АКАДЕМИЯ НАУК СССР
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

На правах рукописи

КОРОЛЕВА
Зоя Ефимовна

АНАЛИЗ ЭФФЕКТИВНОСТИ НЕКОТОРЫХ КОМБИНАТОРНО-ЛОГИЧЕСКИХ АЛГОРИТМОВ РАСПОЗНАВАНИЯ (НА ПРИМЕРЕ РЕШЕНИЯ ГЕОЛОГИЧЕСКИХ ЗАДАЧ)

(05.13.03 – исследование операций и системный анализ)

На русском языке

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

МОСКВА, 1975

АКАДЕМИЯ НАУК СССР
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

На правах рукописи

КОРОЛЕВА
Зоя Ефимовна

АНАЛИЗ ЭФФЕКТИВНОСТИ НЕКОТОРЫХ КОМБИНАТОРНО-ЛОГИЧЕСКИХ АЛГОРИТМОВ РАСПОЗНАВАНИЯ (НА ПРИМЕРЕ РЕШЕНИЯ ГЕОЛОГИЧЕСКИХ ЗАДАЧ)

(05.13.03 – исследование операций и системный анализ)

На русском языке

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

МОСКВА, 1975

Работа выполнена в Вычислительном центре Академии наук СССР

Научный руководитель - кандидат физико-математических наук, старший научный сотрудник В.Б.КУДРЯВЦЕВ.

Официальные оппоненты:

доктор геолого-минералогических наук,
профессор Д.А.РОДИОНОВ

кандидат физико-математических наук
С.В.АЛЕШИН

Диссертация направлена на отзыв в Институт геологии рудных месторождений, петрографии, минералогии и геохимии АН СССР

Автореферат разослан "14" IV 1975 г.

Защита диссертации состоится "15" V 1975 г. в "15" часов на заседании Ученого Совета Вычислительного центра Академии наук СССР (Москва, В-333, ул. Вавилова, 40, конференц-зал).

С диссертацией можно ознакомиться в библиотеке института.

Ученый секретарь Совета

Во многих областях естествознания важное место занимают задачи классификации (распознавания) тех или иных явлений.

В самых общих чертах задачу распознавания можно сформулировать следующим образом. Пусть известно, что рассматриваемый объект может находиться в одном из конечного числа состояний. Требуется по некоторой информации об этом объекте определить, в каком состоянии находится объект.

К задачам распознавания относятся, например, задача об оценке масштабов оруденения в геологии, диагностика заболеваний в медицине, определение экономической перспективности районов в экономике, задача контроля правильности работы устройства в технике, прогнозирование свойств химических соединений в химии, задача построения автоматов, умеющих различать звуковые сигналы, рукописные буквы, геометрические образы и т.д. Ясно, что список подобных задач трудно исчерпать.

К настоящему времени решение задач распознавания накопило большой практический и теоретический опыт. Вообще говоря, проблема распознавания включает в себя ряд проблем, которые можно рассматривать и как важные самостоятельные задачи. К ним относятся, например, задача выбора системы признаков, с помощью которых описывается рассматриваемый объект, задача классификации с эталонами (учителем), задача самопроизвольной классификации, задача вычисления меры важности признака или группы признаков и др.

Исторически теория и практика решения задач распознавания развивалась по различным направлениям, и в настоящее время существует много используемых на практике алгоритмов, основанных на различных подходах к решению задач распознавания.

Рассмотрим, например, как решается основная задача теории распознавания: задача классификации с эталонами – с помощью некоторых из известных алгоритмов.

Алгоритмы, реализующие Байесово правило минимизации среднего риска, предполагают для каждого класса объективное существование функций распределения вероятностей распознаваемых ситуаций и сводятся к восстановлению этих функций [8].

Алгоритмы, основанные на методе потенциальных функций, реализуют рекуррентные процедуры построения разделяющих гиперповерхностей в пространстве признаков. Теория этого метода близка теории стохастической аппроксимации [10].

Алгоритмы метода обобщенного портрета сводятся к построению специальных разделяющих гиперплоскостей, обладающих некоторыми экстремальными свойствами [8].

Ряд алгоритмов (например, "Кора-3", "Арифметика") реализуют идею преобразования пространства признаков с тем, чтобы в новом признаковом пространстве происходило более четкое разделение рассматриваемых ситуаций на классы [9].

В предлагаемой работе подробно рассматривается еще один подход к решению задач распознавания, в котором используются идеи и методы различных разделов дискретной математики. Будем называть этот подход комбинаторно-логическим. К алгоритмам комбинаторно-логического подхода относятся тестовые алгоритмы распознавания и класс алгоритмов, основанных на вычислении оценок. В основе тестовых алгоритмов лежит понятие тупикового теста [1], играющего роль некоторой единицы информации, по которой полностью различимы классы материала обучения. В алгоритмах типа вычисления оценок ту роль единичной информации, по которой выносится суждение о различимости классов, выполняет не обязательно тест, а любая выборка из множества признаков ("опорное множество") [2, 3, 4].

Уже из столь краткого перечня существующих типов задач и алгоритмов следует, что большое практическое значение имело бы решение следующей проблемы: для заданного класса задач распознавания выделить класс алгоритмов, наиболее успешно решающих эти задачи. Но число существующих типов задач и алгоритмов слишком велико, чтобы эта проблема была решена сразу и в полной мере.

Исследование некоторых алгоритмов по мощности, по качеству на определенном круге задач было бы шагом на пути к решению этой проблемы.

В этой работе рассматриваются три алгоритма распознавания, основанные на комбинаторно-логическом подходе: алгоритм нахождения информационных весов эталонов, алгоритм голосования по тупиковым тестам, алгоритм голосования по выборкам длины K. Эти алгоритмы были разработаны в ВЦ АН СССР и в течении ряда лет

использовались при решении различных практических задач по распознаванию (см., например, [4, 6]). Это дало возможность исследовать эти алгоритмы с точки зрения надежности распознавания, затрат машинного времени и определения практически максимальных значений параметров, характеризующих размеры обрабатываемых таблиц.

В данной работе приводится решение при помощи этих алгоритмов одной из актуальных задач в геологии - задачи об оценке масштабов оруденения месторождений ценного минерального сырья на примере трех групп месторождений: ртутных месторождений Северо-Востока СССР, месторождений бокситов Тургайского прогиба и месторождений флюорита Восточного Забайкалья. На задачах этого же типа проводится сравнительная оценка некоторых параметров, характеризующих работу алгоритмов.

Перейдем теперь к более подробному изложению результатов. Сформулируем две задачи, решаемые здесь: задачу классификации с эталонами и задачу нахождения меры важности признака.

Пусть известно, что объект A может находиться в одном из S состояний. Пусть задана матрица

$$T_{n,m,s} = \| x_{ij} \|, \quad (1 \le i \le m, \quad 1 \le j \le n)$$

Каждая строка матрицы, называемая эталоном, описывает какое-либо состояние объекта A в системе признаков $\{ a_1, a_2 \ldots a_n \}$: x_{ij} есть значение признака a_j на i-ом эталоне. Вообще говоря, различные строки могут описывать одно и то же состояние.

Все эталоны поделены на S не пересекающихся групп. В l-ую группу $(1 \le l \le S)$ входят те и только те эталоны, которые описывают l-ое состояние объекта. Будем считать, что заданная система признаков такова, что любые два набора, являющиеся описанием различных состояний объекта, различны.

Требуется: I) Найти такое решающее правило (алгоритм), которое позволяло бы по любому набору, описывающему какое-либо состояние объекта и не обязательно принадлежащему множеству эталонов, определить состояние объекта; 2) каждому признаку a_j $(1 \le j \le n)$ поставить в соответствие число, являющееся мерой важности этого признака при определении состояния объекта.

Рассмотрим тестовый подход к решению вышесформулированных задач.

Введем ряд определений.

Тестом матрицы $T_{n,m,s}$ называется такое подмножество ее столбцов, что любые две строки матрицы, образованной этими столбцами, различны, если они принадлежат различным группам.

Тупиковым тестом матрицы $T_{n,m,s}$ называется тест, любое собственное подмножество столбцов которого не является тестом.

Информационным весом признака a_j называется величина $p_j = \frac{\tau_j}{\mathcal{T}}$, где $\mathcal{T}$ - число всех тупиковых тестов матрицы $T_{n,m,s}$, τ_j - число всех тупиковых тестов, в которые входит j-ый столбец матрицы [3]

Вообще говоря, могут представиться два случая.

В первом случае матрица $T_{n,m,s}$ является полным описанием объекта A, т.е. не существует описаний состояний объекта, отличных от эталонных. В этом случае достаточно найти любой тупиковый тест матрицы $T_{n,m,s}$, т.к. знание значений признаков, образующих тест, позвиляет однозначно определить состояние объекта.

На практике обычно встречается другой тип задачи: матрица $T_{n,m,s}$ содержит не все возможные описания состояний объекта A. Существующие тестовые алгоритмы распознавания используют в этом случае для определения состояния объекта все множество тупиковых тестов матрицы $T_{n,m,s}$.

Отметим два свойства, характеризующих такой подход к решению задачи. Во-первых, построение всего множества тупиковых тестов или даже некоторой части его гарантирует построение решающего правила, по которому правильно распознается эталонный материал. Во-вторых, алгориты голосования по тупиковым тестам, например, гарантирует правильное распознавание состояний объекта по набору, не вошедшему в эталонную матрицу, при условии, что множества тупиковых тестов матрицы $T_{n,m,s}$ и матрицы, полностью описывающей объект A, в большей своей части совпадают.

Меру важности признака a_j для различения состояний объекта в тестовых алгоритмах определяет величина $P_j = \frac{\tau_j}{\tau}$.

Такое решение можно обосновать с помощью следующих рассуждений. Предположим, что некоторый признак a_j входит в наибольшее число тупиковых тестов. Тогда удаление этого признака из матрицы приводит к наибольшей потере информации, позволяющей различать состояния объекта. Отсюда следует, что чем больше величина P_j, тем большую роль играет признак a_j в распознавании состояний.

Отметим, что специальной задачей является задача разработки алгоритмов построения всех тупиковых тестов. Существующие оценки числа тупиковых тестов [7] говорят о том, что это число является сильно растущей функцией параметров, характеризующих размеры матрицы. В данной работе приводится алгоритм нахождения всех тупиковых тестов. Этот алгоритм, кроме того, что использовался при решении различных задач по распознаванию, вошел как составная часть в существующий стохастический алгоритм нахождения тупиковых тестов [6].

Тестовый подход позволяет решать еще одну задачу, связанную с предыдущей, а именно, задачу о взаимозависимости признаков. Обозначим через τ_{ij} - число тупиковых тестов, в которые одновременно входят признаки a_i и a_j. Тогда, очевидно, меру зависимости признака a_i от признака a_j будет характеризовать величина $\chi_{i(j)} = \frac{\tau_{ij}}{\tau}$.

Рассмотрим алгоритмы распознавания, основанные на методе вычисления оценок. Теория этого метода по существу дает способ построения довольно широкого класса алгоритмов, решающих раз-

личные задачи теории распознавания. Каждый алгоритм определяется заданием следующих этапов: выбором системы опорных множеств, выбором функции "близости" частей эталонов, высекаемых опорным множеством, способом определения числа "голосов", подаваемых набором за класс по системе опорных множеств, выбором решающего правила. Решение задачи оптимизации по параметрам, характеризующим вышеперечисленные этапы, позволяет строить алгоритм в некотором смысле наилучшим образом решающий рассматриваемую задачу распознавания [2].

Перейдем теперь к описанию трех конкретных алгоритмов распознавания, исследуемых в этой работе. В дальнейшем через $\tilde{x} = (x_1, x_2, \ldots, x_n)$ будем обозначать описание какого-либо состояния объекта в системе признаков $\{a_1, a_2, \ldots, a_n\}$: x_j есть значение признака a_j.

Алгоритм нахождения информационных весов эталонов

заключается в построении некоторой линейной функции $y(\tilde{x})$, при этом используются предварительно найденные веса признаков. Рассматривалось два способа построения функции: либо $y(\tilde{x}) = \sum_{j=1}^{n} p_j \cdot x_j$, где p_j - информационный вес признака, либо $y(\tilde{x}) = \sum_{j=1}^{n} p_j(x_j)$, где $p_j(x_j)$ - разделяющий вес признака a_j. Процесс обучения заключается в определении некоторого ряда чисел $C_0, C_1, \ldots, C_s$. Решающее правило состоит в следующем: $\tilde{x}$ описывает ℓ-ое состояние, если $C_{\ell-1} < y(\tilde{x}) < C_\ell$, решение не принимается, если $y(\tilde{x}) = C_\ell$ $(\ell = 1, 2, \ldots, s)$.

Алгоритм голосования по тупиковым тестам

заключается в построении системы функций:

$$y_\ell(\tilde{x}) = \sum_{\{\mathfrak{I}\}} \sum_{i=m_1(\ell)}^{m_2(\ell)} x_{j_1}^{x_{ij_1}} \ldots x_{j_k}^{x_{ij_k}} \qquad (\ell = 1, 2, \ldots, s)$$

где $\{\mathfrak{I}\}$ - множество всех тупиковых тестов матрицы $T_{n,m,s}$, $\{a_{j_1} \ldots a_{j_k}\} \in \{\mathfrak{I}\}$, $m_2(\ell) - m_1(\ell)$ - мощность множества эталонов, описывающих ℓ-ое состояние объекта,

$$x_{j_t}^{x_{ij_t}} = \begin{cases} x_{j_t}, & \text{если } x_{ij_t} = 1 \\ 1 - x_{j_t}, & \text{если } x_{ij_t} = 0 \end{cases}$$

Решающее правило состоит в следующем: $\tilde{x}$ описывает ℓ-ое состояние, если $y_\ell(\tilde{x}) = \max_{1 \le i \le s} \{y_i(\tilde{x})\}$, решение не принимается, если $y_\ell(\tilde{x}) = y_i(\tilde{x})$ при $i \neq \ell$.

Алгоритм голосования по выборкам длины K

состоит в построении системы функций:

$$y_\ell(\tilde{x}) = \frac{1}{m(\ell)} \sum_{i=m_1(\ell)}^{m_1(\ell)} \sum_{t=0}^{\varepsilon} C_{n-\rho(\tilde{x}_i,\tilde{x})}^{K-t} \cdot C_{\rho(\tilde{x}_i,\tilde{x})}^{t} \qquad (\ell = 1, 2, \ldots, s),$$

где $\rho(\tilde{x}_i, \tilde{x})$ - расстояние Хемминга между двумя векторами, ε и K - параметры, определяемые в процессе обучения. Решающее правило такое же, как и в предыдущем алгоритме.

В данной работе приводятся примеры применения этих алгоритмов для решения задачи об оценке масштабов оруденения месторождений. Рассматривалось три группы месторождений: группа ртутных месторождений Северо-Востока СССР, группа месторождений бокситов Тургайского прогиба, группа месторождений флюорита Восточного Забайкалья.

Для всех трех групп месторождений был представлен хорошо изученный эталонный материал по двум классам месторождений: крупных, имеющих промышленное значение, и мелких, разработка которых является не рентабельной. Требовалось для группы контрольных месторождений определить, к какому классу они относятся.

В случае ртутных месторождений задача решалась методом голосования по тупиковым тестам. Матрица содержала восемь эталонов, по четыре эталона для описания каждого класса. При описании месторождений было использовано двадцать восемь бинарных признаков. Оценивался масштаб оруденения десяти контрольных месторождений, для восьми из них вывод совпал с оценкой специалистов.

Для группы месторождений бокситов Тургайского прогиба задача решалась методом информационных весов эталонов. Месторождения были описаны в системе тридцати бинарных признаков. Эталонная матрица содержала четырнадцать эталонов: по семи эталонов в каждом классе. Контроль проводился для двенадцати месторождений. В десяти случаях оценка масштаба оруденения соответствовала оценке, даваемой специалистами.

Оценка масштаба оруденения месторождений флюорита проводилась методом голосования по выборкам длины k. В эталонную матрицу входило восемь эталонов: четыре эталона, описывающих крупные месторождения, и четыре эталона, описывающих мелкие месторождения. При описании использовались сорок бинарных признаков. Десять месторождений из двенадцати контрольных были отнесены к своему классу.

Задача о нахождении меры важности признака в этой работе решалась в связи с задачей сокращения исходного признакового пространства. Как будет видно из приведенных примеров достаточно сильное сокращение исходного признакового пространства не привело к ухудшению качества распознавания. В работе приводится решение задачи распознавания методом голосования по тупиковым тестам в исходном признаковом пространстве и в сокращенном (оставлялись признаки с высоким информационным весом). Для всех трех групп месторождений качество распознавания не ухудшилось. Видимо, этот факт говорит об избыточности заданной информации и о возможности выбора наиболее существенной части ее.

Исследование параметров, характеризующих работу алгоритмов, проводилось на аналогичных задачах. В работе приводятся результаты решения задачи об оценке масштабов оруденения месторождений по материалам четырнадцати групп месторождений ценного минерального сырья. Для каждой группы месторождений задача была решена тремя алгоритмами. В работе приведены параметры, харак-

теризующие размеры эталонного материала и материала контроля. По каждому из алгоритмов даются зависимости времени решения задачи от двух параметров – числа признаков и числа эталонов в матрице $T_{n,m,s}$. Проводится сравнительная оценка надежности распознавания по этим алгоритмам и исследуются практические возможности алгоритмов.

Как показали вычисления наиболее надежными с точки зрения качества распознавания оказались алгоритм голосования по тупиковым тестам и алгоритм голосования по выборкам длины K.

В заключении отметим, что рассмотренные алгоритмы использовались при решении различных практических задач по распознаванию. В частности, было решено много задач из области геологии; решение некоторых из них имело практическое значение. Так, исследование ртутных месторождений позволило дать практические рекомендации о необходимости постановки дополнительных геолого-разведочных работ на некоторых мало изученных месторождениях Северо-Востока СССР [5]; по материалам Приморского геологического управления о месторождениях олова Приморья были получены высокие оценки масштабов оруденения некоторых мало изученных месторождений, что в дальнейшем подтвердилось при проверке разведочными работами на двух месторождениях: Южном и Смирновском – было установлено, что рудные тела, сравнительно бедные оловом на поверхности, на глубине переходят в богатые руды с высоким содержанием оловянного камня.

Работа состоит из введения, трех глав и приложения.

В первой главе дается постановка двух задач теории распознавания: задачи классификации с эталонами и задачи нахождения меры важности признака. Рассматриваются алгоритмы комбинаторно-логического подхода к решению этих задач: тестовые алгоритмы и класс алгоритмов, основанных на вычислении оценок. Дается описание трех конкретных алгоритмов распознавания. Подробно описывается алгоритм нахождения всех тупиковых тестов матрицы $T_{n,m,s}$.

Во второй главе формулируется задача об оценке масштабов оруденения месторождений. Дается описание трех групп месторождений ценного минерального сырья. Для каждой из этих групп приводится решение двух задач: задачи об оценке масштабов оруденения месторождений и задачи о нахождении меры важности признака.

В третьей главе исследуются практические возможности вышерассмотренных алгоритмов распознавания. Приводятся зависимости времени решения задачи от параметров, характеризующих размеры эталонной матрицы, а также делается сравнительная оценка качества алгоритмов, с точки зрения надежности распознавания.

В приложении приводится реализация в кодах БЭСМ-6 алгоритма нахождения всех тупиковых тестов матрицы $T_{n,m,s}$.

Автор глубоко благодарен Валерию Борисовичу КУДРЯВЦЕВУ, под чьим руководством была выполнена настоящая работа.

Автор также пользуется случаем поблагодарить Сергея Всеволодовича Яблонского, Рэма Михайловича Константинова и Юрия Ивановича Журавлева за помощь и поддержку в работе.

Основные результаты докладывались на III Всесоюзной конференции по проблемам теоретической кибернетики и опубликованы в работах:

С.В.Яблонский, Н.Г.Демидова, Р.М.Константинов, З.Е.Королёва, В.Б.Кудрявцев, С.В.Сиротинская. Тестовый подход к количественной оценке геолого-структурных факторов и масштабов оруденения (на примере ртутных месторождений). Геология рудных месторождений, 1971, 13, № 2, 30-42.

Р.М.Константинов, З.Е.Королёва. Применение тестовых алгоритмов к задачам геологического прогнозирования. Тр.Международного симпозиума по практическим применениям методов распознавания образов 1971г., М., ВЦ АН СССР, 1973г., 194-204.

З.Е.Королёва. О некоторых характеристиках тестовых алгоритмов распознавания. Тр.III Всесоюзной конференции по проблемам теоретической кибернетики. Матем. ин-т СО АН СССР, 1974.

ЛИТЕРАТУРА

1. Чегис И.А., Яблонский С.В. Логические способы контроля электрических схем. Тр. Матем. ин-та АН СССР, 1958, 51.
2. Журавлев Ю.И., Никифоров В.В. Алгоритмы распознавания, основанные на вычислении оценок. Кибернетика, 1971, № 3.
3. Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П. О математических принципах классификации предметов и явлений. В сб. "Дискретный анализ", № 7, Новосибирск, "Наука" СО АН СССР, 1966.
4. Кудрявцев В.Б., Кудрявцев В.Б. О тестовом подходе к задаче о перспективности населённых пунктов. В сб."Исследование операций", М., ВЦ АН СССР, 1972.
5. Константинов Р.М. Прогноз руд и кибернетические модели. Природа, № 11, М., "Наука" АН СССР, 1974.
6. Кузнецов В.Е. Об одном стохастическом алгоритме обработки больших таблиц по методу тестов. В сб."Дискретный анализ", № 24, Новосибирск, "Наука" СО АН СССР, 1973.
7. Слепян В.А. Вероятностные характеристики распределения тупиковых тестов. В сб."Дискретный анализ", № 12, Новосибирск, "Наука" СО АН СССР, 1968.
8. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М., "Наука" АН СССР, 1974.
9. Бонгард М.М. Проблема узнавания. М., "Наука" СО АН СССР, 1967.
10. Айзерман М.А., Браверман Э.М., Розоноэр А.И. Метод потенциальных функций в теории обучения машин. М., "Наука" АН СССР, 1970.

З. Е. Королева

Анализ эффективности некоторых
комбинаторно-логических алгоритмов
распознавания (на примере решения
геологических задач)

Т-01083. Подписано в печать 3/11-75 г. Зак. 10. Тир. 200
Бесплатно

Отпечатано на ротапринтах в ВЦ АН СССР
Москва, В-333, ул. Вавилова, 40