вершинами $\mathbf{f}^{(p)} = \{f_i^{(p)}\}$, где $f_i^{(p)} = \min(f_i^{(q)}, f_i^{(r)})$. Очевидно, что ортант $O^{(p)}$ покрывает все точки, которые покрывают ортанты $O^{(q)}$ и $O^{(r)}$. Из ортантов, полученных в результате объединения, отбирается ортант с наименьшим объемом тревоги $V(p) = |W_2^{(q)} \cup W_2^{(r)}|/S$, а ортанты $O^{(q)}$ и $O^{(r)}$ удаляются. При этом число оставшихся ортантов уменьшается на единицу. Процедура повторяется до тех пор, пока число оставшихся ортантов не окажется равным заранее заданному. В результате коррекции увеличивается объем области тревоги. Если полученный объем тревоги является неприемлемым, то число ортантов приходится увеличить. Приемлемая величина объема тревоги и число ортантов подбираются с помощью нескольких итераций. Далее с помощью лингвистических переменных и заранее заготовленных шаблонов полученное логическое правило можно представить в виде достаточно простого текстового выражения.

3. ТЕСТИРОВАНИЕ

При тестировании рассмотренного метода идентификации аномальных зон применялись первая и вторая версии алгоритма. В качестве меры аномальности использовалось отношение правдоподобия. Применение в качестве меры аномальности объема тревоги дает близкие результаты. Использовалась система GeoTime 3 (<http://www.geo.iitp.ru/GT3/>).

3.1. Построение карты зон возможных очагов землетрясений

Для построения карты зон возможных очагов землетрясений используются две гипотезы: (1) гипотеза о повторяемости землетрясений и (2) гипотеза о том, что области очагов сильных землетрясений определяются свойствами геологической среды. Первая гипотеза основана на хорошо известных фактах о повторяемости землетрясений. Но история сейсмических наблюдений является очень короткой по отношению к скорости тектонических процессов, а землетрясения с магнитудами, близкими к максимальным, происходят относительно редко. Поэтому использование только первой гипотезы может приводить к пропускам зон ВОЗ из-за относительно короткого периода наблюдений [10, 11]. Вторая гипотеза компенсирует эффект недостаточной представительности имеющихся сейсмических наблюдений [12, 13]. Однако обычно доступные для анализа свойства геологической среды являются неполными. Это значительно усложняет нахождение закономерностей по эмпирическим данным.

Динамические процессы в Кавказском регионе определяются конвергенцией Аравийской и Евразийской плит [14]. Сейсмичность региона в большой степени определяются зонами развития надвигов, сдвигов и сбросов. В работе [15] на моделях показано, что напряжения, способные вызвать смещения по надвигам, сдвигам и сбросам, соотносятся приблизительно как 15:5:1. Указанные зоны могут отображать распределение тектонических напряжений, а следовательно, и распределение потенциала, необходимого для смещений земной коры. Поэтому эти зоны могут использоваться в качестве основы для модели связи региональных характеристик геологической среды с сильными землетрясениями.

Для решения задачи анализировались данные, подготовленные в [16]: высоты поверхностей дневного рельефа, консолидированного фундамента и Мохоровичича, амплитуды неотектонических движений, аномалии силы тяжести в редукции Буге, изостатические и глубинные, градиент скорости вертикальных тектонических движений за постсарматское время ($T=17$ млн. лет), отклонения времен пробега первых коровых волн и продольных волн в верхней мантии, тепловой потока, аномалии магнитного поля, разломы, ранжированные по возрасту заложения, по возрасту последней активизации, по типу движений и по тектонической значимости, а также каталоги землетрясений [17, 18].

Зоны возможных очагов землетрясений с магнитудами $m \geq 6.0$

В области анализа имеется 20 эпицентров землетрясений с целевыми магнитудами $m \geq 6.0$. Для вычисления зон ВОЗ выбраны три следующие три поля признака: $\mathbf{F}_1$ — поле градиента скорости вертикальных тектонических движений за постсарматское время [1/год], $\mathbf{F}_2$ — поле модуля градиента высот поверхности земли [м/км] и $\mathbf{F}_3$ — поле градиента аномалий силы тяжести в редукции Буге [мГал/ км]. Поля $\mathbf{F}_1$, $\mathbf{F}_2$ и $\mathbf{F}_3$ показаны на Рис 1. Будем считать, что размер зон ВОЗ составляет 25–35% от области анализа.

Для более надежного решения применялась коррекция области тревоги. При коррекции к основным ортантам с вершинами в точках прецедентов учебной выборки добавлены ортанты с вершинами, смещенными относительно прецедентов на $0.5\sigma_i$. На Рис. 2 показаны зависимости $U(V)$, полученные при тестировании на обучающей выборке и с помощью k -кратной процедуры кросс-валидации при $k=2$. Можно видеть, что для объема области тревоги $V_0 = 0.2$ при обучении по полной выборке в зоны ВОЗ попадает 0.9 эпицентров целевых землетрясений, а при кросс-валидации объема области тревоги $V_0 = 0.35$ соответствует $U = 0.85$.

Полученное логическое решающее правило имеет вид:

ЕСЛИ

$$\begin{aligned} & (f_1 \geq 10.0 \wedge f_2 \geq 38.6 \wedge f_3 \geq 0.33) \\ & \vee (f_1 \geq 2.8 \wedge f_2 \geq 186.3 \wedge f_3 \geq 0.28) \\ & \vee (f_1 \geq 3.9 \wedge f_2 \geq 5.7 \wedge f_3 \geq 1.0), \end{aligned}$$

ТО возможны очаги землетрясений с магнитудой $m \geq 6.0$.

На Рис. 3 показаны карты зон ВОЗ, полученные при обучении по полной выборке и при помощи апроксимации решения логическим правилом. В первом случае область тревоги соответствует $V = 0.20$, во втором случае область тревоги соответствует $V = 0.35$. На рисунках кружками показаны эпицентры землетрясений, которые попали в области тревоги, а закрашенными кружками — те, которые не попали.

Зоны возможных очагов землетрясений с магнитудами $m \geq 5.5$

В области анализа имеется 48 эпицентров землетрясений с целевыми магнитудами $m \geq 5.5$. Для вычисления зон ВОЗ выбраны те же три поля признака, что и для предыдущего примера. При вычислении зон ВОЗ предполагалось, что размер области тревоги составляет около половины от области анализа. Для коррекции решения к основным ортантам с вершинами в точках прецедентов учебной выборки добавлены ортанты с вершинами, смещенными относительно прецедентов на σ_i . На Рис. 4 показаны зависимости $U(V)$, полученные при тестировании на обучающей выборке и с помощью k -кратной процедуры кросс-валидации при $k = 4$. Можно видеть, что для объема области тревоги $V_0 = 0.5$ при обучении по полной выборке в зоны ВОЗ попадает $U = 0.8$ от числа всего эпицентров целевых землетрясений, а при кросс-валидации $U = 0.67$.

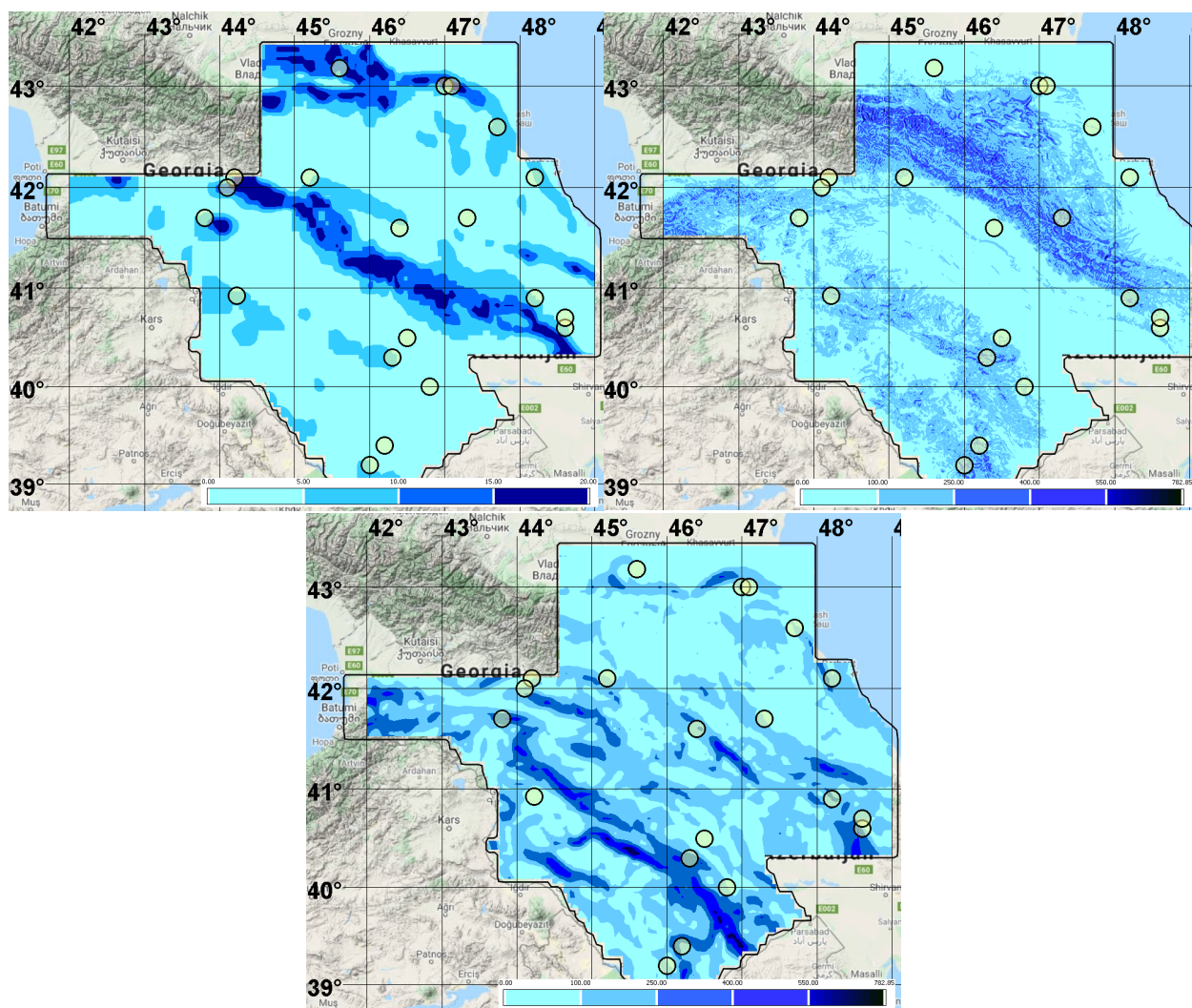


Рис. 1. В области анализа поля признаки прогноза и учебная выборка эпицентров землетрясений с $m \geq 6.0$. Слева сверху F_1 — поле градиента скорости вертикальных тектонических движений за постсарматское время [1/год], справа сверху F_2 — поле модуля градиента высот поверхности земли [м/км] и снизу F_3 — поле градиента аномалий силы тяжести в редукции Буге [мГал/км].

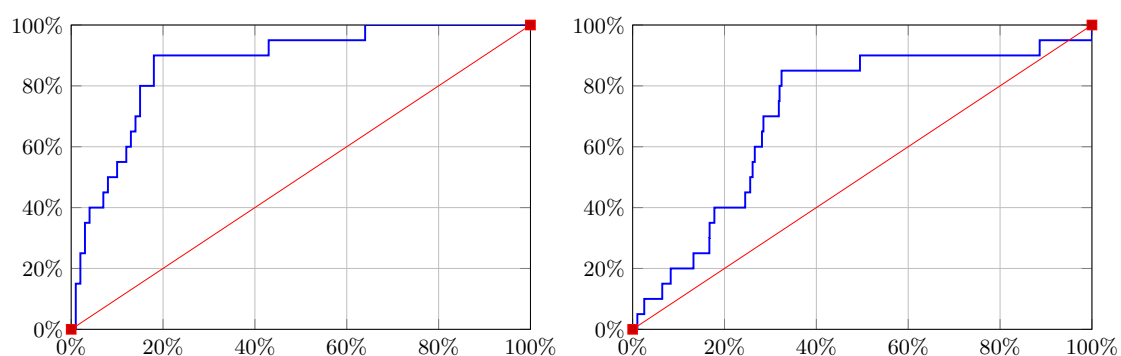


Рис. 2. Графики зависимостей $U(V)$, полученные при тестировании на обучающей выборке и с помощью k -кратной процедуры кросс-валидации при $k=2$.

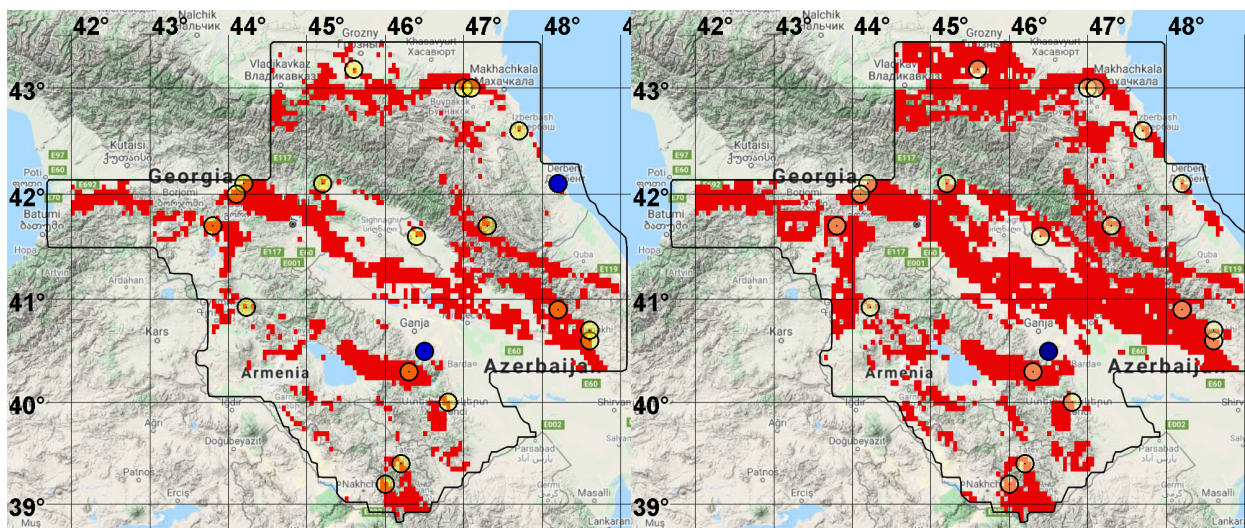


Рис. 3. Зоны ВОЗ. Слева решение по алгоритму обучения, при объеме тревоги $V = 0.20$ вероятность обнаружения $U = 0.9$. Справа решение при помощи решения логического правила, при объеме тревоги $V = 0.35$. Кружки — эпицентры землетрясений с магнитудами $m \geq 6.0$, которые попали в область тревоги, а закрашенные кружки — эпицентры, которые не попали.

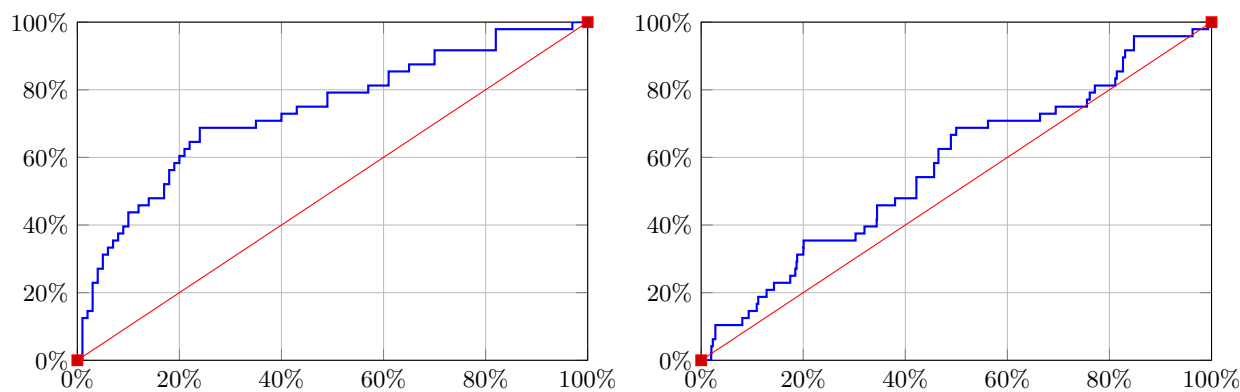


Рис. 4. Графики зависимостей $U(V)$, полученные при тестировании на обучающей выборке и с помощью k -кратной процедуры кросс-валидации при $k=4$.

Полученное логическое решающее правило имеет вид:

ЕСЛИ

$$\begin{aligned}
 & (f_1 \geq 10.5 \wedge f_2 \geq 79.64 \wedge f_3 \geq 0.52) \\
 & \vee (f_1 \geq 2.2 \wedge f_2 \geq 28.02 \wedge f_3 \geq 0.47) \\
 & \vee (f_1 \geq 0.63 \wedge f_2 \geq 219.3 \wedge f_3 \geq 0.71) \\
 & \vee (f_1 \geq 2.00 \wedge f_2 \geq 5.71 \wedge f_3 \geq 1.71),
 \end{aligned}$$

ТО возможны очаги землетрясений с магнитудой $m \geq 5.5$.

На Рис. 5 показаны карты зон ВОЗ, полученные по алгоритму минимальной области тревоги и по аппроксимации зон ВОЗ логическим правилом. В первом случае область тревоги соответствует $V = 0.50$, во втором случае область тревоги соответствует $V = 0.59$. На рисунках кружками показаны эпицентры землетрясений, которые попали в область тревоги, а закрашенными кружками — те, которые не попали. Размер кружков показывает магнитуды землетрясений от 5.5 до 5.9 и больше 6.0.

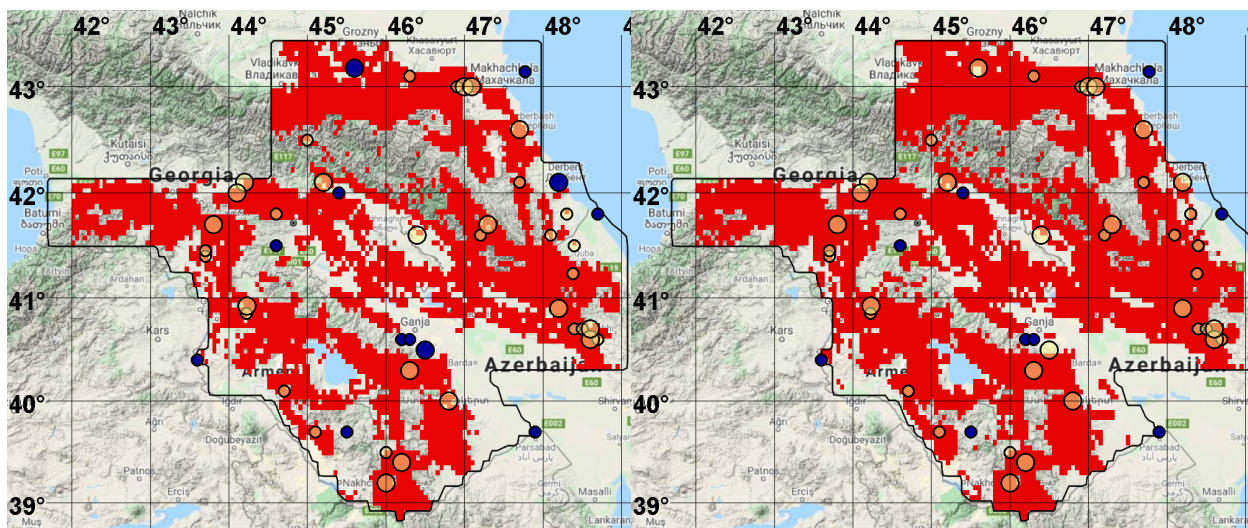


Рис. 5. Зоны ВОЗ. Слева решение по алгоритму обучения, при объеме тревоги $V = 0.50$ вероятность обнаружения $U = 0.8$. Справа решение при помощи решения логического правила, при объеме тревоги $V = 0.59$. Малые кружки — эпицентры землетрясений с магнитудами $5.9 \geq m \geq 5.5$, более крупные $m \geq 6$. Кружками отмечены эпицентры землетрясений, которые попали в область тревоги, а закрашенными кружкам — эпицентры, которые не попали.

3.2. Региональный прогноз рудных месторождений

Гидротермальные свинцово-цинковые месторождения Болгарии

Задача регионального анализа и прогноза рудных месторождений исследовалась в работах [2, 7, 19]. Большая часть свинцово-цинковых месторождений Болгарии принадлежит к гидротермальному среднетемпературному типу. Наиболее крупными месторождениями являются Мадан, Звездел, Рудозем, Устрем, Руен и др. Модель образования месторождений этого типа основана на работах [20, 21, 22, 23]. Она состоит в следующем. Консолидированная кора толщиной 35–40 км подвергается кайнозойской магматической активизации. Гипабиссальные интрузии среднего состава, представленные серией даек и штоков, обеспечивают мобилизацию гидротермальной деятельности и формирование рудных жильных тел на глубинах 1000–1500 м. Протяженность оруденения по глубине составляет около 1000 м. Для того чтобы оруденение стало доступным для эксплуатации, район должен испытать тектоническое поднятие и эрозию. Протяженность вскрытых рудных тел составляет первые километры при ширине до нескольких десятков метров.

Данной модели соответствуют три группы свойств геологической среды. Первая группа охватывает вещественные параметры месторождений и вмещающих пород, а также геофизические свойства, связанные с веществом. Она включает характеристики магнитного поля, состава магматических образований, аномалий силы тяжести в редукции Буге. Вторая группа объединяет структурные свойства вмещающей среды: градиенты потенциальных полей, плотность разломов, гетерогенность, глубину поверхности Мохоровичича, мощность гранитно-метаморфического слоя и др. Третью группу составляют геодинамические признаки, определяющие вскрытие месторождений. К ним относятся неотектонические, движения, характеристики рельефа, изостатические аномалии силы тяжести, тепловой поток и др.

Результаты прогноза могут быть уточнены при дополнении исходных данных геохимическими признаками, использовании более детальных геологических материалов и привлечении более совершенных моделей формирования свинцово-цинковых месторождений. Отобранные при нахождении прогнозирующей функции признаки позволяют уточнить модель образова-

ния свинцово-цинковых месторождений Болгарии. Предполагается, что существенная роль принадлежит признакам, связанным с современным геодинамическим состоянием, таким как амплитуда неотектонических движений и аномалии силы тяжести, обычно не используемым при прогнозировании месторождений.

В области анализа имеется 26 точек, соответствующие местам свинцово-цинковых месторождений. Для вычисления зон, где такие месторождения возможны, выбраны следующие три поля признаков:

F₁ — возраст последних проявлений магматизма [24], кодировка: 0 — отсутствие проявлений магматизма, 1 — последние проявления магматизма в докембрии, 2 — в карбоне, 3 — в перми, 4 — в триасе, 5 — в юре, 6 — в мелу, 7 — в палеоцене, 8 — в эоцен-олигоцене, 9 — в миоцене. Перспективность точки считается тем выше, чем моложе последние проявления магматизма (группа вещественных признаков среды).

F₂ — гравитационное поле в редукции Буге в миллигалах [25]. Положительные аномалии поля связаны с интрузивными телами. В районе их контактов могут находиться месторождения (в том числе и Pb,Zn). Большие модули градиента поля связаны с глубинными разломами, по которым могло подниматься мантийное вещество, впоследствии застывавшее. В результате выплавления некоторых металлов (например Pb,Zn) могло образоваться месторождение. Состав рудовмещающих пород позволяет считать наиболее перспективными ячейки со средними значениями поля (группа вещественных признаков среды).

F₃ — амплитуда новейших тектонических движений в метрах. Области опусканий и области поднятий с амплитудой до 200 м не перспективны. Поднятия с амплитудой 200–500 м — слабо перспективны, 500–700 м — высоко перспективны. Поднятия с (10–30 км) амплитудами более 1700 м слабо перспективны в связи с разрушением эрозией значительной части месторождений (группа геодинамических признаков среды).

На Рис. 6 показаны зависимости $U(V)$, полученные при тестировании на обучающей выборке и с помощью k -кратной процедуры кросс-валидации при $k=2$. При этом для более надежного решения применялась коррекция области тревоги. При коррекции к основным ортантам с вершинами в точках прецедентов учебной выборки добавлены ортанты с вершинами, смещенными относительно прецедентов на $0.7\sigma_i$. Можно видеть, что для объема области тревоги $V_0 = 0.075$ при обучении по полной выборке в область тревоги попадаю 100% месторождений, а при кросс-валидации 85%.

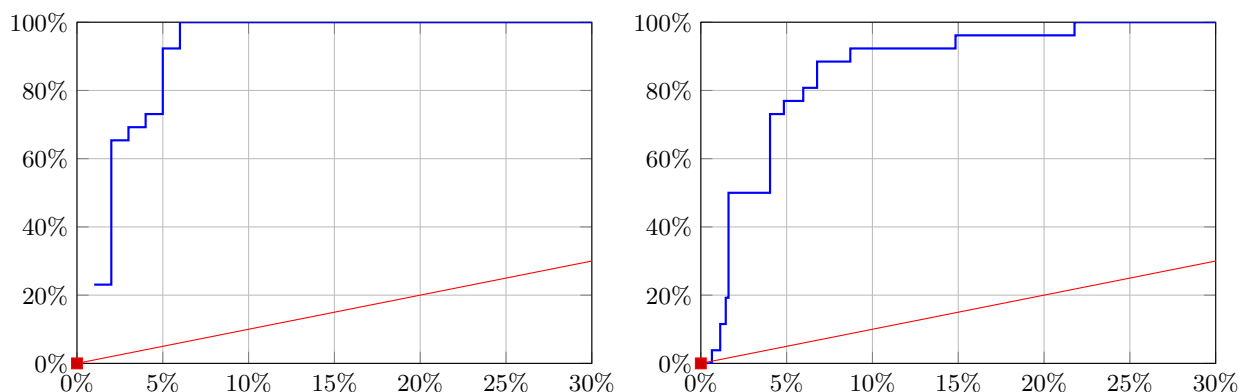


Рис. 6. Графики зависимостей $U(V)$, полученные при тестировании на обучающей выборке и с помощью k -кратной процедуры кросс-валидации при $k=2$.

Полученное логическое решающее правило имеет вид:
ЕСЛИ

$$(f_1 \geq 7 \wedge f_2 \geq -60.1 \wedge f_3 \geq 300) \\ \vee (f_1 \geq 7 \wedge f_2 \geq 20.1 \wedge f_3 \geq 500),$$

ТО возможны гидротермальные свинцово-цинковые месторождения.

На Рис. 7 показаны карты месторождений, полученные при обучении по алгоритму минимальной области тревоги и при помощи решения логическим правилом. В первом случае область тревоги соответствует $V = 0.075$, во втором случае область тревоги соответствует $V = 0.85$. На рисунках кружками показаны свинцово-цинковые месторождения, которые использованы при обучении. Все они попали в область тревоги.

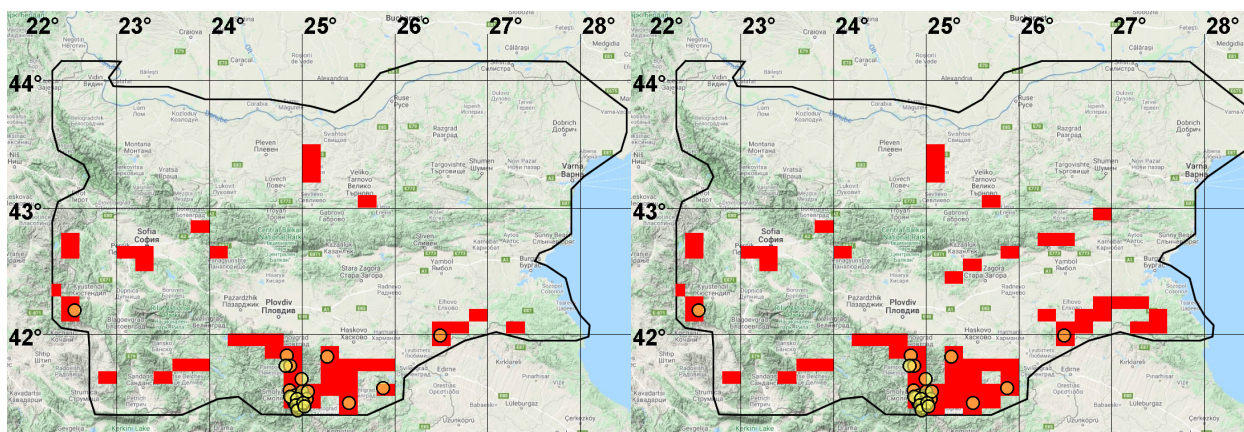


Рис. 7. Прогнозные карты зон свинцово-цинковых месторождений Болгарии, полученные при обучении по полной выборке: слева - результат алгоритма минимальной области тревоги, справа аппроксимация решения логическим правилом. Кружками отмечены месторождения учебной выборки.

Золото-серебряные месторождения Курило-Камчатского вулканического пояса

Золото-серебряные месторождения Курило-Камчатского вулканического пояса относятся к телетермальному типу. Месторождения формировались в период кайнозоя. Поэтому они могут быть приурочены к вулканическим породам кайнозойской эры, петрологический состав которых является кислым (более 70% SiO_2) или средним (от 70% до 50% SiO_2). Вулканогенные растворы, из которых формировалось золото, были низкотемпературными. Одна из причин — удаленность источников тепла, которые, возможно, связаны с вулканическим построением плиоцен-миоценового времени. Выход растворов на поверхность происходил в областях с тектоническими нарушениями земной коры. Можно предполагать, что эти области совпадают с зонами геологических разломов.

При решении задачи в основном использовались данные, подготовленные в ИГЕМ РАН в рамках проекта «Электронная Земля». Цифровые сеточные модели включают поля высот поверхности земли и вулканические породы кайнозоя, четвертичного, неогенового и палеогенового периодов. Векторные данные состоят из полигонов субвулканических интрузий Камчатского звена кайнозойского Курило-Камчатского вулканического пояса, разломов, границ вулканических поясов, точечных полей голоценовых и плиоценовых вулканических построек, а также золото-серебряные месторождений и проявлений.

Обучение производилось на 12 примерах месторождений. Они состоят из 4-х золоторудных собственных, 2-х полиметаллически-золото-серебряных с преобладанием свинца и цинка над медью и 6-и золотосеребряных месторождений. Для вычисления зон, где такие месторождения возможны выбраны следующие три поля признаков:

F_1 — высоты поверхности земли [м].

F_2 — расстояние до геологических разломов [км].

F_3 — расстояние до вулканических построек голоцена [км].

F_4 — расстояние до интрузивных пород кайнозоя [км].

На Рис. 8 показаны зависимости $U(V)$, полученные при тестировании на обучающей выборке и с помощью k -кратной процедуры кросс-валидации при $k=2$. Для более надежного решения применялась коррекция области тревоги. При коррекции к основным ортантам с вершинами в точках прецедентов учебной выборки добавлены ортанты с вершинами, смещенными относительно прецедентов на $0.5\sigma_i$. Можно видеть, что для объема области тревоги $V_0 = 0.02$ при обучении по полной выборке в область тревоги попадало 100% месторождений, а при кросс-валидации 58%.

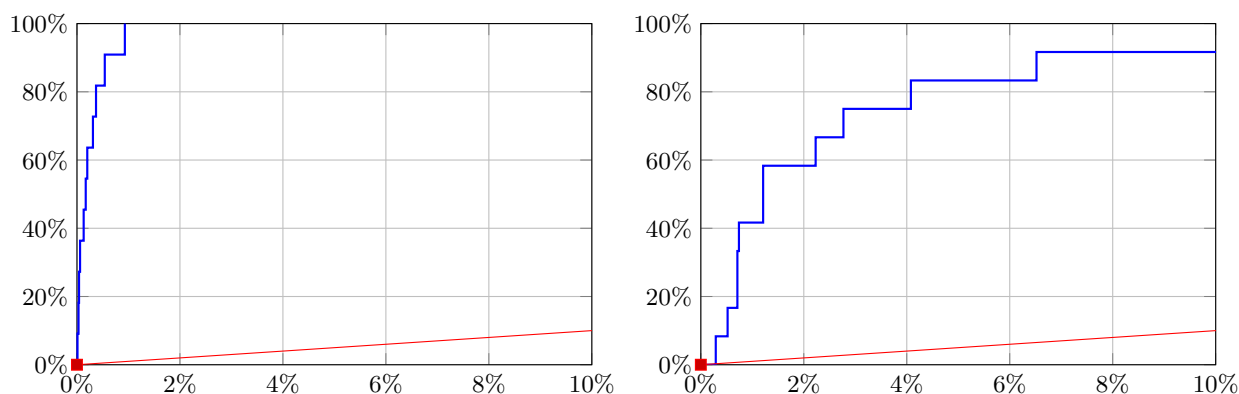


Рис. 8. Графики зависимостей $U(V)$, полученные при тестировании на обучающей выборке и с помощью k -кратной процедуры кросс-валидации при $k=2$.

Полученное логическое решающее правило имеет вид:

ЕСЛИ

$$(f_1 \geq 262 \wedge f_2 \leq 14.2 \wedge f_3 \leq 47.1 \wedge f_4 \leq 10.6) \\ \vee (f_1 \geq 441 \wedge f_2 \leq 0.8 \wedge f_3 \leq 105.9 \wedge f_4 \leq 48.9),$$

ТО возможны телетермальные золотосеребряные месторождения.

На Рис. 9 показаны карты месторождений, полученные при обучении по алгоритму минимальной области тревоги и при помощи решения логическим правилом, по полной выборке и по логическому правилу. В первом случае область тревоги соответствует $V = 0.093$, во втором случае область тревоги соответствует $V = 0.03$. На рисунках желтым цветом показаны свинцово-цинковые месторождения, которые использованы при обучении. Все они попали в область тревоги.

4. ЗАКЛЮЧЕНИЕ

Мы рассмотрели подход к решению одного из типов задач пространственного прогноза. Предлагаемое решение связано с нахождением эмпирических закономерностей в условиях

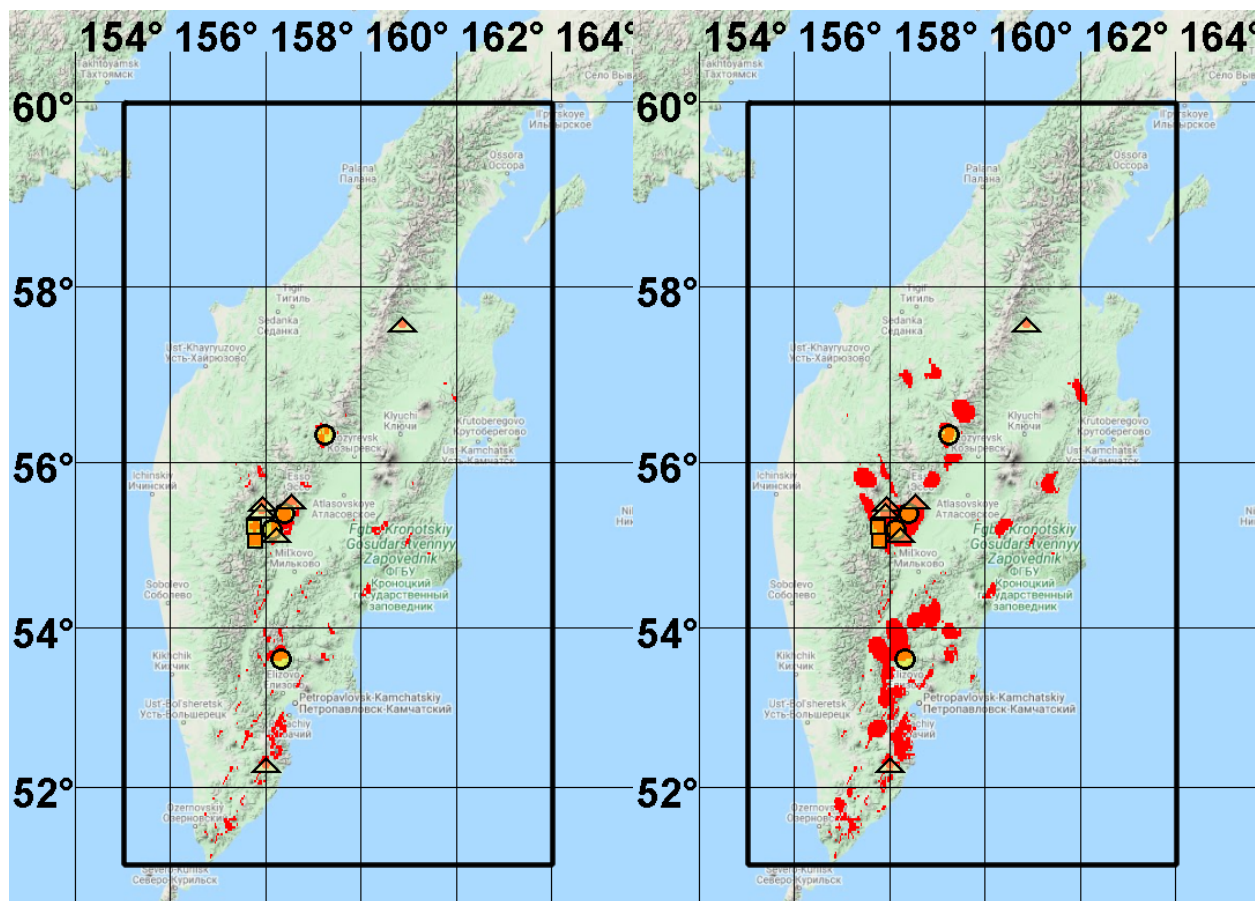


Рис. 9. Прогнозные карты зон золоторудных месторождений Камчатки, полученные при обучении по полной выборке: слева — результат алгоритма минимальной области тревоги, справа аппроксимация решения логическим правилом. Месторождения учебной выборки: кружки — золоторудные собственные, квадраты — полиметаллически-золото-серебряные с преобладанием свинца и цинка над медью, треугольники — золотосеребряные.

неполной информации. В этих задачах отсутствует готовая формула или математическая модель для нахождения прогноза, но предполагается наличие причинно-следственной связи между значением прогноза и значениями переменных, отображающих свойства прогнозирующей величины. При решении таких задач эксперты используют два типа рассуждений: (1) рассуждения по аналогии: объекты со сходными свойствами имеют близкие значения прогнозируемой величины и (2) рассуждения в виде правил: объекты, свойства которых подчиняются некоторому правилу, имеют близкие значения прогнозируемой величины.

Рассмотренный метод находит прогноз в виде области тревоги. Для этого используется непараметрический алгоритм обучения. Полученный результат имеет эмпирическое подтверждение, но не вербализирован. Для вербализации используется алгоритм аппроксимации области тревоги с помощью достаточно простой логической функции от свойств геологической среды. Этот результат понятен специалистам проблемной области и может быть использован для прогноза в регионах с аналогичными геологическими свойствами.

СПИСОК ЛИТЕРАТУРЫ

1. Schenk V., Schenkova Z., Gitis V.G. Features of geonomic prognostic functions for the maximum possible earthquake. *Natural hazards*, 1994, vol.10, no.1–2, pp.97–115.

2. Гитис В.Г., Ермаков Б.В. *Основы пространственно-временного прогнозирования в геоинформатике*. М.: Физматлит, 2004.
3. Bishop C.M. Machine learning and pattern recognition. *Information science and statistics*. Springer, Heidelberg, 2006.
4. Kotsiantis S.B., Zaharakis I., Pintelas P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 2007, vol.160, pp.3–24.
5. Khan S.S., Madden M.G. A survey of recent trends in one class classification. *Irish conference on artificial intelligence and cognitive science*, Berlin: Springer, Heidelberg, 2009, pp. 188–197.
6. Gitis V.G., Derendyaev A.B. Machine Learning Methods for Seismic Hazards Forecast. *Geosciences*, 2019, vol.9, no.7, pp.308.
7. Гитис В.Г. *Основы геоинформационного прогноза полезных ископаемых и анализа геологических процессов*. М.: ВИНТИ, 2012.
8. Bradley A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 1997, vol.30, no.7, pp.1145–1159.
9. Закревский А.Д. Оптимизация покрытий множеств. В кн.: *Логический язык для представления алгоритмов синтеза релейных устройств*, М.: Наука, 1966, pp.136–147.
10. Burton P.W. Seismic risk in southern Europe through to India examined using Gumbel's third distribution of extreme values. *Geophysical Journal International*, 1979, vol.59, no.2, pp.249–280.
11. Kijko A. Estimation of the maximum earthquake magnitude, m_{max} . *Pure and Applied Geophysics*, 2004, vol.161, no.8, pp.1655–1681.
12. Gelfand I.M., Guberman S.A., Keilis-Borok V.I., Knopoff L., Press F., Ranzman E.Y., Rotwain I.M., Sadovsky A.M. Pattern recognition applied to earthquake epicenters in California. *Physics of the earth and planetary interiors*, 1976, vol.11, no.3, pp.227–283.
13. Бунэ В. И., Горшков Г. П. (ред.) *Сейсмическое районирование территории СССР: Методические основы и региональное описание карты 1978 г.*, М.: Наука, 1980.
14. Philip H., Cisternas A., Gvishiani A., Gorshkov A. The Caucasus: an actual example of the initial stages of continental collision. *Tectonophysics*, 1989, vol.161, no.1–2, pp.1–21.
15. Sibson R.H. Frictional constraints on thrust, wrench and normal faults. *Nature*, 1974, vol.249, no.5457, pp.542–544.
16. Gitis V.G., Ermakov B.V., Ivanovskaya L.V., Osher B.V., Trofimov D.P., Shchukin Yu.K. The GEO Expert System: application for seismic hazard analysis of the caucasus region. In: *Cahiers du Centre Europeen de Geodynamique et de Seismologie*, Centre europeen de geodynamique et de seismologie, 1992, pp.136–147.
17. Kondorskaia N.V. and Shebalin N.V. *New catalog of strong earthquakes in the USSR from ancient times through 1977*, vol. 31, World Data Center A for Solid Earth Geophysics, 1982.
18. *Землетрясения в СССР в 1975–1988 гг.*, М.: Наука, 1990.
19. Гитис В.Г., Добрев Т.В., Ермаков Б.В., Кубарева Н.Ю., Щукин Ю.К. Применение экспертной системы ГЕО для регионального прогноза свинцово-цинковых месторождений Болгарии. *Геофизический журнал*, 1989, Том 11, №4, с.36–48.
20. Вольфсон Ф.Н., Дружинин Н.М. *Главнейшие типы рудных месторождений*, М.: Недра, 1975.
21. Т.Н. Шалдун и М.Г. Добровольская Рудные формации свинцово-цинковых месторождений. В кн.: *Рудные формации эндогенных месторождений*, М.: Наука, 1976, с.149–297.
22. Богданов Б., Димитров Д., Стайков М. и др. *Рудные формации Карпато-Балканской области*, М.: Недра, 1978.
23. Косыгин Ю.А., Кулиш Е.А. *Основные типы рудных формаций. Терминологический справочник*, М.: Недра, 1978.

24. Димитрова Е. *Карта магматических формаций Карпато-Балканской области 1:1 000 000*, София: КИПП по картография, 1978.
25. Добрев Т.Б., Шукин Ю.К. *Геофизические поля и сейсмичность восточной части Карпато-Балканского региона*, М.: Наука, 1974.

A method of abnormal geological zone identification

Gitis V.G., Derendyaev A.B., Petrov K.N.

The article proposes a new approach to the identification of zones with rare anomalous event of geological processes. The approach is based on two one-class classification methods of machine learning: the method of minimum area of alarm and the method of preference. the method of minimum area of alarm is nonparametric. It is trained on a sample of anomalous events and computes the field of anomalous zones. The knowledge obtained by this method is non-verbalized. The method of preference allows you to approximate the resulting solution using a simple logical rule that defines the anomalous region in terms of the analyzed properties of the geological environment. There are examples of the application of this approach: to find areas of possible foci of strong earthquakes, and for the regional forecast of deposits are considered.

KEYWORDS: machine learning, spatial forecast, seismic hazard, mineral deposits.