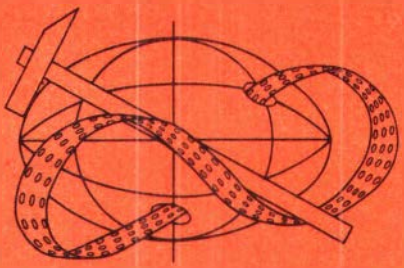
АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
ИНСТИТУТ ГЕОЛОГИИ И ГЕОФИЗИКИ

МЕТОД СОГЛАСОВАННЫХ ОЦЕНОК



НОВОСИБИРСК – 1982

АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
ИНСТИТУТ ГЕОЛОГИИ И ГЕОФИЗИКИ

МЕТОД СОГЛАСОВАННЫХ ОЦЕНОК

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ

НОВОСИБИРСК—1982

УДК 51.553.062.98

Метод согласованных оценок: Метод.рекомендации. /Составители: А.Н.Дмитриев, С.В.Макаров, Е.А. Смертин и др. -Новосибирск: Издание ИГиГ СО АН СССР. - 133 с.

Изложены методические указания для практического применения метода согласованных оценок. Приведены рекомендации по употреблению основных процедур метода в геологических задачах прогнозно-поискового профиля. На ряде практических задач показана общая схема решения и эффективность метода. Даны приемы приложения метода к задачам распознавания, упорядочения объектов, восстановления пропусков в таблицах данных, приведен вариант программ.

Методические рекомендации адресованы геологам, применяющим математические методы и ЭВМ для решения задач прогнозирования и поиска полезных ископаемых, а также математикам, работающим в сфере обработки геолого-геофизической информации.

Составители:

А.Н.Дмитриев, С.В.Макаров, Е.А.Смертин,
А.С.Вакуленко, В.Н.Кандыба, В.Д.Карбышев, Т.И.Штатнова

Материал рекомендован к печати Секцией стратиграфии, тектоники, литологии и осадочных полезных ископаемых ученого совета института геологии и геофизики СО АН СССР в качестве методических рекомендаций.

В В Е Д Е Н И Е

В настоящее время фактор точности в геологии приобретает все большее значение. Ориентация на точность исследовательских процедур поощряет внедрение математических идей и методов. Особое значение при этом имеет процесс математизации операций элементарных актов распознавания, из которых и складывается основной вид сравнительного изучения в геологии. Процедуры обработки информации в геологических задачах прогнозно-поискового профиля нами связываются с проблемой детального исследования принципов и правил сравнительного изучения объектов.

Этот класс задач особенно важен в нефтегеологии и рудопрогнозе в связи с проблемой разбраковки перспективных и неперспективных площадей и районов до их ввода в бурение. Информационная обстановка в сфере указанной проблемы позволила осуществить ряд интересных математических разработок и получить практические результаты при решении конкретных задач.

Наряду с известным тестовым подходом, метод согласованных оценок (МСО) явился основой для разработки целого ряда методик, процедуры которых базируются на итерационных приемах.

В проведенных исследованиях помимо разработки математических методов значительное внимание было уделено этапу сбора, подготовки и предварительной обработки геолого-геофизической информации. Как показал опыт, практических применений метода согласованных оценок, достоверность получаемых результатов решения и эффективность их истолкования зависят от тщательности мобилизации исходных данных и приведения их к виду, уместному для обработки на ЭВМ.

Созданный метод согласованных оценок имеет отношение к следующим вопросам: выяснению логической сцепленности и типов содержательно заданных признаков; формированию признаковых пространств и выяснению количественных оценок общей схемы родства характеристических признаков; упорядочению и сортировке совокупности исследуемых объектов согласно общему целеуказанию. Особое внимание уделено способам оценки нагрузок (существенности) при-

знаков и объектов для ранжирования и распознавания.

Излагаемый материал подразделен на параграфы. В первом и втором параграфах (Дмитриев А.Н.) даны сведения о возникновении МСО, его общие характеристики и основные математические процедуры. Рассмотрение вопроса связи ранжирования объектов исследования с их целевой упорядоченностью приводится в третьем параграфе (Макаров С.В.). В четвертом параграфе (Смертин Е.А.) подробно описан центрированный вариант метода. Согласованные оценки для неоднородных выборок освещены в пятом параграфе (Макаров С.В.). Шестой параграф посвящен вопросу вычисления весовых коэффициентов и согласованных оценок (Макаров С.В.). Алгоритмы и их программные реализации включены в седьмой параграф (Смертин Е.А., Вакуленко А.С.). Вопросу взаимосвязей МСО с родственными ему методами, а также с тестовым подходом посвящен восьмой параграф (Смертин Е.А., Дмитриев А.Н.). Примеры конкретных решений задач прогнозно-поискового профиля (Дмитриев А.Н., Штатнова Т.И., Кандыба В.Н., Карбышев В.Д., Смертин Е.А.) даны в девятом параграфе. Работа заключается десятым параграфом (Макаров С.В.), где дан краткий обзор метода главных компонент. Основа программного обеспечения метода осуществлена программистами: Васильевой Е.Н., Кандыбой В.Н., Вакуленко А.С. Конкретное решение многочисленных задач и трудоемкие работы оформительского характера проводились Штатновой Т.И. при содействии Шишкиной Л.Н.

Составители пользуются случаем поблагодарить геологов ряда экспедиций, а также сотрудников Института геологии и геофизики, Института математики и Вычислительного центра СО АН СССР за оказанную помощь в работе, особенно на заключительных этапах создания и апробирования метода согласованных оценок.

§I ОБЩИЕ СВЕДЕНИЯ О МЕТОДЕ

Разнообразие геологических сведений количественной и логической природы, а также целей обработки геологических данных, объем которых непрерывно возрастает, составляют специфику "информационной среды" геологии. Следует также иметь в виду и то, что информация геолого-геофизического характера весьма неравноценна по достоверности. Именно для такой сложноорганизованной информационной среды и предназначается метод согласованных оценок (МСО).

I. Возникновение и предназначение метода

Название метода отражает особенность основной вычислительной процедуры, при которой оценки строк (столбцов) выражаются через аналогичные оценки столбцов (строк). Согласованные оценки строк и столбцов представляют собой, с математической точки зрения, неподвижную точку некоторого оператора, для нахождения которой и предназначена основная вычислительная процедура. Отправным пунктом для развития метода и его названия послужила работа [9], в которой был изложен математический результат Ю.Л.Васильева, названный "качельной процедурой". Этот вариант метода явился основополагающим для возникновения новых вариантов, дополнений, обоснований и практических приложений.

Оценка столбцов (признаков) и строк (объектов) в таблицах данных производится в предположении, что исследуемые данные представлены в виде прямоугольной таблицы X "объект-признак" и представляют собой характеристики исследуемых объектов. Причем каждая строка в таблице X представляет собой набор значений, характеризующих некоторый объект. Каждый столбец соответствует некоторому признаку. Таким образом, элемент x_{ik} представляет собой значение k-го признака на i-м объекте.

Основополагающими идеями в поиске метода были разработки по теории тестов [46]. В последующем был проведен сравнительный анализ метода согласованных оценок и метода главных компо-

нент (МГК) [30]. Попытка углубить содержательную трактовку согласованности и выяснить природу оценок привела к дополнительным разработкам [29], а также к построению "центрированной качельной процедуры", по которой оценки отклонений выраженности признаков на объектах производятся от средних величин, что в статистике соответствует дисперсии [28]. По мере развития практических приложений как основного, так и центрированного вариантов метода, возникла необходимость в выявлении взаимосвязей МСО с родственными ему вариантами итерационного подхода. Кроме того, при сопоставлении МСО с хорошо изученным тестовым подходом [16, 17, 19] обнаружилась их взаимная дополнительность и то, что в зависимости от структуры массива данных один из подходов оказывается предпочтительнее. Выяснилось также, что оценки объектов, вычисленные по МСО, нередко коррелируют с проявленностью (на тех же объектах) важных в практическом отношении свойств [10-12, 14, 20, 40, 42].

Метод предназначен для решения задач диагностики и классификации, для отбора наиболее информативных признаков, для выделения подмножества представительных объектов, для предсказания значений выделенного (целевого) признака, а также для сжатия информации, в соответствии с конкретной (содержательной) постановкой задачи. По соответствующим алгоритмам были отлажены реализующие их программы [15, 22-24, 33, 34].

2. Особенности метода

Метод согласованных оценок, возникший из практических задач математической обработки геологических данных, с первых шагов подвергался разносторонней критике. В частности, указывалось на очевидную формальную близость МСО к хорошо изученному методу главных компонент (МГК). Поэтому прежде всего было необходимо выяснить, насколько глубока и содержательна связь МСО и МГК и в чем их существенное различие. Последнее установить нетрудно: МГК (в классическом варианте) работает с центрированными данными (так, что среднее значение каждого признака равно 0), а МСО имеет дело с нецентрированными данными, отсчитанными от фиксированных, физически обоснованных нуль-пунктов. Поэтому ни один из этих методов не является частным случаем другого. Наглядные

интерпретации МГК и МСО тоже разные. Геометрический смысл МГК (проектирование на направление "наибольшей вытянутости" выборки) описан во многих руководствах (см.например[3I]). Что же касается МСО, то с ним связаны качественные толкования совсем иного рода [9,I0]. В МСО в явном виде использован принцип согласования весов строк и весов столбцов таблицы между собой. Поэтому в первом варианте МСО данные предполагались неотрицательными, чтобы гарантировать неотрицательность коэффициентов - весов , находимых по МСО. Для выявления глубокой связи между МСО и МГК необходимо с самого начала отказаться от центрирования исходных данных (и, следовательно, от проекционной интерпретации), и вывести МГК из других предпосылок. Такая возможность представляется теоремами Экарта-Янга о сингулярном разложении, а также теоремой Рао-Дарроча [48,50,52]. Нецентрированный вариант МГК, основанный на сингулярном разложении таблицы данных, представляет собой очень содержательную математическую модель, из которой, как частные случаи, следуют МГК (в классическом, центрированном варианте), а также и улучшенный вариант МСО.

Соответствующие соотношения, включающие процедурную часть МСО, могут рассматриваться и как модель получения оценок,и как содержательно интерпретируемая модель. В последнем случае анализ эмпирического материала состоит в применении процедур к массиву исходных данных, а также к его специфическим частям, выявившимся при анализе. Результатом применения метода является преобразование и перестройка массива данных, в частности,исходный массив может оказаться разделенным на несколько классов. Этот анализ является многоэтапным, причем выбор последовательности этапов зачастую нетривиален и зависит как от исходного материала, так и от целей обработки. Зачастую цель обработки материала тесно увязана с целевым признаком, который лишь в редких случаях включается в характеристическую совокупность признаков. Как правило,целевой признак не включается в процедуру обработки данных, а масштабирует ценность исследуемых объектов (например, масштаб запасов полезных ископаемых в данной совокупности месторождений). В совокупности же характеристических признаков целевой признак присутствует "рассеянно", в неявном виде, а цель обработки может состоять в том, чтобы по косвенным (по отношению

к цели) характеристическим признакам для ряда объектов (проб) установить значения целевого признака. Из совокупности характеристических признаков выделяются - приближенные по значению к целевому признаку - поисковые. Объекты, упорядоченные по значениям поисковых признаков, часто согласуются с упорядоченностью по целевому признаку [10, 14, 40].

Оценки объектов и признаков используются как непосредственно, так и для решения задач распознавания, сортировки объектов по заданным классам.

Метод в целом характерен тем что:

а) позволяет устанавливать некоторое естественное соотношение между исследуемой совокупностью объектов сравнения;

б) отличается малой трудоемкостью;

в) допускает задачи с большим числом бинарных и многозначных признаков;

г) позволяет устанавливать числовую меру для столбцов и строк таблиц решения;

д) требует ряд структурных нежестких ограничений на характер таблиц;

е) находится на достаточно строгом уровне математического обоснования в сравнении с другими методами подобного предназначения.

Следует отметить и тот факт, что МСО все еще находится в состоянии дальнейшего обоснования и разработки. Его дальнейшие точки роста очевидны и в этом авторы усматривают нетривиальную перспективу метода.

3. Практические применения

Как уже отмечалось, само возникновение метода обязано практическому запросу, а именно: необходимости обрабатывать большие форматные таблицы, значения признаков которых имеют разнообразную природу (бинарные, шкала наименований, количественные и др.). Кроме того, в большом числе случаев у геологов возникает необходимость в получении прикидочных результатов решения задач в краткие сроки. Причем часто эти прикидки являются вполне удов-

летворительными для реализации поставленной цели. Именно этот метод и послужил основой для организации широкой обработки геолого-геофизической информации по большому перечню целей.

Применение решающих программ по МСО относится к сфере задач прогнозно-поискового профиля. Конкретные задачи производственного характера принимались к решению по двум направлениям:

а) задачи по нефте- и газопрогнозу [10,20,39,40,42];

б) задачи по рудопрогнозу [11,14,32].

Практическому решению задачи предшествует геологическая постановка задачи с последующим ее преобразованием в формализованную [18,41]. Как правило, решение задач-поэтапное, многошаговое с обязательным уточнением и коррекцией исходной информации. Это препарирование исходных данных осуществляется в содружестве с заказчиком, геологом-постановщиком. Результаты выдаются в строгом соответствии с поставленными целями обработки информации и характером требований геологов. Оценка результатов решения, как правило, производится на соответствующих заседаниях научно-технических советов.

Перечень задач, решенных МСО, достаточно обширен. В соответствии с увеличением алгоритмических и программных возможностей круг задач, подлежащих решению, возрастает. В параграфе о практических приложениях будут приведены конкретные примеры решения задач с помощью МСО.

§2 ОПИСАНИЕ МЕТОДА СОГЛАСОВАННЫХ ОЦЕНОК

Данный параграф посвящен изложению общей характеристики метода. Изложение ведется в строгом соответствии той последовательности разработок метода, которая имела место как в теоретическом, так и в практическом отношениях. Прочтение этого параграфа, по замыслу авторов, должно ознакомить читателя не только с общими содержательными и математическими свойствами метода согласованных оценок, но и с рядом подробностей, неизбежно возникающих при данном изложении материала.

Изложение начинается с описания вычислительных процедур, объединенных в некоторую математическую модель. Вслед за освещением основного приема МСО в последующем разделе параграфа приводится краткая алгоритмическая справка о дополнении по центрированным качельным процедурам метода. Далее характеризуется часть метода, которая прямо относится к решению практических задач в сфере упорядочения исследуемых объектов и схем распознавания. Заканчивается параграф изложением представлений о возможности приложения метода (нагрузок объектов) для выяснения вопросов характеристичности и типичности исследуемых совокупностей объектов.

Параграф может иметь и самостоятельное значение, кроме того, в целях сохранения приемственности с первыми публикациями в нем сохранена терминологическая и аппаратная схемы.

I. Вычислительные процедуры

Метод согласованных оценок строк и столбцов, представляемый в рамках нижеизложенной математической модели, тяготеет к классическим подходам математической статистики. Однако, несмотря на эту "классичность", он, в ряде своих особенностей, имеет самостоятельное и обобщающее для статистических решений значение.

Пусть имеется m объектов $s_1, s_2, \ldots, s_m$, охарактеризованных двузначными признаками $x_1, x_2, \ldots, x_n$. Пусть $T = [t_{ij}]$ - таблица $m \times n$, которая составлена из единиц и нулей и в ко-

торой i-я строка (t_{i1}, t_{i2}, ... , t_{in}) отвечает объекту s_i, $i = 1,2, \dots, m$, j-й столбец (t_{1j}, t_{2j}, ... , t_{mj}) отвечает признаку x_j, $j = 1,2, \dots, n$; таблица T отражает выраженность признаков у объектов - если для i-го объекта и j-го признака она превышает некоторый уровень, то $t_{ij} = 1$, а если меньше, то $t_{ij} = 0$. Условимся, что $m \geq 2$, $n \geq 2$ и что в таблице T нет строк и нет столбцов, составленных сплошь из нулей.

Выраженность признаков у объектов может быть измерена и в шкале интервалов [32,36]. В этом случае признак изменяется на отрезке [0,1]. Пока ограничимся случаем, когда $t_{ij} \in \{0,1\}$. Пусть задана $T = [t_{ij}]$ - таблица размера $m \times n$, составленная из единиц и нулей. Дополнительно предположим, что таблица T является связной, т.е. перестановками строк и столбцов ее нельзя представить в виде (рис.1),

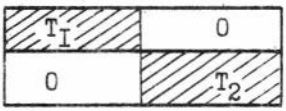


Рис.1.

где через "0" обозначены части таблицы T, заполненные сплошь нулями. Для реальных геологических таблиц, отвечающих значительной переплетенности влияний признаков, такое представление исключено, они заведомо связаны.

Определим для n столбцов таблицы T такие положительные числа π_1, π_2, ... , π_n и для m строк таблицы T такие положительные числа ω_1, ω_2, ... , ω_n, что аналогично тестовому подходу [16] , числа для строк получаются из чисел для столбцов по формуле

$$\lambda \omega_i = t_{i1}\pi_1 + t_{i2}\pi_2 + \dots + t_{in}\pi_n \quad \text{для} \quad i = 1,2, \dots, m, \quad (1)$$

где λ - некоторый множитель, один и тот же для всех строк таблицы T. Однако в отличие от тестового потребуем, чтобы имел место и обратный подход, т.е. чтобы числа столбцов получались

из чисел строк по формуле:

$\mu\pi_j = t_{1j}\omega_1 + t_{2j}\omega_2 + \ldots + t_{mj}\omega_n$ для $j = 1,2, \ldots, n$, (2)

где μ – некоторый множитель, один и тот же для всех столбцов таблицы T .

Так как на самом деле важны не сами числа $\pi_1, \pi_2, \ldots, \pi_n$, а соотношения между ними, то будем считать, что эти числа не превосходят I, и аналогичное предположение сделаем для чисел $\omega_1, \omega_2, \ldots, \omega_m$. Такого рода наборы чисел будем называть нормированными.

Оказывается, в указанных предположениях числа $\pi_1, \pi_2, \ldots, \pi_n$ и $\omega_1, \omega_2, \ldots, \omega_m$, удовлетворяющие соотношениям (I) и (2), для таблицы T единственны [9] и в предлагаемом подходе в качестве оценок объектов и признаков принимаются именно они. Числа $\pi_1, \pi_2, \ldots, \pi_n$ и $\omega_1, \omega_2, \ldots, \omega_m$ называются нагрузками строк и нагрузками столбцов таблицы T .

Итак, в качестве оценок для строк и столбцов таблицы T будут браться нормированный набор положительных чисел $\omega_1, \omega_2, \ldots, \omega_m$ и нормированный набор положительных чисел $\pi_1, \pi_2, \ldots, \pi_n$, удовлетворяющих соотношениям (I) и (2). Числа λ и μ играют при этом роль нормирующих множителей. Соотношения (I) и (2) еще не указывают как найти эти числа для конкретной таблицы. Ниже мы рассмотрим еще эту связь, ибо в ней заключен смысл введенных оценок. Пример. Для указанной таблицы размера 3 x 4 (рис.2) нагрузки строк равны $\omega_1 = \sqrt{2}/2$, $\omega_2 = 1$, $\omega_3 = \sqrt{2}/2$, а нагрузки столбцов равны $\pi_1 = \sqrt{2} - 1$, $\pi_2 = 1$, $\pi_3 = 1$, $\pi_4 = \sqrt{2} - 1$.

1	1	0	0	ω_1
0	1	1	0	ω_2
0	0	1	1	ω_3
π_1	π_2	π_3	π_4	

Рис.2

Нетрудно проверить, что числа:

$1\cdot\pi_1 + 1\cdot\pi_2 + 0\cdot\pi_3 + 0\cdot\pi_4$

$0\cdot\pi_1 + 1\cdot\pi_2 + 1\cdot\pi_3 + 0\cdot\pi_4$

$0\cdot\pi_1 + 0\cdot\pi_2 + 1\cdot\pi_3 + 1\cdot\pi_4$

пропорциональны числам $\omega_1, \omega_2, \omega_3$ (множителем служит $\lambda = 2$), а числа:

$$\begin{array}{cccc} 1\cdot\omega_1 & 1\cdot\omega_1 & 0\cdot\omega_1 & 0\cdot\omega_1 \\ + & + & + & + \\ 0\cdot\omega_2 & 1\cdot\omega_2 & 1\cdot\omega_2 & 0\cdot\omega_2 \\ + & + & + & + \\ 0\cdot\omega_3 & 0\cdot\omega_3 & 1\cdot\omega_3 & 1\cdot\omega_3 \end{array}$$

пропорциональны числам π_1, π_2, π_3, π_4. (Множителем служит $\mu = I + (\sqrt{2}/2)$. Таким образом, данные числа удовлетворяют указанным выше соотношениям.

Рассматриваемые взаимоотношения схематически отражены на рис.3.

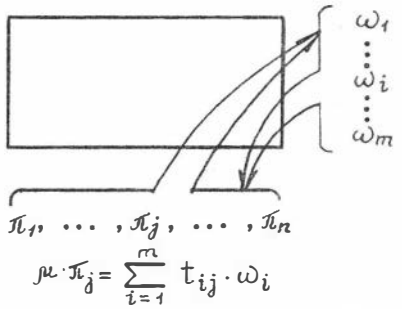


Рис.3

Отметим, что в соотношения (I) и (2) вовлечены все члены таблицы T и притом, в некотором смысле, "равноправно". Различия между величинами нагрузок образуются лишь за счет различий в самих членах таблицы T (некоторые члены оказываются нулями, другие – единицами, см. выше пример), точнее за счет различий в расположении единиц и нулей по строкам и столбцам таблицы.

Нетрудно заметить, что нагрузки отражают характер перекрытий строк и столбцов таблицы T – при одинаковом числе единиц в двух строках большую нагрузку будет иметь та строка, которая больше перекрывается по единицам с другими строками. Нагрузки выделяют типичные в этом смысле строки и столбцы таблицы T и потому в известном смысле отражают комбинаторные, или информационные свойства таблицы.

Соотношения $\lambda \vec{\omega} = T\vec{\pi}$, $\vec{\omega} > \vec{0}$; (I)

$\mu \vec{\pi} = T^*\vec{\omega}$, $\vec{\pi} > \vec{0}$ (2)

можно рассматривать поэтому как модель оценивания строк и столбцов таблицы T по содержащейся в них друг о друге информации.

Вместе с тем, эти соотношения можно рассматривать и как модель, отражающую связи между содержательной "значимостью" объ-

ектов и содержательной "ролью" в таблице признаков этих объектов. Разумеется, переход к такой трактовке должен сопровождаться максимумом точно оговариваемых предосторожностей. Необходимо, чтобы оценки были в этом случае содержательно-интерпретируемыми; желательно, чтобы были содержательно-интерпретируемыми также и группы объектов и группы признаков, на которые подразделяются объекты и признаки при упорядочении по величине нагрузок и т.п.

Как уже отмечалось, нагрузки $\omega_1, \omega_2, \ldots, \omega_m$ и $\pi_1, \pi_2, \ldots, \pi_n$ отражают довольно естественное соотношение между сравниваемыми объектами, и его следует иметь в виду как в комбинаторно-информационном, так и в содержательном плане.

Вычисление нагрузок подсказывается схемой на рис.3. Нетрудно обнаружить, что их можно получать последовательными приближениями (итерациями), опираясь на данную схему.

Начальные значения безразличны (лишь бы не сплошь нули), и можно считать, что нулевое приближение к нагрузкам есть нормированный набор чисел $\omega_1^{(0)} = 1, \ldots, \omega_m^{(0)} = 1$ и нормированный набор чисел $\pi_1^{(0)} = 1, \ldots, \pi_n^{(0)} = 1$.

Для отыскания первого приближения возьмем наборы чисел, которые получаются, если в выражения, стоящие в правых частях соотношений (1) и (2), подставить нулевое приближение, т.е. возьмем набор из m чисел, равных

$$\sum_{j=1}^{n} t_{1j} \cdot \pi_j^{(0)}, \ \ldots, \ \sum_{j=1}^{n} t_{mj} \cdot \pi_j^{(0)} \qquad (3)$$

(m сумм, отвечающих m строкам), набор из n чисел, равных

$$\sum_{i=1}^{m} t_{i1} \cdot \omega_i^{(0)}, \ \sum_{i=1}^{m} t_{i2} \cdot \omega_i^{(0)}, \ \ldots, \ \sum_{i=1}^{m} t_{in} \cdot \omega_i^{(0)} \qquad (4)$$

(n сумм, отвечающих n столбцам). Эти два набора чисел нормируем (т.е. в каждом из этих наборов разделим все числа на число, являющееся наибольшим в этом наборе), и полученные два набора чисел, которые обозначим через $\omega_1^{(1)}, \omega_2^{(1)}, \ldots, \omega_m^{(1)}$ и $\pi_1^{(1)}, \pi_2^{(1)}, \ldots, \pi_n^{(1)}$, возьмем в качестве первого приближения.

Для отыскания второго приближения будем поступать с первым приближением точно так же, как мы поступали с нулевым – подставлять в упомянутые выражения и затем нормировать полученные два набора чисел. Они будут вторым приближением и т.д.. Можно дока-

зать [9], что даже при сравнительно небольшом числе таких итераций получаются числа, весьма близкие к искомым нагрузкам ; эти числа можно поэтому взять в качестве нагрузок. Описанная процедура реализуется программой (Кандыба , 1977).

Пример 3. (Продолжение). Для таблицы размера 3 x 4 на нижеследующем рис.4 представлены результаты первых четырех приближений. Уже после четвертой итерации числа в наборе $\omega_1^{(4)} = 0{,}714$, $\omega_2^{(4)} = 1$, $\omega_3^{(4)} = 0{,}714$ довольно близки соответственно к числам $\omega_1 = \sqrt{2}/2$, $\omega_2 = 1$, $\omega_3 = \sqrt{2}/2$, а числа в наборе $\pi_1^{(4)} = 0{,}429$, $\pi_2^{(4)} = 1$, $\pi_3^{(4)} = 1$, $\pi_4^{(4)} = 0{,}429$ довольно близки соответственно к числам $\pi_1 = \sqrt{2} - 1$, $\pi_2 = 1$, $\pi_3 = 1$, $\pi_4 = \sqrt{2} - 1$. После 10-й итерации получается значение с точностью до 5-го знака.

Скорость сходимости итераций. Следует отметить быструю сходимость итераций для всех без исключения геологических таблиц (около 80), которые были изучены данным методом. Во всех случаях в получаемых на итерациях числах третий знак устанавливается между 5-й и 10-й итерациями, а к 30-й итерации устанавливались 8-й и 9-й знаки. Это обеспечивает весьма малую трудоемкость, так как отдельная итерация весьма проста, а число итераций невелико. Время счета для таблиц размера 10 x 40 около минуты на М-20 , М-220 и несколько секунд на БЭСМ-6.

Желательно не только эмпирическое подтверждение быстрой сходимости. В линейной алгебре известны примеры матриц [13,45], требующие даже при малой точности огромного числа итераций. Нужны доводы, что такого рода примеры среди геологических таблиц маловероятны или невозможны.

Скорость сходимости итераций тем больше, чем меньше величина $|\lambda_2 / \lambda_1|$, где λ_1 - максимальное по модулю собственное значение, а λ_2 - следующее за ним по величине собственное значение матрицы $\Phi = T \cdot T^*$, где T - исходная таблица размера $m \times n$, а T^* - таблица размера $n \times m$, в которой строками служат столбцы таблицы T. Известная теорема А.М.Островского [45]*) для положительной матрицы $T = [t_{ij}]$ ($t_{ij} > 0$ для всех i, j) дает оценку

*) См. там же несколько более сильные оценки Г.Биркгофа и Е.Хопфа.

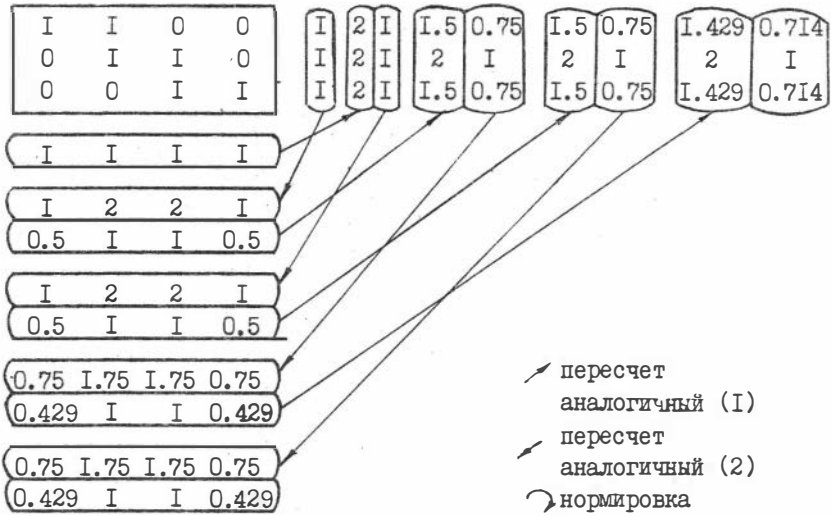


Рис. 4 . Итерационный процесс ("Качели").

$$\frac{|\lambda_2|}{|\lambda_1|} \leq \frac{(\max_{i,j}\{t_{ij}\})^2 - (\min_{i,j}\{t_{ij}\})^2}{(\max_{i,j}\{t_{ij}\})^2 + (\min_{i,j}\{t_{ij}\})^2},$$

где $|\lambda_1| > |\lambda_2|$.

В применении к таблице T это означает, что сходимость итераций для таблицы T будет весьма хорошей, если всевозможные скалярные произведения ее строк будут ненулевыми и будут по возможности мало отличаться друг от друга. Но эти свойства можно трактовать как математическое выражение имеющего место в геологических таблицах значительного переплетения, перекрытия строк по расположению в них единиц, т.е. значительного переплетения влияний признаков на фактор x_{n+1} . Тем самым получаем, что быстрота сходимости спецификой геологических таблиц в известной мере обеспечивается.

2. Центрированные качельные процедуры

Пусть дана таблица T из числовых элементов t_{ij}, стоящих на пересечении i-й строки – s_i и j-го столбца – x_j таблицы T, $i = 1,2, \ldots, m$, $j = 1,2, \ldots, n$; m – число строк таблицы T, n – число столбцов. Пусть t_{ij} – любые вещественные числа $-\infty < t_{ij} < \infty$. Строки таблицы T соответствуют некоторым объектам, а столбцы – признакам объектов. Итак, значение t_{ij} отражает выраженность признака x_j на объекте s_i (для удобства мы отождествляем обозначения строк s_i и объектов, а столбцов x_j – с признаками). Условимся, что $m \geq 2$, $n \geq 2$ и что в таблице T отсутствуют столбцы, заполненные полностью одним и тем же значением.*) Введем числовую меру для оценки строк и столбцов таблицы T.

Числовую меру для объектов можно задавать линейной, согласной значениям объектов и признаков, что приводит к получению весов, описанных выше.

Другой способ задания числовой меры – задать ее с помощью согласования значений весов с величинами отклонений значений элементов столбцов и строк от их средних значений. Тогда мы приходим к так называемым центрированным качельным процедурам [28], которые ведут учет не абсолютным значениям элементов таблицы, а их отклонениям от средних значений по строкам или по столбцам.

Предварительно таблица T преобразуется в таблицу U с помощью следующей нормировки ее столбцов:

$$u_{ij} = \frac{t_{ij} - t_{j\min}}{t_{j\max} - t_{j\min}} \qquad \begin{matrix} i = 1,2, \ldots, m; \\ j = 1,2, \ldots, n, \end{matrix}$$

где $t_{j\min}$ – минимальный, $t_{j\max}$ – максимальный элемент столбца x_j, откуда видно, что u_{ij} попадают в промежуток $0 \leq u_{ij} \leq 1$.

Пример преобразования таблицы T в U иллюстрируется на рис. 5.

*) Для бинарных таблиц нет необходимости в этом условии.

$$T = \begin{bmatrix} I & 0 & 0 & II \\ I & 0 & 7 & 7 \\ I & 2 & 9 & 9 \\ I & 2 & 5 & I \\ 0 & 2 & I0 & 2{,}5 \end{bmatrix} \longrightarrow U = \begin{bmatrix} I & 0 & 0 & I \\ I & 0 & 0{,}7 & 0{,}6 \\ I & I & 0{,}9 & 0{,}8 \\ I & I & 0{,}5 & 0 \\ 0 & I & I & 0{,}I5 \end{bmatrix}$$

Рис. 5 .

Веса строк и столбцов для таблицы T будем получать согласно таблице U .

Один вариант подсчета весов по таблице U (рис. 6)

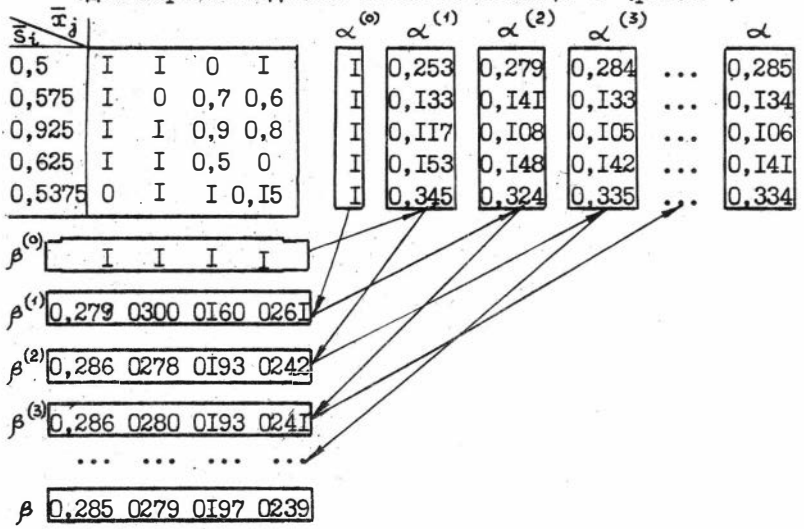


Рис. 6 .

ведется по следующим формулам:

$$\alpha_i^{(z+1)} = \frac{\sum_{j=1}^{n} \beta_j^{(z)} (u_{ij} - \bar{s}_i)^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} \beta_j^{(z)} (u_{ij} - \bar{s}_i)^2}, \qquad \begin{aligned} i &= I,2, \ldots, m ; \\ z &= 0,I,2, \ldots \\ j &= I,2, \ldots, n ; \end{aligned} \qquad (5)$$

$$\beta_j^{(z+1)} = \frac{\sum_{i=1}^{m} \alpha_i^{(z)} (u_{ij} - \bar{x}_j)^2}{\sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_i^{(z)} (u_{ij} - \bar{x}_j)^2} , \qquad (5)$$

где z – номер итерации, $\bar{s}_i$ – среднее арифметическое значение строки s_i, $\bar{x}_j$ – среднее арифметическое значение столбца x_j, причем значения $\alpha_i^{(0)} = \beta_j^{(0)} = I$, $i = I,2, \ldots, m$; $j = I, 2, \ldots, n$, а знаменатели дробей в формулах (I) подобраны так, чтобы для всех $z \geq I$ выполнялось нормировочное условие

$$\sum_{i=1}^{m} \alpha_i^{(z)} = \sum_{j=1}^{n} \beta_j^{(z)} = I.$$

Процесс подсчета весов столбцов и строк по формулам (5) называется центрированной качельной процедурой вычисления согласованных оценок строк и столбцов. Смысл названия "центрированная" задан в способе вычисления весов на основе "операции центрирования", т.е. учета отклонений элементов u_{ij} от их "центров тяжести" по строкам – $\bar{s}_i$ и "центров тяжести" по столбцам – $\bar{x}_j$. Этот учет ведется суммированием квадратов отклонений, умноженных на соответствующие веса, а именно, таких величин: $\beta_j^{(z)} \cdot (u_{ij} - \bar{s}_i)^2$, $\alpha_i^{(z)} (u_{ij} - \bar{x}_j)^2$. Вместо квадратов можно было бы суммировать и другие степени отклонений, причем процедура останется сходящейся. Наш выбор квадратов продиктован соображениями сходства получаемых сумм с формулами для получения дисперсии в математической статистике.

Другой вариант центрирования качельной процедуры задается такими формулами:

$$\alpha_i^{(z+1)} = \frac{\sum_{j=1}^{n} \beta_j^{(z)} \cdot (u_{ij} - \bar{x}_j)^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} \beta_j^{(z)} (u_{ij} - \bar{x}_j)^2} \qquad i = I,2, \ldots, m$$

$$z = 0,1,2, \ldots \qquad (6)$$

$$\beta_j^{(z+1)} = \frac{\sum_{i=1}^{m} \alpha_i^{(z)} (u_{ij} - \bar{s}_i)^2}{\sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_i^{(z)} (u_{ij} - \bar{s}_i)^2} \qquad j = I,2, \ldots, n ;$$

которые отличаются от формул (5) перестановкой центрирования по строкам и центрирования по столбцам.

Смысл названия "качельная" отражен на рис. 6 , где виден способ вычисления оценок, опирающийся на последовательные "качания" от вычисления весов столбцов к весам строк и обратно. То, что оценки получаются согласованными, имеет следующее смысловое выражение: получаемые с помощью формул (5) или (6) оценки столбцов $\beta_j^{(z)}$ и строк $\alpha_i^{(z)}$ сходятся при $z \to \infty$ к пределам β_j и α_i $j = 1,2, \ldots, n$; $i = 1,2, \ldots, m$, которые связаны между собой такими же формулами, т.е. согласованы. В случае формул (6), например, выражение такого согласования имеет вид:

$$\alpha_i = \frac{\sum_{j=1}^{n} \beta_j \, (u_{ij} - \bar{x}_j)^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} \beta_j \, (u_{ij} - \bar{x}_j)^2}$$

$$\beta_j = \frac{\sum_{i=1}^{m} \alpha_i \, (u_{ij} - \bar{s}_i)^2}{\sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_i (u_{ij} - \bar{s}_i)^2} ,$$

т.е. все α_i , $i = 1,2, \ldots, m$, выражены через β_j , $j = 1,2, \ldots, n$, и наоборот, все β_j выражены через α_i.

В программе "Центрированные качели" реализованы оба варианта подсчета весов строк α_i и весов столбцов β_j , как вариант, заданный формулами (5), так и вариант, заданный формулами (6) [33].

Итак, в целом методом согласованных оценок устанавливаются нагрузки строк и столбцов. Причем, как уже отмечалось, нагрузки отражают характер перекрытий строк и столбцов таблицы T - при одинаковом числе единиц в двух строках большую нагрузку будет иметь та строка, которая больше перекрывается по единицам с другими строками. Нагрузки в этом смысле выделяют типичные строки и столбцы T и потому в определенном смысле отображают ком-

бинаторные, или информационные свойства таблицы.

3. Упорядочение и распознавание объектов

Требуется упорядочить объекты $s_1, \ldots, s_m$ по степени проявления целевого признака x_{n+1} на основе их описаний характеристическими признаками $x_1, \ldots, x_n$. Признак x_{n+1} может иметь шкалу отношений или шкалу интервалов [32].

Перед началом обработки информации последняя преобразуется с помощью кодирующего отображения.

Это отображение ставит в соответствие признакам x_j, заданным на объектах s_i, числовую меру, называемую выраженностью признака.

Далее, пусть в данном методе упорядочения по величинам выраженности принимается заключение о степени проявления целевого признака x_{n+1}, что производится на основе косвенной информации, заключающейся в описании s_i признаком x_j. В этом важном случае, когда все признаки - бинарные, значение выраженности $t_{ij} = I$ соответствует следующему: x_j (s_i) свидетельствует о том, что геологически значение x_{n+1} (s_i) велико. Если же $t_{ij} = 0$, то это означает, что ожидаемое значение (в геологическом смысле степень проявления x_{n+1} незначительна), невелико. Заметим, что указанные выше оценки "велико", "невелико" основаны на косвенной предпосылке о том, что на основании одного признака сделать достоверное заключение о степени проявления целевого признака у всех объектов $s_1, \ldots, s_m$ не представляется возможным.

В том случае, когда информация замерена в шкале интервалов, тогда упорядочение объектов по степени проявления x_{n+1} представляет собой предмет, опирающийся на упорядочение по степени проявления суммарной информации, заключенной в x_j, $j = I,2, \ldots, n$. Исключим выполнение следующих условий:
$\max\limits_{i=1,\ldots,m} t_{ij} = \min\limits_{i=1,\ldots,m} t_{ij} = 0$, либо $\max t_{ij} = \min t_{ij} = I$

При решении геологических задач описанное выше кодирующее отображение производится специалистами геологами. Математиками же разработаны специальные алгоритмы и программы, которые осу -

ществляют оценки объектов в том случае, когда известно упорядочение объектов $s_1, \dots, s_m$ по x_{n+1}, и распространяют оценки степени проявления x_{n+1} на те объекты, где эта степень неизвестна [25,27].

Для решения задачи упорядочения по таблице $T = (t_{ij})_{m \times n}$ вычисляются значения $\omega_1, \omega_2, \dots, \omega_m$, и объекты $s_1, \dots, s_m$ упорядочиваются по убыванию ω_i.

Если значения x_{n+1} – запасы минерального сырья, то оценки ω_i, относящиеся к объектам, можно трактовать в случае геологических таблиц для месторождений, как числа, характеризующие соотношения между запасами, не сами запасы, а соотношения между ними. При этом существенно, что эти числа всегда оказываются положительными.

В заключении рассмотрим ситуацию, когда упорядочение $s_1, \dots, s_m$ по x_{n+1} является известным, а степень проявления x_{n+1} у объекта s, охарактеризованного признаками $x_1, \dots, x_n$, неизвестна. Вычислим оценки $\omega_1, \dots, \omega_m$ объектов $s_1, \dots, s_m$, а также оценку $\omega(s)$ объекта s.

Если упорядоченность объектов $s_1, \dots, s_m$ по степени проявления x_{n+1} достаточно хорошо коррелирована с их упорядочением по значению ω_i, то, сравнивая оценку $\omega(s)$ нового объекта s, мы можем найти место объекта s (по степени проявления x_{n+1} на s) в ряду объектов $s_1, \dots, s_m$ [I9]

Описываемый алгоритм и реализующая его программа основаны на приложении метода согласованных оценок ("Качели") и общей схеме распознавания на базе понятия типичной строки ("реплики").

Пусть заданы два класса объектов, чьими эталонами являются $s_1^1, \dots, s_{m_1}^1$ и $s_1^2, \dots, s_{m_2}^2$ соответственно. Описания эталонов бинарными характеристическими признаками $x_1, \dots, x_n$ образуют таблицы T_I и T_2*). Строка $(\tilde{t}_1^\kappa, \dots, \tilde{t}_n^\kappa) = \tilde{s}^\kappa$, где $\kappa = I,2$, называется типичной для T_κ, если

---

*) Где на пересечении i – й строки и j –го столбца стоит значение признака x_j на объекте s_i^κ, $\kappa = I,2$ единица, если x_j выполняется и ноль – в противном случае.

$$\tilde{t}_{ij}^{\kappa} = \begin{cases} 1 \text{ при } \sum_{i=1}^{m_\kappa} t_{ij}^{\kappa} \geq \frac{m_\kappa}{2} \\ 0 \text{ при } \sum_{i=1}^{m_\kappa} t_{ij}^{\kappa} < \frac{m_\kappa}{2}, \end{cases} \qquad (7)$$

где t_{ij}^{κ} – элемент таблицы T_κ. При $\sum_{i=1}^{m_\kappa} t_{ij}^{\kappa} = \frac{m_\kappa}{2}$ значение $\tilde{t}_{ij}^{\kappa}$ может определяться также из неформальных, практических соображений или случайным путем. Рассмотрим следующую схему распознавания на базе понятия "реплика". Пусть $P^{\kappa} = (P_1^{\kappa}, \ldots, P_n^{\kappa})$ – произвольный неотрицательный вектор нагрузок столбцов таблицы T_κ, $\sum_{j=1}^{n} P^{\kappa} = 1$. Для произвольной строки $S = (t_1, \ldots, t_n)$ положим $r(P^{\kappa}, S) = \sum_{j=1}^{n} |t_j + \tilde{t}_j^{\kappa} - 1| P_j^{\kappa}$ и $R(P^1, P^2, S) = \frac{r(P^1, S)}{r(P^2, S)}$ при $r(P^2, S) \neq 0$. Величина $r(P^{\kappa}, S)$ представляет собой взвешенное число совпадений строки S с типичной строкой таблицы T_κ и ее можно принять в качестве меры близости к классу, представленному эталонами $S_1^{\kappa}, \ldots, S_{m_\kappa}^{\kappa}$. Величина $R(P^1, P^2, S)$ оценивает тяготение строки S к одному из указанных двух классов при заданных мерах близости. Если $R(P^1, P^2, S) > 1$, то S тяготеет к первому классу, если $R(P^1, P^2, S) < 1$, то – ко второму. При $R(P^1, P^2, S) = 1$, мы ничего не можем сказать о принадлежности S. Естественно сформировать такое решающее правило для диагностики испытуемых объектов S, описаниями которых являются строки $(t_1, \ldots, t_n)$:

а) при $R(P^1, P^2, S) \geq 1 + \varepsilon_1$, S относится к первому классу;

б) при $R(P^1, P^2, S) \leq 1 - \varepsilon_2$, S относится ко второму классу;

в) при $1 - \varepsilon_2 < R(P^1, P^2, S) < 1 + \varepsilon_1$, S не распознается.

Здесь пороги $\varepsilon_1, \varepsilon_2 > 0$ определяют "решительность" алгоритма распознавания.

Наиболее простой процедурой подобного рода является, по-ви-

димому, следующая: положим $P = P^{к}_{II} = (\frac{1}{n}, \ldots, \frac{I}{n})$, т.е. все признаки считаются в равной степени существенными для диагностики. Тогда $\tau(P^{к}_{II}, S) = \tau_{II}(S) = \frac{I}{n}\sum_{j=1}^{n} |t_j + \tilde{t}^{к}_j - I|$, т.е. равняется числу совпадений строки S с типичной строкой $\tilde{S}^{к}$, поделенному на n, а $R(P^{I}_{II}, P^{2}_{II}, S) = R_{II}(S)$ – отношение числа совпадений с типичной строкой первого класса к числу совпадений с типичной строкой второго класса.

Для того, чтобы указанное решающее правило можно было применять к решению той или иной конкретной задачи диагноза, необходимо убедиться в его способности распознавать объекты обучения $S^{1}_{1}, \ldots, S^{1}_{m_1}, S^{2}_{1}, \ldots, S^{2}_{m_2}$, т.е. проверить, что если $S^{к}_{j}$ – строка, соответствующая объекту $S^{к}_{j}$, где $j = I, \ldots, m_к$, то

$$R(P^{I}, P^{2}, S^{1}_{j}) > I \qquad j = I, \ldots, m_1$$

$$(*) \qquad R(P^{I}, P^{2}, S^{2}_{j}) < I \qquad j = I, \ldots, m_2.$$

Такое распознавание объектов обучения будем называть устойчивым.*) Однако в ряде случаев условия (*) могут не выполняться, тем не менее может быть сформулирована приемлемая с содержательной точки зрения процедура распознавания. Такая ситуация имеет место, если выполняются условия (**) $R(P^{I}, P^{2}, S^{1}_{j}) > R(P^{I}, P^{2}, S^{2}_{j'})$ для всех $j = 1, \ldots, m_1, j' = 1, \ldots, m_2$. Положим $\delta_1 = \min_{j=1,\ldots,m_1} (P, P^{2}, S^{1}_{j})$, $\delta_2 = \max_{j'=1,\ldots,m_2} (P^{I}, P^{2}, S^{2}_{j'})$, тогда (**) – означает, что $\delta_1 > \delta_2$. Сформулируем решающее правило:

а) при $R(P^{I}, P^{2}, \tilde{S}) \geq \delta_1$, S относится к первому классу;

б) при $R(P^{I}, P^{2}, S) \leq \delta_2$, S относится ко второму классу;

в) при $\delta_2 < R(P^{I}, P^{2}, S) < \delta_1$, S не распознается.

Распознавание по этому правилу при невыполнении условия (*) называется неустойчивым, хотя содержательно устойчивое распознавание более приемлемо.

Рассмотрим процедуры распознавания на основе нагрузок, вычисленных по центрированным качельным процедурам.

*) Такая проверка не исключает и обычного внешнего экзамена процедуры распознавания.

Пусть имеются две группы объектов обучения $s_1^1, \ldots, s_{m_1}^1$ и $s_1^2, \ldots, s_m^2$. Из их описаний обычным путем составим таблицы X_I и X_2. Пусть $\rho^\ell = (\rho_1^\ell, \ldots, \rho_n^\ell)$-вектор нагрузок столбцов таблицы X_ℓ, $\ell = I,2$, подсчитанных по МСО. Для произвольной строки $S = (t_1, \ldots, t_n)$ определим

$$\bar{z}(\rho^\ell, S) = \sum_{j=1}^{n} (x_j - \bar{x}_j^\ell)\cdot\rho_j^\ell \quad \text{и}$$

$$\bar{R}(\rho^1, \rho^2, S) = \frac{\bar{z}(\rho^1, S)}{\bar{z}(\rho^2, S)}, \quad (\bar{z}(\rho^2, S) \neq 0), \tag{8}$$

где $\bar{x}_j^\ell$ - среднее арифметическое j-го столбца таблицы X_ℓ.

Содержательно величина $\bar{z}(\rho^\ell, s)$ отражает удаленность строки S от "средней строки" таблицы X_ℓ, $\ell = I,2$. Чем выше значение $\bar{z}(\rho^\ell, S)$, тем дальше строка S отстоит от строки, составленной из значений $\bar{x}_j^\ell$. Причем коэффициенты ρ_j^ℓ характеризуют величину рассеяния значений j-го столбца относительно их среднего арифметического $\bar{x}_j^\ell$ так, что чем выше это рассеяние, тем больше значение ρ_j^ℓ в ряду $\rho_1^\ell, \rho_2^\ell, \ldots, \rho_n^\ell$. Или можно сказать, что чем выше концентрация значений j-го столбца вокруг среднего, тем меньше значение ρ_j^ℓ в ряду $\rho_1^\ell, \ldots, \rho_n^\ell$. То же самое можно сказать о величине $\bar{z}(\rho^\ell, S)$: чем выше концентрация значений компонент строки S относительно соответствующих компонент средней строки таблицы X^ℓ, тем меньше значение $\bar{z}(\rho^\ell, S)$. Поэтому, чем меньше значение отношения $R(\rho^1, \rho^2, S)$, тем больше строка S тяготеет к таблице X_I и наоборот, чем больше значение этого отношения, тем больше S тяготеет к X_2.

Положим $\alpha_1 = \max \bar{R}(\rho^1, \rho^2, s_j^1)$, $\alpha_2 = \min \bar{R}(\rho^1, \rho^2, s_j^2)$ и сформулируем нижеследующее решающее правило для распознавания по нагрузкам, вычисляемым ЦКП.

I) Пусть $\alpha_1 < \alpha_2$, тогда:

а) при $\bar{R}(\rho^1, \rho^2, S) \leq \alpha_1$, S относится к классу, представленному таблицей X_1;

б) при $\bar{R}(\rho^1, \rho^2, S) \geq \alpha_2$, S относится к классу, представленному X_2;

в) при $\alpha_2 > \bar{R}(\rho^1, \rho^2, S) > \alpha_1$, S не распознается.

2) Если $\alpha_1 \geq \alpha_2$, то:

а) при $\bar{R} < \alpha_2$, S относится к X_I;

б) при $\bar{R} > \alpha_1$, S относится к X_2;

в) при $\alpha_1 \geq \bar{R} \geq \alpha_2$, S не распознается.

Таким образом, если $\alpha_1 = \alpha_2 = \alpha$, то:

а) при $\bar{R} < \alpha$, S относится к X_1;

б) при $\bar{R} > \alpha$, S относится к X_2;

в) при $\bar{R} = \alpha$, S не распознается.

4. Нагрузки как меры значимости объектов

Сформулируем основные требования к выбору нагрузок P_I, P_2. Обозначим через E^N, $N = I,2, \ldots$ N - мерный единичный куб. Пусть T - таблица описаний объектов $s_1, \ldots, s_m$ бинарными признаками $x_1, \ldots, x_n$. Произвольному m-мерному вектору-столбцу $x = (\alpha_1, \ldots, \alpha_m)'$, где "$'$" - знак транспонирования, сопоставим его типизированный вариант $\tilde{x} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_m)'$, где $\tilde{\alpha}_i = \alpha_i$ при $\sum_{i=1}^{m} \alpha_i \geq \frac{m}{2}$ и $\tilde{\alpha}_i = I - \alpha_i$ при $\sum_{i=1}^{m} \alpha_i < \frac{m}{2}$ *). Таблице T сопоставим ее типизированный вариант $\tilde{T}$. Столбцами таблицы $\tilde{T}$ являются векторы $\tilde{x}_j$ - типизированные варианты столбцов x_j таблицы T. Для произвольных векторов $a = (a_1, \ldots, a_N)$, $b = (b_1, \ldots, b_N)$, $a \leq b$ означает, что $a_1 \leq b_1, \ldots, a_N \leq b_N$ **). Пусть T составлена из описаний объектов $s_1, \ldots, s_m$ бинарными признаками $x_1, \ldots, x_N$.

---

*) Таким образом при $x = x_j$, $j = I, \ldots, n$ в первом случае типичным значением x_j для объектов $s_1, \ldots, s_m$ является единица, а во втором - ноль. Кроме того, при $x = x_j$ и $\sum_{i=1}^{m} \alpha_i = \frac{m}{2}$ x_j может определяться и иным путем (см.[I9]),но так, чтобы это не приводило к недоразумениям.

**) $a < b$ означает, что $a \leq b$ и $a \neq b$.

Определение I. Неотрицательная функция $\mu(x)$, заданная на E^m, называется мерой характеристичности признака, если для любых x', $x'' \in E^m$ $\tilde{x}' < \tilde{x}''$ влечет $\mu(x') < \mu(x'')$.

Пусть $\tilde{s}^* = (\tilde{t}_1^*, \ldots, \tilde{t}_n^*)$ - реплика T *). Для произвольного вектора $s \in E$ положим $\tilde{s} = (\tilde{t}_1, \ldots, \tilde{t}_n)$, где $\tilde{t}_j = |t_j + \tilde{t}_j^* - I|$.

Определение 2. Неотрицательная функция $\eta(s)$, заданная на E^n, называется мерой типичности объекта, связанной с $\tilde{s}^*$, если для любых s', $s'' \in E^n$ $\tilde{s}' < \tilde{s}''$ влечет $\eta(s') < \eta(s'')$. Если $\alpha = (t_1, \ldots, t_m)'$ - столбец значений признака x на объектах $s_1, \ldots, s_m$, то число $\mu(\alpha)$ называется мерой характеристичности признака x для объектов $s_1, \ldots, s_m$. Положим $\mu(x) = \mu(\alpha)$. Для объекта s число $\eta(s)$ называется мерой типичности s в классе, представленном эталонами $s_1, \ldots, s_m$.

Отметим, что если $\mu(x)$ - мера характеристичности, то $\eta(s) = a \sum_{j=1}^{n} \tilde{t}_j \cdot \mu(x_j)$ будет мерой типичности и наоборот, если $\eta(x)$ - мера типичности, то $\mu(x) = b \cdot \sum_{i=1}^{m} \tilde{\alpha}_i \cdot \eta(s_i)$ - мера характеристичности, где a, b - произвольные положительные числа,

$$\mu(x_j) > 0, \quad j = I, \ldots, n, \qquad \eta(s_i) > 0, \quad i = I, \ldots, m. \tag{9}$$

При определении вектора нагрузок P^I, P^2 естественно потребовать, чтобы числа P_j^I, P_j^2 были мерами характеристичности признаков x_j для объектов $s_1^1, \ldots, s_{m_1}^1$ и $s_1^2, \ldots, s_{m_2}^2$ соответственно. Тогда $r(P^k, s) = \eta^k(s)$ будет мерой типичности объекта s в указанном выше смысле.

Простейшим примером меры характеристичности является частота, с которой признак принимает свое типичное значение. Простейшим примером меры типичности будет число совпадений строки s с репликой таблицы T. Однако при таком способе оценки мера типичности объекта не зависит от того, насколько характеристичны свой-

*) Где $\tilde{t}_{ij}$ представляет собой типичное значение признака $j = I, \ldots, n$.

ственные ему признаки, а мера характеристичности признака не зависит от типичности тех объектов, на которых он принимает свое типичное значение. Для установления взаимосвязи между η и μ потребуем, чтобы для них выполнялись одновременно соотношения согласования. В этом случае мерой характеристичности признака является (с точностью до постоянного множителя) сумма мер типичности тех объектов, где он принимает свое типичное значение. И наоборот, мерой типичности объекта s является сумма мер характеристичности тех признаков, которые на объекте s принимают свое типичное значение. При этом мера типичности, порождаемая по формуле функцией $\mu(x_j)$, совпадает (с точностью до постоянного множителя) с $\eta(s)$, а мера характеристичности, порождаемая $\eta(s)$, совпадает с $\mu(x_j)$. Иначе говоря, меры характеристичности и типичности являются взаимосогласованными. Полагая для $i = 1, \dots, m$, $j = 1, \dots, n$, $\mu(s_i) = \tilde{\omega}_i$, $\mu(x_j) = \tilde{\pi}_j$, $\frac{1}{a} = c$, $\frac{1}{b} = d$ запишем соотношения согласования для строк и столбцов таблицы T:

$$c\,\tilde{\omega}_i = \sum_{j=1}^{n} \tilde{t}_{ij} \cdot \tilde{\pi}_j$$

$$d\,\tilde{\pi}_j = \sum_{i=1}^{m} \tilde{t}_{ij} \cdot \tilde{\omega}_i \,, \qquad (10)$$

где c и d – произвольные положительные числа. Приходим к основным уравнениям метода согласованных оценок для таблицы T. Решение этой системы – "качельные" нагрузки столбцов и строк T – определяют взаимосогласованные меры характеристичности признаков и типичности объектов*).

Пусть $\tilde{\pi}^{\kappa} = (\tilde{\pi}_1^{\kappa}, \dots, \tilde{\pi}_n^{\kappa})$, $\tilde{\omega}^{\kappa} = (\tilde{\omega}_1^{\kappa}, \dots, \tilde{\omega}_{m_\kappa}^{\kappa})$, "качельные" нагрузки столбцов и строк таблицы $\tilde{T}_{\kappa}$, φ^{κ} – частота принятия признаком x_j своего типичного значения. Справед-

*) Для этого, кроме требования измеримости $\tilde{T}$, следует предположить, что $\tilde{T}$ не имеет нулевой строки. Практически любая геологическая таблица удовлетворяет этому требованию.

лича формула

$$\tilde{\pi}_j^{\kappa} = \gamma^{\kappa}\left[\mathcal{P}_j^{\kappa} + \sum_{i=1}^{m_{\kappa}} \left(\frac{\tilde{\omega}_i^{\kappa}}{\gamma^{\kappa}} - \frac{I}{m_{\kappa}}\right) t_{ij}^{\kappa}\right] = \gamma^{\kappa}\left(\mathcal{P}_j^{\kappa} + \sum_{i=1}^{m_{\kappa}} \sigma_i^{\kappa} \cdot t_{ij}^{\kappa}\right), \quad (II)$$

где $\gamma^{\kappa} = \sum_{i=1}^{m_{\kappa}} \tilde{\omega}_i^{\kappa}$, $\sigma_i^{\kappa} = \frac{\tilde{\omega}_i^{\kappa}}{\gamma^{\kappa}} - \frac{I}{m_{\kappa}}$.

Причем $\sigma_i^{\kappa} \geq 0$ означает, что $\tilde{\omega}_i^{\kappa} \geq \frac{\gamma^{\kappa}}{m}$, где $\frac{\gamma^{\kappa}}{m_{\kappa}}$ - среднее арифметическое чисел $\tilde{\omega}_1^{\kappa}, \ldots, \tilde{\omega}_{m_{\kappa}}^{\kappa}$. Отметим также, что $\tilde{\omega}_i^{\kappa} = \varphi(\tilde{\pi}^{\kappa}, s_i^{\kappa})$.

Использование $\tilde{\omega}_i^{\kappa}$ как меры типичности объекта s_i^{κ} в группе объектов $s_1^{\kappa}, \ldots, s_{m_{\kappa}}^{\kappa}$ проводилось на ряде геологических примеров и дало результаты, хорошо согласующиеся с содержательной трактовкой типичности геологических объектов. Таким образом, можно сказать, что при $\sigma_i^{\kappa} > 0$ "типичность" объекта s_i^{κ} в группе $s_1^{\kappa}, \ldots, s_{m_{\kappa}}^{\kappa}$ выше средней, а при $\sigma_i^{\kappa} < 0$ - ниже средней. Если $\mathcal{P}_j^{\kappa}$ оценивает частоту встречаемости типичного значения, то $\sum_{i=1}^{m_{\kappa}} \sigma_i^{\kappa} t_{ij}^{\kappa}$ оценивает "качество" выполнения этого значения. Например, если для x_j , x_j' $\mathcal{P}_j^{\kappa} = \mathcal{P}_{j'}^{\kappa}$, то неравенство $\tilde{\pi}_{j'}^{\kappa} > \tilde{\pi}_j^{\kappa}$ означает, что объекты s_i^{κ} , для которых $t_{ij'} = I$, в среднем более типичны, чем объекты s_i^{κ} , для которых $t_{ij} = I$.

Все вышесказанное позволяет утверждать, что с содержательной точки зрения величина $\tilde{\pi}_j^{\kappa}$ может быть истолкована как мера характеристичности признака x_j для объектов $s_1^{\kappa}, \ldots, s_{m_{\kappa}}^{\kappa}$, совмещающая в себе как частотную, так и "качественную" оценку. Поэтому естественно ожидать, что если мы используем полученные нагрузки столбцов $\tilde{\pi}_j^{\kappa}$ в описанной выше диагностической схеме, положив

$$P_I^{\kappa} = \left(\frac{\tilde{\pi}_1^{\kappa}}{\sum_{j=1}^{n} \tilde{\pi}_j^{\kappa}}, \ldots, \frac{\tilde{\pi}_n^{\kappa}}{\sum_{j=1}^{n} \tilde{\pi}_j^{\kappa}}\right), \quad z_I^{\kappa} = (P_I^{\kappa}, S), R_I(S) = \frac{z(P_I^1, S)}{z(P_I^2, S)}, \quad (I2)$$

то полученный алгоритм распознавания окажется применимым для широкого круга геологических задач. Опыт его применения для решения ряда геологических задач показал следующее: во всех слу -

чаях результат распознавания по R_I был не хуже, чем по R_{II}. Отмечены случаи неразделения классов по R_{II} (т.е. $\delta_1 > \delta_2$) при их разделении по P_I, а также устойчивого распознавания по R_I при неустойчивом распознавании по R_{II}, хотя даже и в таких ситуациях общая картина распределения значений R_I и R_{II} примерно одинаковая.

Заключительные высказывания по МСО в целом сводятся к следующему. В основу метода заложена процедура подсчета оценки строк и столбцов таблиц последовательным приближением (итерациями) этих оценок к предельным. Основополагающий нецентрированный вариант метода, в связи с рядом причин дополнен центрированным вариантом. Взаимная дополнительность вариантов была хорошо вскрыта решением конкретных задач прогноза . В частности, некоторое преимущество центрированных процедур перед нецентрированными состоит в большей строгости распознавальческой схемы. Нецентрированные процедуры имеют преимущество в задаче обнаружения выразительности целевой значимости объектов, а также в вопросах упорядочения строк и столбцов.

Программы вычисления оценок строк и столбцов таблиц, коэффициентов в задачах распознавания (как для центрированного, так и для нецентрированного вариантов) реализованы на ЭВМ, на алгоритмическом языке "Альфа-6". Их достоинством является возможность обрабатывать таблицы заполненные как бинарными, так и количественными значениями признаков [10,14,40,42] . Эти программы позволяют обрабатывать большеформатные таблицы ($m \times n$ =100x200), что посуществу позволяет давать ориентировочные решения для многих задач практического направления, возникающих в сфере геологического прогноза и поиска.

§3 КОРРЕЛЯЦИИ ОЦЕНОК МСО СО ЗНАЧЕНИЯМИ ЦЕЛЕВОГО ПРИЗНАКА

Как уже сообщалось выше (см. §2) в ряде случаев вычисленные по МСО нагрузки строк таблицы (существенности объектов) хорошо коррелируют с упорядоченностью объектов по значениям целевого признака.

Принятое подразделение признаков, которыми характеризуются объекты, следующее. Полная характеристика объектов*) исследования (месторождений) представляется описанием, расчлененным на характеристические признаки. Как правило, количество характеристических признаков является избыточным. Эта избыточность заранее неизвестна и обнаруживается целевыми установками и обработкой таблиц. После решения таблиц, в соответствии с целью обработки, выделяется неизбыточная часть признаков, которая, по аналогии с геологическими определениями, называется поисковой совокупностью признаков. Именно поисковая совокупность признаков родственна целевому, по которому объекты ранжируются в соответствии со своей значимостью. Упорядочение объектов по совокупности поисковых признаков, совпадающее с упорядоченностью по целевому, и является предметом внимания этого параграфа.

В связи с большой практической важностью этого факта представляется уместным провести математическое рассмотрение сравнения оценок МСО с линейной регрессией целевого признака на поисковые. Далее перейдем к формализованным представлениям вопроса. Выявление эффекта корреляции нагрузок объектов с целевым признаком имеет и математический интерес, поскольку расширяются результаты на контакте различных математических приемов.

Пусть наряду с массивом X задан также некоторый целевой признак $z = (z_1, \dots, z_m)^T$. (Термин "целевой" означает, что

---

*) Полнота описания объектов здесь понимается в содержательном смысле, т.е. полная характеристика исчерпывает естественную информацию о месторождении, имеющуюся в распоряжении геолога.

и его значение надо уметь предсказать, но используя данные X, а также некоторую обучающую выборку, в которой значение z задано). В отличие от z , признаки $x_{\cdot\kappa} = (x_{1\kappa}, \ldots, x_{m\kappa})^T$ будем называть поисковыми.

Запишем уравнение линейной регрессии признака z на поисковые признаки:

$$z_i = \gamma_1 x_{i1} + \ldots + \gamma_n x_{in} + \varepsilon_i \quad . \qquad (1)$$

Или в матричном виде:

$$z = X\gamma + \varepsilon; \quad \gamma = (\gamma_1, \ldots, \gamma_n)^T . \qquad (2)$$

Величины γ_i суть коэффициенты регрессии.

Введем обозначение: g_κ есть ковариация z с поисковым признаком $x_{\cdot\kappa}$:

$$g_\kappa = \sum_{i=1}^{m} z_i x_{i\kappa} \; ; \; g = (g_1, \ldots, g_n)^T \; ; \; g = Xz . \qquad (3)$$

Вектор коэффициентов регрессии γ находится из условия минимума функционала погрешности:

$$Q = \sum_{i=1}^{m} \varepsilon_i^2 = \sum_{i=1}^{m} (z_i - \sum_{\kappa=1}^{n} \gamma_\kappa \cdot x_{i\kappa})^2 \longrightarrow min \qquad (4)$$

Решение оптимизационной задачи (4) хорошо изучено. Оно имеет вид:

$$\gamma = [X^T X]^{-1} g = [X^T X]^{-1} X z \quad . \qquad (5)$$

Наилучшая (в смысле метода наименьших квадратов) линейная оценка признака z (по массиву X):

$$u = X\gamma = X[X^T X]^{-1} \cdot X^T z = X[X^T X]^{-1} g \; . \qquad (6)$$

Рассмотрим теперь уравнения МСО:

$$\begin{aligned} \nu\alpha &= X\beta , \\ \mu\beta &= X^T\alpha . \end{aligned} \qquad (7)$$

Здесь $\alpha = (\alpha_1, \ldots, \alpha_m)^T$ – "нагрузки" объектов,

$\beta = (\beta_1, \ldots, \beta_n)$ – "нагрузки" признаков.

(Величины ν , μ суть нормирующие множители, обеспечивающие выполнение условий: $\alpha^T\alpha = \mu\nu$, $\beta^T\beta = I$).

Из (7) следует:

$$\mu\nu\beta = \nu X^T\alpha = X^T X\beta \; ;$$

$$X^T \cdot X\beta = \mu\nu\beta \;; \quad \beta = \mu\nu\left[X^T X\right]^{-1}\beta,$$

$$\nu\alpha = X\beta \Longrightarrow \nu\alpha = \mu\nu\; X\; \left[X^T X\right]^{-1}\beta.$$

В результате получаем:

$$\mu^{-1}\alpha = X\left[X^T X\right]^{-1}\beta. \qquad (8)$$

Множитель μ^{-1} для дальнейшего несущественен, и его можно опустить (считая, что он введен в компоненты вектора α). Введем обозначение: $\mathcal{D} = \| d_{ik} \| = X\left[X^T X\right]^{-1}$. Перепишем формулы (6),(8) в виде:

$$u = \mathcal{D} g, \qquad (9)$$

$$\alpha = \mathcal{D}\beta. \qquad (10)$$

Линейные (регрессионные) оценки u_i признака z естественным образом упорядочивают множество объектов. (На практике задача обычно ставится так, что чем больше u_i, тем "лучше" или "важнее" i-й объект). С другой стороны, объекты можно (чисто формально) упорядочить по возрастанию нагрузок α_i. Если бы эти два упорядочения различались не очень существенно, можно было бы, для грубой классификации объектов по целевому признаку, пользоваться объектными нагрузками. Это может быть полезно в тех случаях, когда линейные оценки u_i получить невозможно, ввиду отсутствия обучающей выборки. Вычислительные эксперименты с конкретными геологическими данными показали, что иногда объектные нагрузки хорошо коррелируют с целевым признаком.

Попробуем дать этому факту формальное обоснование.

Выделим в X два объекта (строки) с номерами i, j. Для них в силу (9), (10) справедливы соотношения:

$$\left.\begin{aligned} u_i &= d_{i1} g_1 + \ldots + d_{in} g_n \\ u_j &= d_{j1} g_1 + \ldots + d_{jn} g_n \end{aligned}\right\}, \qquad (11)$$

$$\left.\begin{aligned} \alpha_i &= d_{i1}\beta_1 + \ldots + d_{in}\beta_n \\ \alpha_j &= d_{j1}\beta_1 + \ldots + d_{jn}\beta_n \end{aligned}\right\}. \qquad (12)$$

Обозначение: $\theta_k(i,j) = d_{ik} - d_{jk}$, $(k = 1, \ldots, n)$.
Образуем разности:

$$\Delta u = u_i - u_j = \theta_1 g_1 + \ldots + \theta_n g_n, \qquad (13)$$

$$\Delta\alpha = \alpha_i - \alpha_j = \theta_1\beta_1 + \ldots + \theta_n\beta_n . \qquad (14)$$

Требуется выяснить, при каких условиях из неравенства $\Delta\alpha>0$ следует $\Delta u > 0$ (и наоборот). Если такое следование имеет место, упорядочение по возрастанию α_i влечет аналогичное упорядочение объектов по возрастанию линейных оценок признака z.

Примем, что данные x_{ik} – неотрицательные (это предположение играло существенную роль в основном варианте МСО, который был использован для вычислительных экспериментов). Кроме того, предположим, что $z_1 \geq 0$.

Прямое и обратное преобразование МГК:

$$X = YC^T , \quad Y = XC . \qquad (15),(16)$$

Здесь $Y = \| y_{ik} \|$ – матрица главных компонент;

y_{ik} есть значение к-той главной компоненты на i-м объекте.

$C = \| C_i^k \|$ – ортогональная матрица, столбцы которой суть собственные векторы матрицы $X^T X$; (напомним, что данные – нецентрированные).

Развернутая запись уравнений (15),(16):

$$x_{ik} = C_k' y_{i1} + \ldots + C_k^n y_{in} , \qquad (17)$$

$$y_{ik} = C_1^k x_{i1} + \ldots + C_n^k x_{in} . \qquad (18)$$

Выразим вектор ковариаций g (см. формулу (3)) через элементы матрицы главных компонент:

$$g_k = \sum_{i=1}^{m} z_i x_{ik} = C_k' \sum_{i=1}^{m} z_i y_{i1} + \ldots + C_k^n \sum_{i=1}^{m} z_i y_{in} \qquad (19)$$

Обозначение:

$$v_s = \sum_{i=1}^{m} z_i y_{is} . \qquad (20)$$

Величины v суть ковариации целевого признака с главными компонентами. Таким образом имеем:

$$g_k = C_k' v + \ldots + C_k^n v_n . \qquad (21)$$

Далее, введя обозначение

$$h_s = \theta_1 C_1^s + \ldots + \theta_n C_n^s , \qquad (s = 1, \ldots, n), \qquad (22)$$

представим разность Δu (см.(13)) в виде:

$$\Delta u = \sum_{s=1}^{n} v_s h_s . \qquad (23)$$

Обратимся к разности $\Delta\alpha$ (см. (I4)). Вектор β есть главный собственный вектор матрицы X^TX (именно такое решение уравнений (7) использовано в практически реализованном варианте МСО). Таким образом, $\beta_K = C'_K$. Следовательно:

$$\Delta\alpha = \theta_1 C'_1 + \theta_2 C'_2 + \ldots + \theta_n C'_n = h_1 . \qquad (24)$$

В силу (23),(24) имеем:

$$\Delta u = v_1 \cdot \Delta\alpha + \sum_{s=2}^{n} v_s h_s \qquad (v_1 = \sum_{i=1}^{m} z_i y_{i_1}) \qquad (25)$$

Так как, по предположению, $x_{iK} \geq 0$, то, по теореме Фробениуса о главном собственном векторе матрицы с неотрицательными элементами, $\beta_K \geq 0$, $1 \leq k \leq n$ и, значит, $y_{i_1} \geq 0$, $1 \leq i \leq m$. Так как $z_i \geq 0$, имеем: $v_1 \geq 0$. Случай $v_1 = 0$ можно не учитывать как вырожденный (поскольку речь идет о практических данных). Итак, при сделанных допущениях $v_1 > 0$. Остальные величины v_s ($s \geq 2$), а также h_s , могут иметь разные знаки. Выражение (25) позволяет сделать качественный вывод: эффект корреляции объектных нагрузок с целевым признаком следует ожидать, прежде всего, когда первая главная компонента сильно доминирует над остальными. Но этот случай не единственный и, быть может, не самый типичный (среди благоприятных для названного эффекта). Величины h_s можно рассматривать как случайные, с нулевым средним, при этом слагаемые $v_s h_s$ ($s \geq 2$), могут "гасить" друг друга. Поэтому требование сильного доминирования первой главной компоненты не является, вообще говоря, необходимым.

§4 ЦЕНТРИРОВАННЫЙ ВАРИАНТ МЕТОДА

Цель данного параграфа – описание центрированных качельных процедур метода согласованных оценок.

I. Согласование

Дана таблица T числовых элементов t_{ij}, стоящих на пересечении i-й строки – (S_i) и j-го столбца – (x_j) таблицы T, $i = I,2, \ldots, m$, $j = I,2, \ldots, n$; m – число строк, а n – число столбцов таблицы. Элементы t_{ij} суть неотрицательные числа. Строки таблицы T соответствуют некоторым объектам, а столбцы – признакам этих объектов. Величина t_{ij} есть значение признака x_j на объекте S_i (для удобства мы отождествили обозначения строк S_i и объектов, которым эти строки соответствуют, а также обозначения столбцов x_j и признаков). Предполагается, что $m \geq 2$, $n \geq 2$ (рис. 7).

$$
T = \begin{array}{c}
\begin{array}{cccccc} (x_1) & (x_2) & \ldots & (x_j) & \ldots & (x_n) \end{array} \\
\left[\begin{array}{cccccc}
t_{11} & t_{12} & \ldots & t_{1j} & \ldots & t_{1n} \\
t_{21} & t_{22} & \ldots & t_{2j} & \ldots & t_{2n} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
t_{i1} & t_{i2} & \ldots & t_{ij} & \ldots & t_{in} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
t_{m1} & t_{m2} & \ldots & t_{mj} & \ldots & t_{mn}
\end{array}\right]
\begin{array}{c} (S_1) \\ (S_2) \\ \ldots \\ (S_i) \\ \ldots \\ (S_m) \end{array}
\end{array}
$$

Рис. 7

Для получения численных оценок строк и столбцов таблицы T вводятся следующие итерационные процедуры. Последующие приближения $\pi_i^{(k+1)}$, $\omega_j^{(k+1)}$ получаются через предыдущие $\pi_i^{(k)}$, $\omega_j^{(k)}$ по формулам:

$$\pi_i^{(k+1)} = \frac{I}{\lambda_k} \sum_{j=1}^{n} t_{ij} \cdot \omega_j^{(k)}, \quad i = I,2,\ldots,m; \qquad (I)$$

$$\omega_j^{(k+1)} = \frac{I}{\mu_k} \sum_{i=1}^{m} t_{ij} \cdot \pi_i^{(k)}, \quad j = I,2,\ldots,n, \quad k \geq 0; \qquad (2)$$

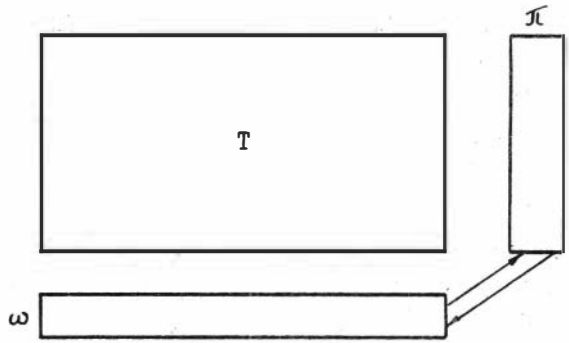


Рис. 8 Иллюстрация к формулам (4),(5)

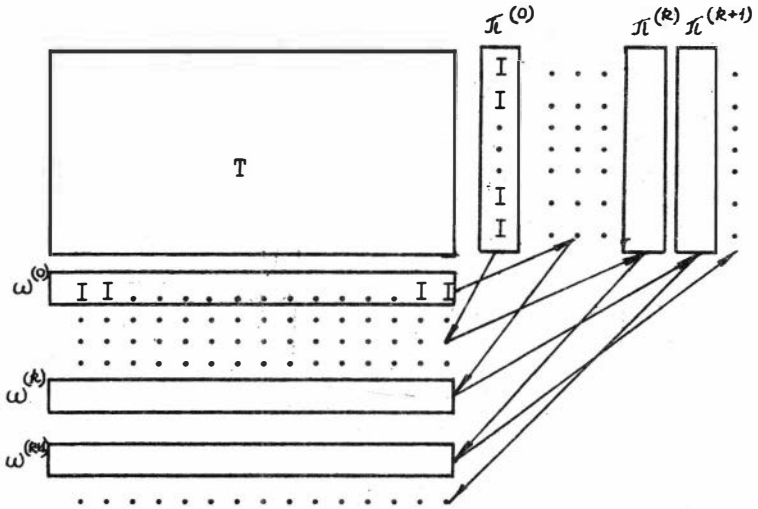


Рис. 9 Иллюстрация к формулам (I),(2),(3)

где λ_k, μ_k – некоторые нормирующие множители; например, в работе [22] употребляются множители вида:

$$\lambda_k = \max_{1 \le i \le m} \left\{ \sum_{j=1}^{n} t_{ij} \cdot \omega_j^{(k)} \right\}, \quad \mu_k = \max_{1 \le j \le n} \left\{ \sum_{i=1}^{m} t_{ij} \cdot \pi_i^{(k)} \right\},$$

а в работе [33] – вида:

$$\lambda_k = \sum_{i=1}^{m} \sum_{j=1}^{n} t_{ij} \cdot \omega_j^{(k)}, \quad \mu_k = \sum_{j=1}^{n} \sum_{i=1}^{m} t_{ij} \cdot \pi_i^{(k)}.$$

В качестве начальных приближений для оценок строк и столбцов берутся значения:

$$\pi_i^{(0)} = 1, \quad i = 1,2, \ldots, m \; ; \quad \omega_j^{(0)} = 1, \quad j = 1,2, \ldots, n. \qquad (3)$$

Как доказано в работе [9], для почти всех таблиц существуют пределы:

$$\pi_i = \lim_{k \to \infty} \pi_i^{(k)}, \quad \omega_j = \lim_{k \to \infty} \omega_j^{(k)},$$
$$\lambda = \lim_{k \to \infty} \lambda_k, \quad \mu = \lim_{k \to \infty} \mu_k;$$

тогда, учитывая (1), (2), имеем:

$$\pi_i = \frac{1}{\lambda} \sum_{j=1}^{n} t_{ij} \cdot \omega_j , \qquad (4)$$

$$\omega_j = \frac{1}{\mu} \sum_{i=1}^{m} t_{ij} \cdot \pi_i . \qquad (5)$$

Равенства (4), (5) отражают согласование оценок π_i, $i = 1,2, \ldots, m$, с оценками ω_j, $j = 1,2, \ldots, n$, которое в том, что π_i получаются через ω_j так же, как ω_j – через π_i.

Для выяснения того, как происходит согласование, рассмотрим, что содержательно выражают формулы (1) – (3).

Формулы (3) показывают, что целесообразно принять равными начальные приближения к оценкам всех строк и столбцов, ибо заранее неизвестно, какие строки или столбцы предпочтительнее. За счет различия значений t_{ij}, $i = 1,2, \ldots, m$; $j = 1,2, \ldots, n$, происходит дифференциация значений $\pi_i^{(k)}$, $\omega_j^{(k)}$ при $k \ge 1$, как видно из формул (1),(2).

Роль множителей λ_k, μ_k – ограничивать значения $\pi_i^{(k+1)}$, $\omega_j^{(k+1)}$ сверху так, чтобы суммы вида $\sum_{j=1}^{n} \omega_j^{(k+1)} + \sum_{i=1}^{m} \pi_i^{(k+1)}$ не росли до бесконечности с ростом k, а были ограничены сверху конечным числом, например $m+n$.

Итак, "внешнее" ограничение суммы $\sum_{i=1}^{m} \pi_i^{(k+1)} + \sum_{j=1}^{n} \omega_j^{(k+1)}$

осуществляется парой нормирующих множителей (λ_k, μ_k), а "внутреннее" распределение этой суммы на значения $\pi_i^{(k+1)}, \omega_j^{(k+1)}$ зависит от $m \cdot n$ значений t_{ij} , $i = 1,2, \ldots, m$; $j = 1, 2, \ldots, n$.

Согласование же проявляется в том, что "внешне" пара (λ_k, μ_k) имеет пределом пару (λ , μ), а "внутренне" $m + n$ значений $\pi_i^{(k+1)}$, $\omega_j^{(k+1)}$ имеют пределом $m+n$ значений π_i, ω_j , $i = 1,2, \ldots, m$, $j = 1,2, \ldots, n$.

Такие качественные рассуждения позволят нам дать идейную сторону центрирования качельных процедур.

2. Центрирование

Как отмечено выше, перераспределение суммы оценок строк и столбцов по отдельным строкам и столбцам определяется совокупностью $m \cdot n$ значений t_{ij} признака x_j на объекте S_i , $i = 1, 2, \ldots, m$, $j = 1,2, \ldots, n$. Но само значение t_{ij} есть, другими словами, отклонение выраженности x_j на S_i от нулевой, что, формализованно, можно выразить так:

$$\pi_i^{(k+1)} = \frac{1}{\lambda_k} \sum_{j=1}^{n} (t_{ij} - 0) \cdot \omega_j^{(k)} , \qquad (1')$$

$$\omega_j^{(k+1)} = \frac{1}{\mu_k} \sum_{i=1}^{m} (t_{ij} - 0) \cdot \pi_i^{(k)} \qquad (2')$$

Так, центрирование, как отклонение t_{ij} от нулевых центрирующих значений, можно условно приписать даже первоначальному варианту качельной процедуры, данному в [9], а затем и модифицированным вариантам [15].

В работах [28,34] в качестве центрирующих значений выбраны средние арифметические из значений t_{ij} по строкам и по столбцам, то есть значения $\frac{1}{n} \sum_{j=1}^{n} t_{ij}$, $\frac{1}{m} \sum_{i=1}^{m} t_{ij}$. В этом случае центрирование вида $t_{ij} - \frac{1}{n} \sum_{j=1}^{n} t_{ij}$ или вида $t_{ij} - \frac{1}{m} \sum_{i=1}^{m} t_{ij}$ давало бы для некоторых t_{ij} отрицательные значения. Чтобы добиться положительности, эти выражения берутся в квадрате $(t_{ij} - \frac{1}{n} \sum_{j=1}^{n} t_{ij})^2$ и $(t_{ij} - \frac{1}{m} \sum_{i=1}^{m} t_{ij})^2$, чем

достигается аналогия с выражениями для дисперсии.

Теперь, когда идея центрирования объяснена на простейших примерах, дадим общие формулы:

$$\pi_i^{(k+1)} = \frac{I}{\Lambda_k} \sum_{j=1}^{n} \varphi\left[(t_{ij} - \alpha_{ij});\ \omega_j^{(k)}\right], \qquad (6)$$

$$\omega_j^{(k+1)} = \frac{I}{\mu_k} \sum_{i=1}^{m} \psi\left[(t_{ij} - \beta_{ij}),\ \pi_i^{(k)}\right], \qquad (7)$$

где α_{ij}, β_{ij} – центрирующие значения, φ , ψ – некоторые неотрицательные функции. Несколько примеров конкретного вида функций φ , ψ приведено в статье [29]. Нами ниже будет рассмотрен более частный, важный случай, предложенный в работах [28,34].

3. Реализация процедур

В данной реализации центрированных качельных процедур элементы t_{ij} таблицы T свободны от условия неотрицательности. Кроме того, предполагается, что в таблице T отсутствуют строки и столбцы, заполненные целиком одним и тем же значением.

Таблица T преобразуется в таблицу U , состоящую из неотрицательных элементов u_{ij} , путем следующей нормировки ее столбцов:

$$u_{ij} = \frac{t_{ij} - \min\limits_{1 \le \alpha \le m} \{t_{\alpha j}\}}{\max\limits_{1 \le \alpha \le m} \{t_{\alpha j}\} - \min\limits_{1 \le \alpha \le m} t_{\alpha j}}, \quad i = 1,2,\ldots,m \ ; \ j = 1,2,\ldots,n ,$$

откуда видно, что $0 \le u_{ij} \le 1$. Оценки строк и столбцов таблицы T получаются по таблице U , в которой столбцы и строки становятся сопоставимыми между собой за счет неравенства $0 \le u_{ij} \le 1$.

Имеется два варианта подсчета оценок строк и столбцов по таблице U . Один задается формулами:

$$\pi_i^{(k+1)} = \left\{ \frac{\sum\limits_{j=1}^{n} \left[\omega_j^{(k)} \cdot (u_{ij} - \bar{x}_j)\right]^2}{\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} \left[\omega_j^{(k)} (u_{ij} - \bar{x}_j)\right]^2} \right\}^{1/2}, \qquad (8)$$

$$\omega_j^{(k+1)} = \left\{ \frac{\sum_{i=1}^{m} \left[\pi_i^{(k)} (u_{ij} - \bar{s}_i) \right]^2}{\sum_{j=1}^{n} \sum_{i=1}^{m} \left[\pi_i^{(k)} (u_{ij} - \bar{s}_i) \right]^2} \right\}^{1/2}, \qquad (9)$$

где $\bar{x}_j = \frac{I}{m} \sum_{i=1}^{m} u_{ij}$, $\bar{s}_i = \frac{I}{n} \sum_{j=1}^{n} u_{ij}$, k – номер итерации, причем $\pi_i^{(0)} = \omega_j^{(0)} = I$, $i = I,2,\ldots,m$; $j = I,2,\ldots,n$.

В этом варианте, как видно из формул (9), (8), оценки строк растут с ростом отклонений компонент строк от компонент "средней" строки таблицы T, а оценки столбцов тем выше, чем выше покомпонентные отклонения столбцов от "среднего" столбца.

Другой вариант центрированной качельной процедуры получится, если в формулах (8), (9) провести перестановку величин $\bar{x}_j$ и $\bar{s}_i$:

$$\pi_i^{(k+1)} = \left\{ \frac{\sum_{j=1}^{n} \left[\omega_j^{(k)} (u_{ij} - \bar{s}_i) \right]^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} \left[\omega_j^{(k)} (u_{ij} - \bar{s}_i) \right]^2} \right\}^{1/2}, \qquad (10)$$

$$\omega_j^{(k+1)} = \left\{ \frac{\sum_{i=1}^{m} \left[\pi_i^{(k)} (u_{ij} - \bar{x}_j) \right]^2}{\sum_{j=1}^{n} \sum_{i=1}^{m} \left[\pi_i^{(k)} (u_{ij} - \bar{x}_j) \right]^2} \right\}^{1/2} \qquad (11)$$

В этом варианте оценки строк и столбцов тем выше, чем выше разброс значений компонент строки или столбца самих по себе, то есть по отношению к среднему арифметическому значений компонент оцениваемой строки или столбца.

Рассмотренные формулы (8), (9) и (10), (11) представляют собой частный случай формул (6), (7). Оценки, получаемые по программам на ЭВМ, описанным в статье [24], суть оценки, задаваемые формулами (8), (9) и (10), (11).

4. Вопросы применения

Как указывалось выше, формулы (8) служат для оценки отклонений строк таблицы от "средней" строки (состоящей из средних арифметических значений по каждому столбцу), а формулы (9) – для оценки отклонений столбцов от "среднего" столбца (в котором каждая компонента – среднее арифметическое из значений соответствующей ей строки таблицы T).

Поэтому величины π_i , ω_j , получаемые по формулам (8),(9), дают возможность ранжировать строки и столбцы таблицы T по их типичности: чем типичнее для таблицы T строка S_i или столбец x_j , тем меньше значения π_i по отношению к $\pi_1, \pi_2, \ldots, \pi_m$ и ω_j по отношению к $\pi_1, \pi_2, \ldots, \pi_m$. Самые большие значения оценок π_i и ω_j соответствуют самым нетипичным строкам и столбцам таблицы T . Типичность, в данном случае, понимается, как малость отклонений от "средних" строки и столбца.

Оценки π_i, ω_j, извлекаемые из формул (I0), (II), отражают величину разброса значений строк и столбцов самих по себе, то есть суммируются отклонения компонент оцениваемой строки (или столбца) от среднего арифметического ее компонент.

Чем выше величины разброса, тем больше значения оценок π_i , ω_j . Поэтому представляется целесообразным применять эти оценки для подразделения столбцов таблицы на группы столбцов, состоящие из столбцов с близкими значениями ω_j. Группа с наименьшими значениями ω_j суть самые близкие к отождествляющим столбцы, группа с наибольшими значениями ω_j – самые контрастные столбцы. Из строк с близкими значениями π_i также можно образовать группы. Эти группы характеризуют стабильность строк: чем меньше значения π_i в группе, тем стабильнее по своим значениям объекты в этой группе. Самые высокие значения π_i имеют объекты "с крайностями". Исходя из подразделения объектов на группы с близкими значениями π_i , можно попытаться устанавливать оптимальное число градаций столбцов.

Не останавливаясь более на вопросах применения центрированных качельных процедур для одной таблицы, перейдем к изложению способа их употребления для случая двух таблиц, что сво-

дится к алгоритмам распознавания по ЦКП.

5. Распознавание

Задача распознавания формулируется таким образом.

Пусть даны две таблицы эталонов – T_1, T_2 и таблица проб – $T_П$; все три таблицы имеют одно и то же число столбцов – n. Требуется для каждой строки $S_П$ из $T_П$ указать, к какой из таблиц T_1 и T_2 строка $S_П$ тяготеет в большей степени.

Прежде, чем воспользоваться процедурами ЦКП, выполняются следующие подготовительные операции. Составляется таблица T, в которую входят таблицы T_1, T_2, $T_П$, записанные одна под другой, причем так, что строки таблицы T_1 размещаются выше, а строки из $T_П$ ниже, чем строки таблицы T_2.

Так таблицы T_1, T_2, $T_П$ записываются одна под другой в указанном порядке и образуется таблица T. В таблице T строки нумеруются так:

строка $б$ получает номер

i_1, если $б$ – строка из T_1 с номером i_1;

$i_2 + m_1$, если $б$ – из T_2 с номером i_2;

$i_3 + m_1 + m_2$, если $б$ из $T_П$ с номером i_3;

здесь m_1 – число строк в T_1, m_2 – число строк в T_2.

Далее проводится нормировка значений таблицы T по формуле:

$$u_{ij} = \frac{t_{ij} - t_{jmin}}{t_{jmax} - t_{jmin}},$$

где t_{jmin} – минимальный, а t_{jmax} – максимальный элементы j – го столбца пары таблиц (T_1, T_2). Получившиеся в результате нормировки таблицы обозначаются так же: T_1, T_2, $T_П$. Заметим, что в результате этой нормировки элементы таблиц T_1, T_2 попадают в интервал $[0,1]$, а элементы $T_П$ могут быть меньше 0 или больше 1.

Затем таблицы T_1, T_2 подлежат обработке с помощью ЦКП по формулам (8), (9), то есть вычисляются оценки строк π_i^{ℓ} и оценки столбцов ω_j^{ℓ}, $\ell = 1,2$, удовлетворяющие равенствам:

$$\pi_i^{\ell} = \left\{ \frac{\sum_{j=1}^{n} [\omega_j^{\ell} (u_{ij}^{\ell} - \bar{x}_j^{\ell})]^2}{\sum_{i=1}^{m_\ell} \sum_{j=1}^{n} [\omega_j^{\ell} (u_{ij}^{\ell} - \bar{x}_j^{\ell})]^2} \right\}^{1/2}, \quad (I2)$$

$$\omega_j^{\ell} = \left\{ \frac{\sum_{i=1}^{m_\ell} [\pi_i^{\ell} (u_{ij}^{\ell} - \bar{s}_i)]^2}{\sum_{j=1}^{n} \sum_{i=1}^{m_\ell} [\pi_i^{\ell} (u_{ij}^{\ell} - \bar{s}_i)]^2} \right\}^{1/2}, \quad (I3)$$

где $\bar{x}_j^{\ell}$ – среднее арифметическое значений j – го столбца, а $\bar{s}_i^{\ell}$ – среднее арифметическое значений i – й строчки таблицы T_ℓ, m_ℓ – число строк таблицы T_ℓ, $\ell = 1,2$.

Для распознавания употребляются величины:

$$\bar{R}(S) = \frac{\bar{z}(1,S)}{\bar{z}(2,S)}, \quad \text{где} \quad (I4)$$

$$\bar{z}(\ell, S) = \sum_{j=1}^{n} [(t_j - \bar{x}_j^{\ell}) \omega_j^{\ell}]^2, \quad \ell = 1,2, \quad (I5)$$

характеризующие произвольную строку $S = (t_1, t_2, \ldots, t_n)$; величина $\bar{R}(S)$ имеет смысл при $\bar{z}(2,S) \neq 0$ и отражает удаленность S от T_1.

Поясним смысл употребления отношения (I4) для распознавания строки S на принадлежность к таблицам T_1 или T_2. Из формул (I2), (I5) видно, что величины $\bar{z}(\ell, S_i)$ (для строк S_i таблицы T_ℓ) пропорциональны величинам $(\pi_i^{\ell})^2$, то есть ранжируют строки таблицы T_ℓ по их типичности. Если считать, что и все другие строки S не принадлежащие T_ℓ, также таковы, что величины $\bar{z}(\ell, S)$ отражают их типичность по отношению к строкам таблицы T_ℓ, то, все-таки, эта типичность подсчитывается в единицах таблицы T_ℓ. Когда же берется отношение (I4), то в нем типичность выражена в единицах, учитываю-

щих обе таблицы как T_1, так и T_2. Если строка S_1 – типичная строка таблицы T_1, причем нетипична для таблицы T_2, то $\bar{R}(S_1) < I$, но для строки S_2, типичной таблице T_2 и нетипичной таблице T_1, имеет место неравенство $\bar{R}(S_2) > I$. Эти соображения позволяют сформулировать следующий способ употребления отношения (I4) для распознавания.

Производится подсчет чисел

$$\bar{\alpha}_1 = \max_{S \in T_1} \bar{R}(S), \quad \bar{\alpha}_2 = \min_{S \in T_2} \bar{R}(S).$$

Возможны 2 случая: либо $\bar{\alpha}_1 < \bar{\alpha}_2$, либо $\bar{\alpha}_1 \geq \bar{\alpha}_2$. В первом случае, когда $\bar{\alpha}_1 < \bar{\alpha}_2$, принимаются следующие решения:

а) если $\bar{R}(S_n) \leq \bar{\alpha}_1$, то S_n относится к T_1;
б) если $\bar{R}(S_n) \geq \bar{\alpha}_2$, то S_n относится к T_2;
в) если $\bar{\alpha}_1 < \bar{R}(S_n) < \bar{\alpha}_2$, то S_n относится к зоне отказа от решения; здесь $S_n \in T_n$.

Во втором случае, когда $\bar{\alpha}_1 \geq \bar{\alpha}_2$, решение таково:

а) если $\bar{R}(S_n) < \bar{\alpha}_2$, то S_n относится к T_1;
б) если $\bar{R}(S_n) > \bar{\alpha}_1$, то S_n относится к T_2;
в) если $\bar{\alpha}_2 \leq \bar{R}(S_n) \leq \bar{\alpha}_1$, то S_n попадает в зону неопределенности.

В изложенном выше очерке центрированных качельных процедур метода согласованных оценок охарактеризован ряд важных элементов этих процедур: согласование, центрирование, реализация, а также вопросы применения и алгоритм распознавания с помощью ЦКП. Приведенные характеристики этих элементов могут послужить стимулом для дальнейшей разработки в этой области, так как в ряде моментов усматривается возможность такой разработки.

§5 СОГЛАСОВАННЫЕ ОЦЕНКИ ДЛЯ НЕОДНОРОДНЫХ ВЫБОРОК

Дана таблица данных X размера $m \times n$; строка – объект, столбец – признак; x_{ik} есть значение k -го признака на i -м объекте. Пусть объекты (строки) x_i разбиты на K классов ($1 \leqslant K \leqslant m$). Таблице X соответствует n – мерная обучающая выборка объема m , обозначаемая тем же символом X . Надо спроектировать выборку X на вектор $\alpha = (\alpha_1, \ldots, \alpha_n)$ так, чтобы проекции $u_i = \alpha_1 x_{i_1} + \ldots + \alpha x_{in}$ имели относительно небольшую дисперсию внутри каждого класса и в то же время хорошо разделяли бы объекты из разных классов. Иными словами, надо свести n – мерную задачу классификации к одномерной. Прикладная цель: распознавание (типизация) новых объектов, не принадлежащих обучающей выборке X , с помощью некоторого решающего правила с параметрами, зависящими от вектора α .

Одну из конкретизаций данной задачи представляет собой дискриминантный анализ [43]. Однако вычислительный алгоритм дискриминантного анализа довольно сложен даже для двух классов; если же задано более двух классов, возникают, помимо известных вычислительных, также и дополнительные диагностические трудности.

Ниже предлагается другой подход к проблеме, более простой в вычислительном отношении, являющийся, с одной стороны, вариантом известной методики канонических корреляций, а с другой – обобщением метода главных компонент (МГК), а также – для неотрицательных данных – метода согласованных оценок [II , I9] .

I. Канонические корреляции

Начнем с изложения одной классической задачи (Хотеллинг, I936), [3I] .

Даны две таблицы данных: X ($m \times n$) и Y ($m \times s$). Объекты $x_{i\cdot} \in X$ занумерованы также, как и объекты $y_{i\cdot} \in Y$; т.е. между строками X и Y установлено взаимно-однозначное соответствие: $x_{i\cdot} \sim y_{i\cdot}$.

Возьмем линейные комбинации:

$$u_i = \alpha_1 x_{i1} + \ldots + \alpha_n x_{in}; v_i = \beta_1 y_{i1} + \ldots + \beta_s y_{is}$$

$$(\ i = 1, \ldots, m\ ) \qquad (1),(2)$$

Введем меру взаимосвязи между наборами и :

$$R = \sum u_i v_i = \alpha^T X^T Y \cdot \beta \qquad (3)$$

или в развернутом виде:

$$R = \begin{cases} \alpha_1 \beta_1 w_{11} + \ldots + \alpha_1 \beta_s w_{1s} \\ \ldots\ldots\ldots\ldots \\ \alpha_n \beta_1 w_{n1} + \ldots + \alpha_n \beta_s w_{ns}\ , \end{cases} \qquad (4)$$

где обозначено: $w_{pq} = \sum x_{ip} y_{iq}$

Строки x_i , y_i можно считать реализациями случайных векторов. Если, кроме того, центрировать и нормировать величины x_i , y_i , u_i , v_i , то функционал R будет являться выборочным коэффициентом корреляции. В дальнейшем мы будем избегать вероятностных трактовок, сохраняя в ряде случаев статистическую терминологию.

Найдем векторы $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, $\beta = (\beta_1, \ldots, \beta_n)^T$, обращающие R в максимум, удовлетворяющие при этом условию нормировки:

$$\alpha^T X^T X \alpha = I\ , \quad \beta^T Y^T Y \beta = I \qquad (5)$$

(таким образом, мы требуем, чтобы дисперсии величин u_i и v_i равнялись единице).

Введя множители Лагранжа λ , μ , приходим к задаче на безусловный экстемум функционала:

$$P = \alpha^T X^T Y \beta - \frac{\lambda}{2} (\alpha^T X^T X \alpha - I) - \frac{\mu}{2} (\beta^T Y^T Y \beta - I) \qquad (6)$$

Необходимые условия экстремума (6):

$$\left.\begin{aligned} X^T Y \beta - \lambda X^T X \alpha &= 0 \\ -\mu Y^T Y \beta + Y^T X \alpha &= 0 \end{aligned}\right\} \qquad \begin{matrix} (7) \\ (8) \end{matrix}$$

Из (7) , (8), с учетом (5) , следует: $\lambda = \alpha^T X^T Y \beta = \beta^T Y^T X \alpha = \mu$. Общее значение λ , μ обозначим через ν . Домножим (7) слева на $Y^T X$ $(X^T X)$ и сложим с (8), умноженным на ν ; аналогично домножим (8) слева на $X^T Y (Y^T Y)^{-1}$ и сложим с (7) , умноженным на ν . В результате будем иметь:

$$(Y^T X (X^T X)^{-1} X^T Y - \nu^2 Y^T Y)\beta = 0\ , \qquad (9)$$

$$(X^{T}Y \ (Y^{T}Y)^{-1}Y^{T}X - \nu^{2}X^{T}X)\alpha = 0 \qquad (10)$$

Ненулевые собственные числа системы (9),(10) по определению суть <u>канонические корреляции</u>. Число их равно рангу матрицы $X^{T}Y$; в модели факторного анализа это есть эффективное число общих факторов [3I].

Пусть $\nu_1^2, \ldots, \nu_n^2$ и $\nu_1^2, \ldots, \nu_s^2$ суть наборы собственных значений (включая нулевые) уравнений (9) и (10) (ненулевые части их совпадают). Соответствующие наборы собственных векторов: $(\alpha^1, \ldots, \alpha^n)$, $(\beta^1, \ldots, \beta^s)$. Линейные функции $X\alpha^1, \ldots, X\alpha^n$, а также $Y\beta^1, \ldots, Y\beta^s$, называются <u>каноническими величинами</u>. Известно ([3I]), что величины $X\alpha^i$ ($i = 1, \ldots, n$), равно как и $Y\beta^j$ ($j = 1, \ldots, s$), попарно некоррелированы и имеют единичное стандартное отклонение; ковариация $X\alpha^i$ и $Y\beta^j$ ($i \neq j$) равна нулю; ковариация $X\alpha^i$ и $Y\beta^i$ равна ν_i. Эти свойства канонических величин делают их полезными для приложений.

Эта теоретическая схема является довольно общей. В частности, как нетрудно показать, при $s = 1$ она превращается в схему линейной регрессии. Другие модификации даются ниже.

2. Связь канонических корреляций с МГК

Сохраняя прежнюю постановку задачи (§2), введем, вместо (5), новую нормировку:

$$\alpha^{T}\alpha = 1 , \qquad \beta^{T}\beta = 1 , \qquad (11)$$

что приводит к другому, сравнительно с (6), функционалу:

$$Q = \alpha^{T}X^{T}Y\beta - \frac{\lambda}{2}\alpha^{T}\alpha - \frac{\mu}{2}\beta^{T}\beta \qquad (12)$$

Условия экстремальности:

$$X^{T}Y\beta = \lambda\alpha \qquad (13)$$

$$Y^{T}X\alpha = \mu\beta \qquad (14)$$

Из (13),(14) следует: ($\nu = \lambda\mu$).

$$X^{T}YY^{T}X\alpha = \nu\alpha , \qquad (15)$$

$$Y^{T}XX^{T}Y\beta = \nu\beta \qquad (16)$$

Перепишем (13) - (14), полагая $Y = I$ $(m \times n)$ (I - еди-

ничная матрица):

$$X^T\beta = \lambda\alpha \qquad (17)$$

$$X\alpha = \mu\beta \qquad (18)$$

$$X^TX\alpha = \nu\alpha \qquad (19)$$

$$X X^T\beta = \nu\beta \qquad (20)$$

Система вида (19), (20) возникает в методе главных компонент (МГК). Для центрированных данных (когда среднее по столбцам равно нулю) главный вектор α определяет направление наибольшего разброса выборки X, а соответствующий ему вектор есть первая главная компонента (проекция выборки на главное направление). Если данные $x_{ik} \in X$ неотрицательные, уравнения (17), (18) совпадают с основными уравнениями так называемого метода согласованных оценок, который, по существу, является вычислительным вариантом МГК, но ввиду неотрицательности данных допускает специфическую интерпретацию результатов. В частности, компоненты векторов α, β можно рассматривать как весовые коэффициенты признаков и объектов [19, 29].

Таким образом, схема канонических корреляций охватывает, как частные случаи, основные уравнения МГК и метода согласованных оценок.

Вернемся к нормировке (5). Пусть $Y = I$ ($m \times m$). Тогда условия экстремальности (7), (8) примут вид:

$$X^T\beta = \lambda X^TX\alpha \qquad (21)$$

$$X\alpha = \mu\beta \qquad (22)$$

Покажем, что $\lambda = \mu = I$; (возможно также $\lambda = \mu = -I$, но в таком случае, домножив α или β на $-I$, мы опять приходим к задаче с положительными значениями: $\lambda = \mu = I$). Из (22) следует:

$$\mu^2\beta^T\beta = \alpha^TX^TX\alpha = I; \qquad \mu^2\beta^T\beta = I \qquad (23)$$

Так как $Y^TY = I$, из нормировки (5) следует: $\beta^T\beta = I$. Поэтому, в силу (23), имеем: $\mu^2 = I$. Далее, в силу (21),(22), справедливы равенства:

$$\mu X^T\beta = \lambda\mu X^TX\alpha; \quad X^TX\alpha = \mu X^T\beta, \qquad (24)$$

из которых следует: $\lambda\mu = I$. Итак, $\lambda = \mu = I$. Система (21), (22) приобретает вид:

$$X^T\beta = X^T X\alpha \qquad (25)$$
$$X\alpha = \beta$$

или:

$$[X^T X]^{-1} X^T \beta = \alpha \qquad (26)$$
$$X\alpha = \beta$$

Система (26) определяет другой вариант метода согласованных оценок. Если данные $x_{ik} \in X$ – неотрицательные, то, при выполнении тех же весьма нежестких условий, что и в старом варианте метода, посредством аналогичной вычислительной (итерационной) процедуры, можно найти положительные векторы α , β , компоненты которых играют роль весовых коэффициентов. Отметим, что нормировка (5) может оказаться более удачной по сравнению с (II), ибо она может компенсировать неудачный выбор масштабов (единиц измерения) признаков X , в то время как условие (II) прямой связи с масштабами не имеет.

Примечание. Если $Y = I$, мера взаимосвязи (4) принимает вид:

$$R = \sum\sum \alpha_k \beta_i x_{ik} \qquad (27)$$

Такая запись наводит на несколько иную формулировку задачи. Домножить строки x_i таблицы X на такие коэффициенты β_i , и одновременно домножить столбцы $x_{\cdot k}$ на коэффициенты α_k , чтобы обратить (27) в максимум. Для неотрицательных данных такое домножение равносильно приписыванию объектам и признакам индивидуальных весовых коэффициентов.

3. Матрица сходства и обобщение МГК

Рассмотрим неоднородную выборку X , разбитую на k классов ($k \leq m$). Мы не потеряем в общности, если предположим, что объекты, принадлежащие одному и тому же i -му классу, образуют в таблице X сплошной массив X^i. Для иллюстрации примем, что $k = 3$. Зададим на множестве пар объектов простейшее отношение близости, или сходства: каждой паре представителей одного и того же класса ставится в соответствие I , а каждой паре объектов из разных классов – 0. Это отношение определяет матрицу сходства (размера $m \times m$), имеющую блочно-диагональную структу-

ру, в соответствии с блочной структурой таблицы X :

$$X = \begin{vmatrix} X^1 \\ X^2 \\ X^3 \end{vmatrix} \quad \begin{vmatrix} Y_{11} & & \\ & Y_{22} & \\ & & Y_{33} \end{vmatrix} = Y \qquad (28)$$

Заштрихованные блоки заполнены единицами, прочие – нулями; классу X с m_i элементами сопоставлен "единичный" блок Y_{ii} размера ($m_i \times m_i$).

Решим задачу на максимум:

$$R = \alpha^T X^T Y \beta \qquad (29)$$

$$\alpha^T \alpha = I, \qquad \beta^T \beta = I , \qquad (30)$$

где X , Y задаются схемой (28).

Вектор β также запишем в блочном виде: $\beta^T = (\beta^1, \beta^2, \beta^3)$, где $\beta^i = (\beta_1^i, \ldots , \beta_{m_i}^i)$.

Условия экстремальности (I3),(I4):

$$| X^{1T} Y_{11} \; | \; X^{2T} Y_{22} \; | \; X^{3T} Y_{33} | \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^3 \end{pmatrix} = \lambda \alpha \qquad (31)$$

$$\begin{vmatrix} Y_{11} & & \\ & Y_{22} & \\ & & Y_{33} \end{vmatrix} \begin{vmatrix} X^1 \\ X^2 \\ X^3 \end{vmatrix} \alpha = \mu \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^3 \end{pmatrix} \qquad (32)$$

Система (32) очевидным образом распадается на три подсистемы; выпишем, например, первую из них, обозначив через u'_k сумму элементов k -го столбца подматрицы X^1 :

$$\begin{aligned} u'_1 \alpha_1 + \ldots + u'_n \alpha_n &= \mu \beta_1^1 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \qquad (33) \\ u'_1 \alpha_1 + \ldots + u'_n \alpha_n &= \mu \beta_{m_1}^1 \end{aligned}$$

Отсюда следует, что все координаты подвектора β^1 равны между собой. Очевидно, то же самое справедливо и для β^2, β^3 .

Итак:

$$\begin{aligned} \beta_1^1 = \ldots = \beta_{m_1}^1 &= \hat{\beta}_1 \\ \beta_1^2 = \ldots = \beta_{m_2}^2 &= \hat{\beta}_2 \qquad (34) \\ \beta_1^3 = \ldots = \beta_{m_3}^3 &= \hat{\beta}_3 \end{aligned}$$

Здесь $\hat{\beta}_i$ – общее значение координаты подвектора β^i . Поэтому подсистема (33) вырождается в одно уравнение. Если, кроме

того, упростить (3I) с учетом (34), система (3I),(32) примет вид:

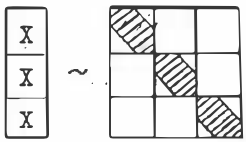
$$m_1 u_\kappa^1 \hat{\beta}_1 + m_2 u_\kappa^2 \hat{\beta}_2 + m_3 u_\kappa^3 \hat{\beta}_3 = \lambda \alpha_\kappa \qquad (35)$$

$$m_1^i \alpha_1 + \ldots + u_n^i \alpha_n = \mu \hat{\beta}_i \qquad (36)$$

Заметим, что систему (35),(36) можно получить непосредственно из уравнений МСО (I7),(I8), если дополнительно потребовать (34). Таким образом, константы β_i (весовые коэффициенты) характеризуют классовую принадлежность объектов. Можно предположить, что и для других матриц сходства, задаваемых с помощью какой-либо естественной метрики, близость коэффициентов β_i, получаемых из уравнений вида (I3),(I4), будет свидетельствовать о близости соответствующих объектов по некоторому характеристическому свойству (быть может, неявному), которое можно положить в основу одномерной классификации (и следовательно свести задачу распознавания к одномерной).

Уравнения (35),(36) можно рассматривать как систему МГК относительно величин u_κ^i (от которых легко перейти к средним значениям признаков по каждому классу: $m_i^{-1} u_\kappa^i$); сравните с уравнениями (I9),(20). В этой связи обычная задача МГК ((I9),(20)) соответствует случаю, когда число классов равно числу объектов ($\kappa^i = m$), т.е. когда каждый объект является единственным представителем своего класса.

Рассмотрим еще один пример матрицы сходства (заштрихованные части блоков заполнены единицами; остальные элементы – нули):



$$X = \sim = Y \qquad (37)$$

Схема (37) выражает случай, когда каждый объект похож на некоторое количество (меньшее, чем m_i) соседних элементов своего класса X^i. Из-за "краевого эффекта", состоящего в том, что крайние объекты в классе имеют, по (37), меньше сходных соседей, по сравнению с внутренними объектами, характеристические константы β_i (веса), будучи, как правило, близкими для внутренних объектов, для крайних (т.е. "нетипичных") объектов класса будут относительно меньшими.

В общем случае элементы матрицы сходства $y_{ij} \in Y$ суть значения некоторой меры сходства объектов x_i и x_j выборки $X(i,j = I, \ldots, m)$; можно потребовать, чтобы эта мера принимала значения из интервала (0,I).

Естественно предполагать, что, при удачном задании классов и меры сходства, диагональные блоки Y (являющиеся внутриклассовыми матрицами сходства), будут заполнены, в основном, величинами, близкими к I, в то время как остальные блоки - величинами, близкими к 0.

В примере с матрицей (37) при умножении X^T на Y (если решать систему (I3),(I4)), фактически происходит "скользящее усреднение" (сглаживание) по столбцам. Можно ожидать, что аналогичный эффект будет проявляться и в общем случае (кроме, быть может, очень искусственных примеров). Поэтому весьма вероятно, что классификация, выполненная с помощью характеристических констант β_i , будет отчасти гасить ошибки первичной классификации.

В качестве меры сходства можно взять, например, функционал:

$$r(x_i, x_j) = \frac{I}{n} \sum_{\kappa=1}^{n} exp - \left\{ -\rho\alpha_\kappa (\beta_i x_i - \beta_j x_j)^2 \right\} \quad (38)$$

Здесь α_κ , β_i , β_j - некоторые начальные значения весовых коэффициентов; ρ - нормирующий множитель, подбираемый экспериментально..

Расстояние от произвольного объекта $z = (z_1, \ldots, z_n)^T$ до класса X^κ (необходимое для диагностики) определяется естественным образом как среднее (по всем $x_i \in X^\kappa$) значение величин $r(z, x_i)$.

§6 ВЫЧИСЛЕНИЕ ВЕСОВЫХ КОЭФФИЦИЕНТОВ И СОГЛАСОВАННЫХ ОЦЕНОК

Дано некоторое множество объектов, характеризуемых набором признаков. Имеется таблица X размера ($m \times n$), составленная из оценок признаков для всех объектов. Элемент $x_{ik} \in X$ есть значение k-го признака на i-м объекте. Таблица X представляет собой либо сводку экспериментальных данных, либо сводку экспертных оценок (в последнем случае "признак" есть мнение эксперта). Предполагается, что таблица обладает достаточной однородностью, так что усреднение оценок x_{ik} как по строкам (объектам), так и по столбцам (признакам) имеет физический смысл. Требуется получить среднее взвешенное значение по каждой строке и по каждому столбцу. Для этого надо знать весовые коэффициенты (веса) признаков и объектов. Обычно веса вводятся априорно, на основе каких-либо внешних соображений. Мы рассмотрим серию способов нахождения весовых коэффициентов на основе имеющихся данных (т.е. по таблице X), снижающих влияние субъективного фактора.

Примеры задач:

1. Экзаменационная комиссия (или судейская коллегия на соревнованиях); x_{ik} есть оценка в баллах, поставленная k-м экзаменатором (судьей) i-му студенту (спортсмену). Требуется вывести средний балл.
2. Группа экспертов, оценивающих затраты на различные варианты ведения разведки (или разработки) месторождения; x_{ik} есть прогнозируемая стоимость i-й разведки (или разработки) по мнению k-го эксперта. Получить согласованные оценки.
3. Эффективность промышленных предприятий, выпускающих несколько разных видов продукции; x_{ik} есть процент выполнения плана i-м предприятием по k-му виду продукции. Получить сводную оценку эффективности.
4. Геологические данные; x_{ik} есть оценка рудоносности i-го района по k-му способу поиска. Упорядочить исследуемые районы по степени перспективности.

На неформальном уровне очевидно, что веса объектов и признаков должны быть связаны между собой; согласование весов состо-

ит в учете этой связи посредством аналитических выражений.

Обозначения:

α – весовой коэффициент (вес) i-го объекта ($i = I,\dots,m$) $(0 \leqslant \alpha_i \leqslant I)$;

β – весовой коэффициент (вес) κ-го признака ($\kappa = I,\dots,n$) $(0 \leqslant \beta_\kappa \leqslant I)$;

$\bar{x}_{\cdot\kappa} = \sum \alpha_i x_{i\kappa}$ – среднее взвешенное значение κ-го признака;

$\bar{x}_{i\cdot} = \sum \beta_n x_{i\kappa}$ – среднее взвешенное значение i-го объекта;

$u_{i\kappa} = x_{i\kappa} - \bar{x}_{\cdot\kappa}$ – отклонение $x_{i\kappa}$ от среднего по объектам;

$v_{i\kappa} = x_{i\kappa} - \bar{x}_{i\cdot}$ – отклонение $x_{i\kappa}$ от среднего по признакам;

$w_{i\kappa}$ – общее наименование для $u_{i\kappa}$, $v_{i\kappa}$ (т.е. $w_{i\kappa} = u_{i\kappa}$ или $w_{i\kappa} = v_{i\kappa}$);

$\alpha = (\alpha_1, \dots, \alpha_m)^T$ – весовой вектор объектов $\sum \alpha_i = I$;

$\beta = (\beta_1, \dots, \beta_n)^T$ – весовой вектор признаков $\sum \beta_\kappa = I$;

Типичные формулы для весовых коэффициентов:

$$(\alpha)\begin{cases} \mu\alpha_i = \sum_\kappa \beta_\kappa \cdot x_{i\kappa} + h \\ \mu\alpha_i = \left(\sum_\kappa \beta_\kappa \cdot w_\kappa^2 \right)^\tau \\ \mu\alpha_i = \exp\left\{ -\gamma \sum_\kappa \beta_\kappa \cdot x_{i\kappa} + h \right\} \\ \mu\alpha_i = \exp\left\{ -\gamma \left(\sum_\kappa \beta_\kappa w_{i\kappa}^2 \right)^\tau \right\} \end{cases} \qquad (\beta)\begin{cases} \nu\beta_\kappa = \sum_i \alpha_i x_{i\kappa} + h \\ \nu\beta_\kappa = \left(\sum_i \alpha_i^2 w_{i\kappa}^2 \right)^\tau \\ \nu\beta_\kappa = \exp\left\{ -\gamma \sum_i \alpha_i x_{i\kappa} + h \right\} \\ \nu\beta_\kappa = \exp\left\{ -\gamma \sum_i (\alpha_i w_{i\kappa}^2)^\tau \right\} \end{cases}$$

Здесь μ, ν – нормирующие множители; h, τ, γ – параметры, подбираемые экспериментально. Константа h задает нуль-пункт отсчета весов.

Комбинируя формулы из (α) и (β), можно получить системы уравнений для вычисления согласованных весовых векторов.

Например:

$$\left.\begin{aligned} \mu\alpha_i &= \sum \beta_\kappa x_{i\kappa} \\ \nu\beta_\kappa &= \sum \alpha_i x_{i\kappa} \end{aligned}\right\} \qquad (I)$$

$$\left.\begin{aligned} \mu\alpha_i &= \left(\sum \beta_\kappa w_{i\kappa}^2 \right)^\tau \\ \nu\beta_\kappa &= \left(\sum \alpha_i w_{i\kappa}^2 \right)^\tau \end{aligned}\right\} \qquad (2)$$

$$\left.\begin{aligned} \mu\alpha_i &= \sum \beta_\kappa x_{i\kappa} + h \\ \nu\beta_\kappa &= \exp\left\{ -\gamma \sum \alpha_i w_{i\kappa}^2 \right\} \end{aligned}\right\} \qquad (3)$$

$$\left.\begin{aligned} \mu\alpha_i &= \exp\{-\gamma_1 \sum \beta_\kappa \cdot w_{i\kappa}\} \\ \nu\beta_\kappa &= \exp\{-\gamma_2 \sum \alpha_i w_{i\kappa}\} \end{aligned}\right\} \qquad (4)$$

Выбор монотонно-возрастающей или монотонно-убывающей функции при задании весов соответствует предпочтению больших или малых отклонений, что определяется содержанием задачи. Например, в задаче о судейской коллегии можно принять систему (3): весовой коэффициент спортсмена возрастает вместе с оценками, а весовой коэффициент судьи убывает с ростом отклонения поставленной им оценки от среднего значения. Тем самым фактически вводится "штраф за необъективность".

Систему (I) рассмотрим более детально. Такой принцип задания весовых векторов предложен в работе [9]. В этом случае вес i-го объекта (признака) равен, с точностью до постоянного множителя, среднему взвешенному по строке (столбцу).

Из (I) следует:

$$\left.\begin{aligned} X X^T &= \lambda\alpha \\ X^T X &= \lambda\beta \end{aligned}\right\} \qquad (\lambda = \mu\nu) \qquad (5)$$

Если все элементы $x_{i\kappa}$ – неотрицательные и матрица $X^T X$ – неразложимая, то (5) имеет неотрицательное решение $\alpha_i \geq 0, \beta_\kappa \geq 0$ [I3].

Интересно отметить, что система (5) выражает условия экстремума функционала S, представляющего собой взвешенную сумму всех элементов:

$$S = \sum_i \sum_\kappa \alpha_i \beta_\kappa x_{i\kappa} \qquad (6)$$

При этом не обязательно, чтобы элементы $x_{i\kappa}$ были неотрицательными.

Спектральная задача вида $X^T X \beta = \lambda\beta$ возникает в методе главных компонент [3I], причем таблица X предполагается центрированной (средние по столбцам равны нулю). Тогда $X^T X$ есть ковариационная матрица признаков, а собственные векторы β задают направление главных осей выборочного эллипсоида (вторые моменты которого совпадают со вторыми моментами данной выборки объектов). Главный собственный вектор может иметь координаты с

лтными знаками, вследствие чего он не может быть принятым как весовой вектор. Таким образом, аналогия между методом главных компонент и вариантом (I) МСО не является полной.

Рассмотрим два квадратичных функционала:

$$Q_\alpha = \sum_\kappa \left\{ \sum_i \alpha_i x_{i\kappa} \right\}^2 = \alpha^T X X^T \alpha \qquad (7)$$

$$Q_\beta = \sum_i \left\{ \sum_\kappa \beta_\kappa x_{i\kappa} \right\}^2 = \beta^T X^T X \beta \qquad (8)$$

Нетрудно доказать, что максимум Q_α при условии $\sum \alpha_i^2 = I$ и максимум при условии $\sum \beta_\kappa^2 = I$ достигаются на решении α, β уравнений (5). (В методе главных компонент функционал Q_β характеризует разнос наблюдений вдоль главной оси выборочного эллипсоида).

Система (I) применялась для обработки геологических данных [I9]. Было обнаружено, что в ряде случаев весовые коэффициенты, полученные как решение (5), задают на множестве объектов порядок, допускающий содержательную интерпретацию. В частности, веса α_i имели значимую корреляцию с размером запасов полезных ископае - мых.

Уравнения (I),(5) симметричны относительно α, β - т.е. относительно объектов и признаков. Эта симметрия при центрировании теряется, ибо нуль-пункт определяется усреднением только по объектам и только по признакам. Попытка синтезировать вариант системы (I) с двумя типами усреднений (по строкам и столбцам одновременно) привела к системе, не имеющей нетривиальных решений.

"Согласованность" весовых векторов α и β, обусловленная уравнениями (I), является вырожденной, ибо система (5) распадается на две подсистемы: $X X^T \alpha = \lambda \alpha$ и $X^T X \beta = \lambda \beta$; таким образом, весовые векторы α и β можно искать независимо друг от друга.

Более сильная связь между весовыми векторами может быть установлена, если вместо (7),(8) взять два других функционала:

$$P_\alpha = \sum_\kappa \beta_\kappa^2 \left\{ \sum \alpha_i x_{i\kappa} \right\}^2 = \alpha^T X B^2 X^T \alpha \; ; \qquad (9)$$

$$P_\beta = \sum_i \alpha_i^2 \left\{ \sum \beta_\kappa x_{i\kappa} \right\}^2 = \beta^T X^T A^2 X \; ; \qquad (10)$$

$$A = \begin{vmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_m \end{vmatrix} ; \qquad B = \begin{vmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_n \end{vmatrix} \qquad (II)$$

Нормированные векторы α и β ($\sum \alpha_i^2 = I$, $\sum \beta_k^2 = I$), обращающие P_α и P_β в максимум, удовлетворяют системе:

$$\begin{aligned} X B^2 X^T \alpha &= \lambda_\alpha \alpha \\ X^T A^2 X \beta &= \lambda_\beta \beta \end{aligned} \qquad (I2)$$

Уравнения (I2) задают новый принцип согласования весов (в данном случае весами являются величины α_i^2, β_k^2). В отличие от (5) при отыскании экстремальных направлений учитываются веса объектов (в пространстве признаков) и веса признаков (в пространстве объектов). В нецентрированном варианте сохраняется симметрия "объект-признак". Поскольку такая симметрия довольно редко имеет физическое обоснование, этот принцип можно применить только для весьма узкого класса задач. Но, во всяком случае, интересно проверить эффективность такой методики для тех же ситуаций, где был эффективен подход с использованием уравнений (I), (5).

Перейдем от исходной таблицы X к центрированной таблице U : ее элементы $u_{ik} \in U$ суть отклонения от среднего по объектам (столбцам):

$$u_{ik} = x_{ik} - \sum \alpha_i x_{ik} \qquad (I3)$$

Вместо (I0) будем иметь:

$$P = \beta^T U^T A^2 U \beta . \qquad (I4)$$

Условие максимума P по β ($\sum \beta_k^2 = I$):

$$U^T A^2 U \beta = \lambda \beta . \qquad (I5)$$

Вес α_i определим выражением:

$$\mu \alpha_i = \Big| \sum_k \beta_k \cdot u_{ik} \Big| \Big/ \sum \alpha_i = I . \qquad (I6)$$

Такое соглашение о весах объектов приводит к растяжению выборки в направлении β , поскольку предпочтение (больший вес) отдается объектам, более удаленным от центра (в проекции на главную ось). Это позволяет выделить две крайние группы объектов. Уравнения

(I5),(I6) следует решать совместно, например методом итерации. Рассмотрим систему (2) при $\tau = -1$, $\mu = \frac{1}{n}$, $\nu = \frac{1}{m}$

$$\alpha_i^{-1} = \frac{1}{n} \sum_{k=1}^{n} \beta_k w_{ik}^2 \,, \qquad \beta_k^{-1} = \frac{1}{m} \sum_{i=1}^{m} \alpha_i w_{ik}^2 \tag{I7}$$

Такие уравнения возникают в задаче об уравновешивании матрицы данных, которая состоит в следующем ([47]).

Дана таблица неотрицательных чисел (экспериментальных данных) $W = \| w_{ik} \|$ размера $(m \times n)$. Строки таблицы W домножаются на множители $\alpha_i > 0$ $(i = 1, \ldots, m)$, столбцы на множители $\beta_k > 0$ $(k = 1, \ldots, n)$. Таким образом, осуществляется переход к преобразованной таблице $Z = \| z_{ik}^2 \|$:

$$z_{ik}^2 = \alpha_i w_{ik}^2 \beta_k \tag{I8}$$

Требуется найти такие векторы $\alpha = (\alpha_1, \ldots, \alpha_m)^T$, $\beta = (\beta_1, \ldots, \beta_n)^T$ с положительными компонентами, чтобы выполнялись одновременно два условия нормировки (по строкам и по столбцам):

$$\frac{1}{m} \sum_{i=1}^{m} z_{ik}^2 = 1 \quad (k = 1, \ldots, n); \tag{I9}$$

$$\frac{1}{n} \sum_{k=1}^{n} z_{ik}^2 = 1 \quad (i = 1, \ldots, m) \tag{20}$$

Легко убедиться, что из (I8),(I9),(20) следует (I7).

В работе [47] доказано, что эта задача имеет решение, если W не содержит нулевых элементов (достаточное условие). (Теорема: "Всякая прямоугольная матрица без нулевых элементов уравновешиваема"). Однако решение существует и для многих матриц, содержащих нулевые элементы, что подтверждается примерами.

Для нахождения α, β, обладающих требуемыми свойствами, в [47] предложен специальный прием, который может применяться также для уравнений, выражающих связь между весовыми коэффициентами. Вводится потенциал (или потенциальная функция):

$$\varphi(\alpha) = \frac{1}{n} \sum_{i} \sum_{j} \ln \sum_{k} \frac{\alpha_k}{\alpha_i} \frac{w_{kj}^2}{\sum_s w_{sj}^2} \tag{2I}$$

Градиент $\Delta\varphi(\alpha)$ потенциала $\varphi(\alpha)$ есть вектор с компонентами:

$$g_i(\alpha) = \frac{n_i}{n} \sum_{j=1}^{n} \frac{w_{ij}^2}{\sum_{\kappa} \alpha_\kappa w_{\kappa j}^2} - \frac{I}{\alpha_i}; \quad (i = I, \ldots, m) \qquad (22)$$

На искомом решении $g_i(\alpha) = 0$.
Аналогично можно ввести потенциал $\psi(\beta)$:

$$\psi(\beta) = \frac{I}{m} \sum_{\kappa} \sum_{j} \ell n \sum_{i} \frac{\beta_\kappa}{\beta_j} \frac{w_{j\kappa}^2}{\sum w_{js}^2} \qquad (23)$$

с градиентом $h(\beta) = \nabla\psi(\beta)$

$$h_\kappa(\beta) = \frac{n}{m} \sum_{i=1}^{m} \frac{w_{ij}^2}{\sum \beta_s \cdot w_{sj}^2} - \frac{I}{\beta_\kappa}; \quad (\kappa = I, \ldots, n) \qquad (24)$$

Векторы α, β, удовлетворяющие (I9),(20), суть стационарные точки потенциалов $\varphi(\alpha)$, $\psi(\beta)$ соответственно. Можно показать, что в этих точках $\varphi(\alpha)$ и $\psi(\beta)$ достигают минимума. Имеется два подхода к минимизации $\varphi(\alpha)$ или $\psi(\beta)$:

I. Решение уравнений (I9),(20) с помощью итерационного метода;
2. Непосредственный поиск минимума потенциальной функции методом покоординатного спуска или каким-либо иным градиентным методом.

Второй подход обладает достаточной общностью и может быть применен для нахождения весовых коэффициентов. Потенциальные функции во многих случаях, представляющих практический интерес, могут быть найдены элементарным интегрированием. В частности, для системы (I) они имеют вид:

$$\varphi(\alpha) = \frac{I}{2} \sum_{i=1}^{m} (I + w_{ii}^2)\alpha_i^2 - \frac{I}{2} \sum_{i=1}^{m} \sum_{s=1}^{m} w_{is}\alpha_i\alpha_s; \qquad (25)$$

$$\psi(\beta) = \frac{I}{2} \sum_{\kappa=1}^{n} (I + \tilde{w}_{ii}^2)\beta_\kappa^2 - \frac{I}{2} \sum_{\kappa=1}^{n} \sum_{s=1}^{n} \tilde{w}_{is}\beta_\kappa\beta_s, \qquad (26)$$

где w_{is}, $\tilde{w}_{is}$ - элементы ковариационных матриц $X X^T$, $X^T X$.

Замечание. Система (I) не нуждается в специальной методике, поскольку она сводится к спектральной задаче (5), для решения которой имеются типовые программы на ЭВМ. Но для других систем, выражающих связь между весовыми коэффициентами, градиентный метод поиска экстремума потенциальной функции может оказаться предпочтительнее, поскольку отпадает необходимость в разработке особой вычислительной процедуры в каждом конкретном случае.

В задаче о судейской коллегии можно принять (помимо (3)) следующую систему связи между весовыми коэффициентами:

$$\left.\begin{aligned} \mu\alpha_i &= \sum \beta_\kappa x_{i\kappa} \\ \tilde{\nu}\beta_\kappa &= \Big(\sum \alpha_i w_{i\kappa}^2\Big)^{-1} \end{aligned}\right\} \qquad (27)$$

или в матричном виде:

$$\left.\begin{aligned} \mu\alpha &= X\beta \\ \nu\beta^{-1} &= W^T\alpha \end{aligned}\right\} \qquad (28)$$

Здесь обозначено: $\nu = \tilde{\nu}^{-1}$; $W = \|w_{i\kappa}^2\|$; компоненты вектора β^{-1} суть обратные величины по отношению к соответствующим компонентам вектора β. Из (28) следует:

$$W^T X\beta = \mu\nu\beta^{-1}. \qquad (29)$$

Обозначив $W^T X = A$; $\mu\nu = \lambda$, получаем:

$$A\beta = \lambda\beta^{-1}. \qquad (30)$$

Потенциальная функция для этой системы имеет вид:

$$\psi(\beta) = \sum_{i=1}^{n}\sum_{s=1}^{n} a_{is}\beta_i\beta_s - \frac{I}{2}\sum_{i=1}^{n} a_{ii}\beta_i^2 - \lambda\sum \ln\beta_i \qquad (3I)$$

Если домножить каждое из уравнений (30) на β_κ, в левой части получим квадратичные формы относительно β_κ. Выполнив приведение их к каноническому виду (сумме квадратов), получим линейную алгебраическую систему относительно β_κ, которую можно решать обычными методами.

§7. АЛГОРИТМЫ И ПРОГРАММНЫЕ РЕАЛИЗАЦИИ ПО МЕТОДУ СОГЛАСОВАННЫХ ОЦЕНОК

I. Вводные замечания

Здесь рассматриваются такие процедуры, которые отражают тот или иной перечень целевых условий [35], в рамках метода согласованных оценок. Предварительно строится вспомогательная таблица, по которой производится вычисление оценок строк и столбцов для исходной таблицы. Упомянутая вспомогательная таблица называется таблицей целевых особенностей и строится после предварительной нормировки значений исходной таблицы, которая производится следующим образом.

2. Нормировка исходной таблицы

Пусть дана первоначальная таблица X (исходная) размерности $m \times n$ из элементов x_{ij}, стоящих в пересечении i – й строки и j –го столбца, а также целевой признак Y (исходный) из m элементов y_i.

Нормировка таблицы X сводится к построению таблицы U из элементов u_{ij} вида:

$$u_{ij} = \frac{x_{ij} - \min\limits_{1 \le \alpha \le m} \{x_{\alpha j}\}}{\max\limits_{1 \le \alpha \le m} \{x_{\alpha j}\} - \min\limits_{1 \le \alpha \le m} \{x_{\alpha j}\}}, \quad 1 \le i \le m, \quad 1 \le j \le n.$$

Аналогично проводится нормировка значений целевого признака Y, приводя к построению признака z из элементов z_i, равных

$$z_i = \frac{y_i - \min\limits_{1 \le \alpha \le m} \{y_\alpha\}}{\max\limits_{1 \le \alpha \le m} \{y_\alpha\} - \min\limits_{1 \le \alpha \le m} \{y_\alpha\}}.$$

Затем строится таблица целевых особенностей, как описано ниже. Рассмотрим построение таблицы целевых особенностей.

Пусть дана таблица U и перечень $\mathcal{D}$ (целевой), в котором для каждой строки u_i из таблицы U указывается

$\mathcal{D}_{i1}$) с какими из строк u_1, u_2, ... , u_{i-1} строка u_i должна быть сходна, $2 \le i \le m$;

$\mathcal{D}_{i2}$) с какими из строк u_{i+1} , u_{i+2} , ... , u_m строка u_i должна быть различима , $1 \le i \le m-1$.

Таблица целевых особенностей, обозначаемая $\Delta_{\mathcal{D}} U$, строится таким путем. Для каждой пары строк (u_i , u_k) из таблицы U , такой, что $1 \le i \le k \le m$, в таблицу $\Delta_{\mathcal{D}} U$ включается строка $[\Delta (i , k)]$, если u_k принадлежит $\mathcal{D}_{i2}$ (если u_k не принадлежит $\mathcal{D}_{i2}$, то строка $\Delta (i , k)$ не включается в таблицу $\Delta_{\mathcal{D}} U$); кроме того, если u_i принадлежит $\mathcal{D}_{k1}$, то в таблицу $\Delta_{\mathcal{D}} U$ включается строка $[1 - \Delta (i , k)]$ (если же u_i не принадлежит $\mathcal{D}_{k1}$, то строка $[1 - \Delta (i , k)]$ не включается в таблицу $\Delta_{\mathcal{D}} U$); никакие другие строки не включаются в таблицу $\Delta_{\mathcal{D}} U$.

Строка $[\Delta (i , k)]$ составляется из элементов

$$\Delta (i , k)_j = u_{ij} - u_{kj} ,$$

а строка $[1 - \Delta (i , k)]$ - из элементов

$$1 - \Delta (i , k)_j = 1 - u_{ij} + u_{kj} ,$$

" u_k принадлежит $\mathcal{D}_{i2}$ " означает, что u_i и u_k должны быть различимы; " u_i принадлежит $\mathcal{D}_{k1}$ " означает, что u_i и u_k должны быть сходными.

Аналогично построению таблицы $\Delta_{\mathcal{D}} U$ строится и столбец $\Delta_{\mathcal{D}} z$.

Далее опишем процедуру получения оценок строк и столбцов.

3. Вычисление согласованных оценок

Пусть даны таблицы U , $\Delta_{\mathcal{D}} U$ и признаки z , $\Delta_{\mathcal{D}} z$. Таблицу $\begin{bmatrix} U \\ \Delta_{\mathcal{D}} U \end{bmatrix}$, в которой строки из $\Delta_{\mathcal{D}} U$ расположены под строками из U , обозначим через T , а соответствующий столбец $\begin{bmatrix} z \\ \Delta_{\mathcal{D}} z \end{bmatrix}$ обозначим через C .

Пусть b - число строк таблицы T (столбца C), а C_i -

i -я компонента столбца C, $i = 1,2, \ldots, b$.

Оценки строк $-\pi_i$ и оценки столбцов $-\omega_j$ получаются, исходя из C и T, с помощью следующего итерационного процесса.

Шаг 0. Пола ается $\pi_i^{(0)} = 1$, $\omega_j^{(0)} = 1$, $i = 1,2, \ldots, b$; $j = 1,2, \ldots, n$.

Шаг $k+1$, $k \geq 0$. Приближения $\pi_i^{(k+1)}$, $\omega_j^{(k+1)}$ получаются из приближений $\pi_i^{(k)}$, $\omega_j^{(k)}$ по формулам:

$$\pi_i^{(k+1)} = \left\{ \frac{\sum_{j=1}^{n} \left[\omega_j^{(k)} \left(1 - \left| t_{ij} - c_i \right| \right) \right]^2}{\sum_{j=1}^{n} \left[\rho_j^{(k)} \right]^2} \right\}^{1/2}, \quad i = 1,2,\ldots,b ,$$

$$= \left\{ \frac{\sum_{i=1}^{b} \left[\pi_i^{(k)} \left(1 - \left| t_{ij} - c_i \right| \right) \right]^2}{\sum_{i=1}^{b} \left[\pi_i^{(k)} \right]^2} \right\}^{1/2}, \quad j = 1,2,\ldots,n .$$

В качестве оценок π_i , ω_j принимаются предельные значения

$$\pi_i = \lim_{k \to \infty} \pi_i^{(k+1)} , \quad \omega_j = \lim_{k \to \infty} \omega_j^{(k+1)} .$$

Оценки π_i , ω_j и представляют собой искомые целевые согласованные оценки строк и столбцов таблицы T .

В ряде случаев оценки столбцов могут быть известны и нужно получить только оценки строк.

Ниже дается три варианта подсчета.

4. Получение оценок строк при заданных оценках столбцов

Пусть даны таблица X размерности $m \times n$, строка $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ и строка весов ω_j , $j = 1,2,\ldots, n$.

Целесообразно провести подсчет таких вариантов весов строк, как набор весов X вида

$$x_i = \sum_{j=1}^{n} x_{ij}\,\omega_j \qquad , \qquad i = 1,2, \ldots , m \quad ,$$

набор α_i вида

$$\alpha_i = \left\{ \sum_{j=1}^{n} \left[(x_{ij} - y_j)\,\omega_j \right]^2 \right\}^{1/2} \qquad , \quad i = 1,2, \ldots , m$$

и набор весов β_i вида

$$\beta_i = \sum_{j=1}^{n} \frac{\left[(x_{ij} - x_{j\min}) \cdot (\bar{x}_j - x_{j\min}) + (x_{j\max} - x_{ij})(x_{j\max} - \bar{x}_j) \right] \cdot \omega_j}{(x_{j\max} - x_{j\min})^2},$$

где $\bar{x}_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij}$, $x_{j\min} = \min \{x_{\alpha j}\}$, $x_{j\max} = \max_{1 \le \alpha \le m} \{x_{ij}\}$.

До сих пор рассматривались процедуры над таблицами, в которых все элементы суть известные численные значения. Однако часто встречаются таблицы, в которых некоторые значения неизвестны, представляют собой "прочерки" (–).

Ниже излагается алгоритм оценки неизвестных значений.

5. Алгоритм восстановления прочерков

Пусть дана таблица T из M строк и N столбцов, имеющая в наличии ℓ прочерков:

$(I_1, J_1), (I_2, J_2), \ldots, (I_\ell, J_\ell)$, где I_t – номер строки, а J_t – номер столбца, в котором стоит прочерк с номером t , $1 \le t \le \ell$.

Алгоритм состоит из пяти частей последовательного приближения.

Шаг 1) Пусть $\varepsilon_{IJ} = \begin{cases} 1, & \text{если } t_{IJ} \text{ – известное число,} \\ 0, & \text{если } t_{IJ} \text{ – прочерк,} \end{cases}$

где t_{IJ} – элемент I-й строки таблицы T , стоящий в J-м столбце. Для получения первого приближения к оценке неизвестных значений таблица T обрабатывается по методу согласованных оце-

нок с помощью следующего итерационного процесса.

Задается начальное приближение весов столбцов таблицы T:

$$T - \beta_s^{(0)} = I, \qquad s = I,2, \ldots, N.$$

Веса строк $\alpha_r^{(q)}$ начального и последующих приближений находятся по формуле

$$\alpha_r^{(q)} = \frac{\sum_{k=1}^{N} \varepsilon_{rk} \cdot \beta_k^{(q)} \cdot t_{rk}}{\sum_{k=1}^{N} \varepsilon_{rk} \cdot \beta_k^{2}}, \quad r = I,2,\ldots,M, \quad q \geq 0. \quad (I)$$

Веса же столбцов получаются так:

$$\beta_s^{(q+1)} = \frac{\sum_{i=1}^{M} \varepsilon_{is} \cdot \alpha_i^{(q)} \cdot t_{is}}{\sum_{i=1}^{M} \varepsilon_{is} \cdot \alpha_i^{(q)2}}, \quad s = I,2,\ldots,N, \quad q \geq 0; \quad (2)$$

для того, чтобы числитель формул (I), (2) был определен, полагаем, что 0 "-" = 0; это и естественно, ибо произведение нуля на любое конечное число дает 0.

В качестве окончательных весов столбцов берутся такие веса $\beta_s^{(\tau)}$, что

$$\left| \beta_s^{(\tau)} - \beta_s^{(\tau-1)} \right| < \delta, \qquad s = I,2, \ldots, N, \qquad (3)$$

где δ - порог точности получения весов столбцов.

В качестве окончательных весов строк берутся веса $\alpha_r^{(\tau-1)}$, $r = I,2, \ldots, M$, где τ определяется, как указано выше, по весам $\beta_s^{(q)}$.

Прочерк на месте (I_t, J_t) в таблице T замещается на значение $\alpha_{I_t}^{(\tau-1)} \cdot \beta_{J_t}^{(\tau)}$ для всех t, $I \leq t \leq \ell$. Значения $\alpha_{I_t}^{(\tau-1)} \cdot \beta_{J_t}^{(\tau)}$ суть первые приближения к оценке неизвестных значений, а полученная из T таблица обозначается через T_I.

Шаг 2) Таблица T алгебраически умножается "сама на себя" по формуле $U = T_I \cdot T_I'$, где T_I' означает транспонирование таблицы T_I.

Если t_{IJ} обозначает элемент I-й строки J -го столбца таблицы T_I, то элемент u_{IK} таблицы U равен:

$$u_{IK} = \sum_{J=1}^{N} t_{IJ} \cdot t_{KJ} \quad , \quad I,K = I,2, \ \dots \ ,M .$$

Для таблицы U находятся собственные числа $\lambda_1 > \lambda_2 > \dots > \lambda_r$ и соответствующие им собственные векторы $\tilde{\alpha}_1$, $\tilde{\alpha}_2$, ... , $\tilde{\alpha}_r$.

Шаг 3) Таблица T_I преобразуется в T_2 вычитанием из каждого элемента t_{IJ} таблицы T_I произведения весов $\alpha_I^{(\tau-1)} \cdot \beta_J^{(\tau)}$, полученных на шаге I).

Шаг 4) Пусть собственный вектор $\tilde{\alpha}_2 = (\alpha_1^2, \alpha_2^2, \ \dots \ , \alpha_M^2)$, соответствующий собственному числу λ_2 , полученный на шаге 2), представляет собой начальное приближение весов строк таблицы T_2, т.е. $\alpha_r^{(0)} = \alpha_r^2$, $r = I,2, \ \dots \ ,M$, а все последующие приближения весов столбцов и строк вычисляются по формулам (I),(2) . При пороге точности ε получим, на некотором шаге, как по формуле (3), заключительное приближение весов строк и весов столбцов таблицы T_2. Обозначим через $(\gamma_1^{(2)}, \gamma_2^{(2)}, \ \dots \ , \gamma_M^{(2)})$, $(\delta_1^{(2)}, \delta_2^{(2)}, \ \dots \ , \delta_M^{(2)})$ наборы заключительных приближений весов строк и столбцов таблицы T_2. С их помощью второе приближение к оценке прочерков дается суммой:

$$\alpha_I^{(\tau-1)} \cdot \beta_J^{(\tau)} + \gamma_I^{(2)} \delta_J^{(2)} .$$

Шаг 5) Каждая последующая таблица T_{q+1} : $T_3, T_4, \ \dots \ , T_r$ образуется из предыдущей вычитанием из каждого элемента $t_{IJ}^{(q)}$ предыдущей таблицы T_q произведения ее весов $\gamma_I^{(q)} \delta_J^{(q)}$, $q = 2,3, \ \dots \ , r - I$.

В качестве начального приближения весов строк таблицы T_{q+I} принимается собственный вектор $\tilde{\alpha}_{q+1}$, а все последующие приближения получаются по формулам (I),(2), с той оговоркой, что вместо t_{rK} , t_{is} берутся элементы $t_{rK}^{(q+1)}$, $t_{is}^{(q+1)}$ из таблицы T_{q+I} . Заключительное приближение, получаемое по формулам (I),(2),(3) при некотором пороге ε , обозначается через $(\gamma_1^{(q+1)}, \gamma_2^{(q+1)}, \dots, \gamma_M^{(q+1)})$, $(\delta_1^{(q+1)}, \delta_2^{(q+1)}, \ \dots \ , \delta_N^{(q+1)})$.

Окончательная оценка неизвестного значения (I_t, J_t) представляется в виде величины $x_{I_t J_t}$, равной

$$x_{I_t J_t} = \alpha_{I_t}^{(\tau-1)} \cdot \beta_{J_t}^{(\tau)} + \sum_{q=2}^{z} \gamma_{I_t}^{(q)} \cdot \delta_{J_t}^{(q)} .$$

Изложив алгоритмы, перейдем теперь к изложению программных реализаций.

6. Программные реализации

Комплекс программ по методу согласованных оценок состоит из программ МС0I – МС05, написанных на языке α для ЭВМ М-222. Программы МС0I – МС04 предназначены для получения числовых оценок строк и столбцов таблиц описаний. Точность оценок $-(10^{-6})$. Программа МС05 предназначена для восстановления пропусков. Программы МС0I – МС03 выполняются одна за другой за один пуск. Программа МС04 может выполняться отдельно, но после выполнения программ МС0I – МС03. Программа МС04 выполняется по мере необ – ходимости. В ходе работы в программах используются устройства ввода, печати, магнитный барабан. Распечатки текстов программ и блок-схема комплекса МС0 приведены в приложении.

Программа МС0I
"Нормировка"

Назначение: Программа МС0I предназначена для нормировки исходного массива и записи нормированного массива U на магнитный барабан.

Входные данные:

1. M – число строк X;
2. N – число столбцов X;
3. X – исходный массив.

На печать выводится:

1. Значения параметров M, N;
2. Значения элементов массива X.

Программа МС02.
"Построение массива целевых особенностей"

Назначение: Программа МС02 предназначена для построения массива целевых сравнений и слияния построенного массива с нормированным массивом U . Полученный массив T выводится на МБ.

Входные данные:

1. M – число строк U;
2. N – число столбцов U ;
3. Д – массив целевых требований;
4. U – нормированный массив.

На печать выводится:

1. Значения параметров M,N;
2. K – число сходств и различий массива U ;
3. Значения элементов массива Д .

Программа МС03
"Вычисление согласованных оценок"

Назначение: Программа МС03 предназначена для вычисления согласованных оценок строк и столбцов массива целевых особенностей.

Входные данные:

1. B – число строк T (B = M + K);
2. N – число столбцов T;
3. T – массив целевых особенностей.

На печать выводится:

1. Значения параметров B,N;
2. Оценки строк ω_i, $i = I, \ldots, B$;
3. Оценки столбцов ω_j, $j = I, \ldots, N$.

Программа МС04
"Распознавание"

Назначение: Программа МС04 предназначена для вычисления оценок строк по заданным оценкам столбцов.

Входные данные:

1. MI – число строк Y ;

2. N – число столбцов Y ;
3. ω – массив весов размерности N;
4. γ – типичная строка (массив размерности N).

На печать выводится:

1. Значения оценок строк ψ_i ;
2. Значения оценок строк α_i ;
3. Значения оценок строк β_i (где i = I, ... , MI).

Инструкция к пользованию

I. Подготовка данных.

Параметры, входные массивы перфорируются на КУ-I. Каждый параметр перфорируется на отдельной перфокарте. Массивы перфорируются построчно.

2. Порядок составления пакета:

Паспорт MC0I
Рабочая программа MC0I
Параметры M,N
Массив X
Паспорт MC02
Рабочая программа MC02
Параметры M,N
Массив Д
Паспорт MC03
Рабочая программа MC03
Параметры B,N
Паспорт MC04
Рабочая программа MC04
Параметры MI и N
Массив Y
Массив W
Массив γ.

Ограничения: Величина $(M+K)x(N+2)+2N \leq 3385$,
для M = N = K имеем M = 40

Программа МС05
"Восстановление прочерков"

Назначение: Программа МС05 предназначена для вычисления значений "пустых" элементов массива и выдачи их на печать.

Входные данные:

1. М2 - число строк;
2. М2 - число столбцов;
3. L - число прочерков;
4. z - массив с прочерками;
5. ПР - массив прочерков размерности L x 2.

На печать выводится:

1. Значение параметров М2, N2, L ;
2. Значение элементов массива z ;
3. На каждую итерацию выводятся:
   а) веса строк и столбцов,
   б) значения прочерков и их индексы (число итераций) $\mu = \min(M2, N2)$.

Инструкция к пользованию

I. Подготовка данных.

Параметры, массивы z , ПР перфорируются на КУ-I. Массив перфорируется построчно. Вместо прочерков перфорируются нули. Структура записи массива ПР имеет вид:

номер строки прочерка
номер столбца прочерка
} в массиве z .

Массив прочерков определяет положение прочерков в массиве z .

Порядок составления пакета.

Паспорт МС05
Рабочая программа
Параметры М2, N2, L .
Массив z .
Массив ПР.

После каждого параметра, после массива z , после каждой записи массива ПР должна стоять п/к "Блокировка К$\overline{\sum}$ ".

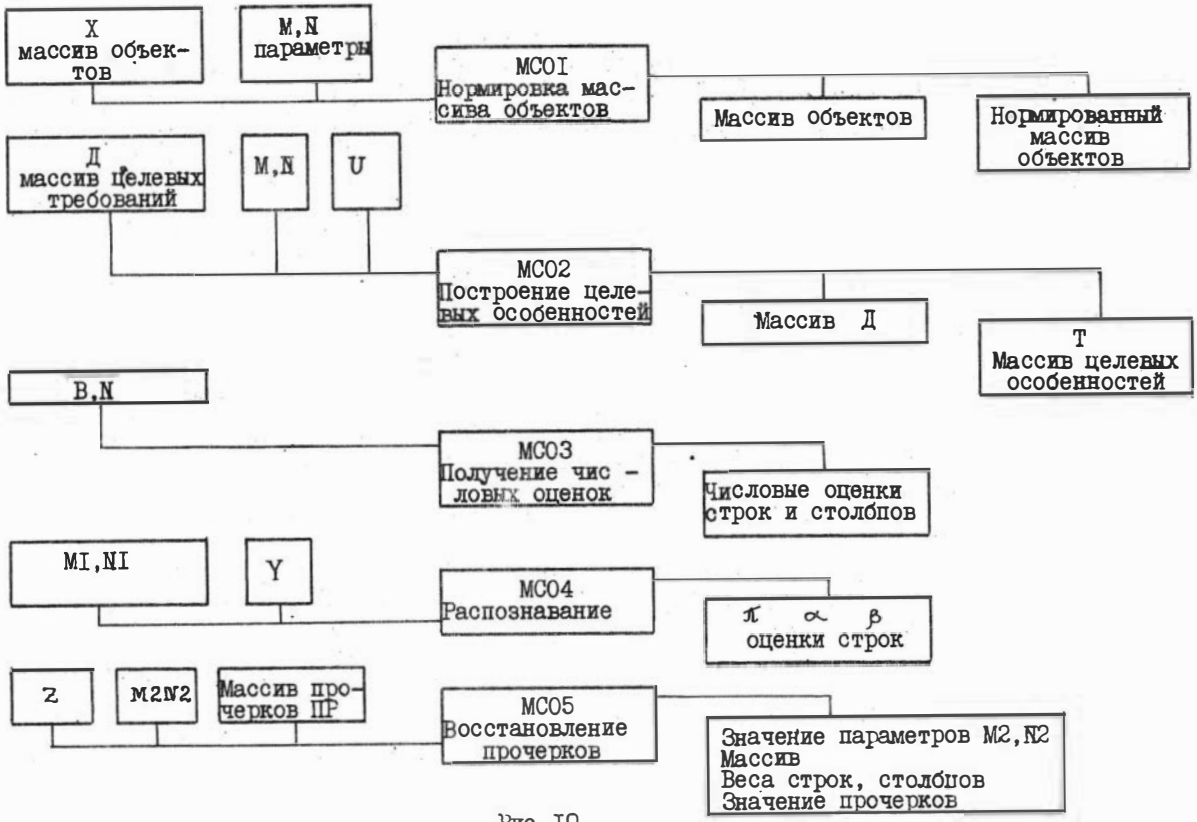

X
массив объектов
M,N
параметры
MC01
Нормировка массива объектов
Массив объектов
Нормированный массив объектов
Д
массив целевых требований
M,N
U
MC02
Построение целевых особенностей
Массив Д
T
Массив целевых особенностей
B,N
MC03
Получение числовых оценок
Числовые оценки строк и столбцов
MI,NI
Y
MC04
Распознавание
π α β
оценки строк
Z
M2N2
Массив прочерков ПР
MC05
Восстановление прочерков
Значение параметров M2,N2
Массив
Веса строк, столбцов
Значение прочерков

Рис. IO

Ограничения: по программе можно обрабатывать массивы, для которых выполняется неравенство

$$M^2 + MN + M + N \leq 1300 \quad \text{или},$$

если $M = N$, то $M \leq 25$

Время решения порядка 10 минут.

В программе использована процедура ГАСОВ, реализованная на ВЦ ИГиГ.

Для образного представления работы изложенного комплекса программ можно воспользоваться схемой, представленной на рис. 10. В центре рисунка располагается наименование программ комплекса, слева от него - исходные данные, а справа - характер получаемых результатов.

§8 НЕКОТОРЫЕ СВЯЗИ МСО С ДРУГИМИ МЕТОДАМИ

Для установления и осмысливания природы связи МСО с другими родственными ему методами отметим, что математической точкой роста данного метода являются возможности, содержащиеся в процедурах сингулярного разложения. Подчеркнем – МСО это прямо вытекающий подход из теоремы о сингулярном разложении Экарта-Янга [50,21], статистическая модификация которой осуществлена в теореме Рао-Даррока [52,48]. Обнаруженные координаты МСО в окрестности других методов линейно-статистического профиля, позволяют более точно выйти на взаимосвязи с другими подходами не только в плане процедурного сходства, но и в плане подобия идеологий. Важно иметь также ввиду, что, насколько нам известно, нецентрированный метод главных компонент для обработки данных не применялся, а если и применялся, то эпизодически и крайне редко. В этом отношении МСО является, в своей основной нецентрированной процедуре, существенным дополнением.

1. Общие соображения

Напомним, что оценки строк и столбцов в таблицах данных производятся "качаниями". Качания эти таковы, что последующие приближения к оценкам строк получаются предыдущими приближениями оценок столбцов, а последующие приближения к оценкам столбцов выявляются предыдущими приближениями оценок строк. Условием конечности числа качаний является прием согласования весов строк и весов столбцов между собой. Это согласование выражается тем, что результат множества последовательных единообразных пересчетов оценок столбцов может дать предельную оценку строк таблицы за один такой пересчет, при условии, что оценки столбцов также предельны. Это справедливо и обратно, по предельной оценке строк в один пересчет получаются предельные оценки столбцов. Поскольку предельные оценки строк принимаются в качестве весов

строк, а предельные оценки столбцов – в качестве весов столбцов, то мы и приходим к выводу, что веса строк и столбцов таблицы суть согласованные оценки. При решении ряда задач по методу согласованных оценок оказалось, что веса строк хорошо коррелируют со значениями целевого признака. Таким образом выясняется, что исходный материал, в неявном виде, уже подготовлен для такой корреляции, причем веса строк, подсчитанные по методу согласованных оценок, как бы проявляют уже заложенную в материале корреляцию, а другими словами – соответствие значениям целевого признака весов строк.

Эти соображения позволяют объяснить эффективность метода согласованных оценок в решении задач целевой классификации объектов и наметить пути поиска связей с другими сходными методами. Другое свойство метода – быстрота сходимости оценок строк и столбцов к предельным – позволяет решать, в приемлемое время, задачи с большим количеством строк и столбцов. Указанные свойства метода согласованных оценок привели к проявлению близких по форме алгоритмов типа "Цикл", разработанных А.А.Бишаевым, который сделал следующие два дополнения к имевшимся построениям, описанным в предыдущем разделе. В работе [5] предложено проводить нормирующее преобразование исходных таблиц на основе вычисления расстояний прежде, чем производить оценку их строк и столбцов. Это преобразование проводится с учетом разных типов шкал признаков и пропусков значений. Кроме того, для возможно полного соответствия оценок строк значениям целевого признака по некоторому критерию качества предложено также ввести значения целевого признака явным образом в процедуру вычисления оценок строк. Назовем первое предложение А.А.Бишаева правилом нормирующего преобразования исходного материала, а второе – итеративным правилом подсчета оценок строк и столбцов с включением целевого признака в явной форме.

2. Формализованное выражение приемов обработки таблиц

Дадим краткое обсуждение этих правил. Пусть T – исходная таблица, z – ее целевой признак, $x_1, x_2, \ldots, x_n$ – ее столб-

цы, а S_1 , S_2 , ... , S_m – ее строки; пусть ω_1, ω_2, ... , ω_n – веса столбцов таблицы T , а π_1, π_2 , ... , π_m – веса строк.

Тогда первое правило – прием согласования – можно записать следующим образом:

Веса ω_1 , ω_2 , ... , ω_n, π_1, π_2, ... , π_m так соответствуют исходной таблице T , что пара наборов (ω , π) вида

$$((\omega_1 , \omega_2 , \dots , \omega_n), (\pi_1, \pi_2, \dots , \pi_m))$$

получается один из другого однократным применением пары вектор-функций (φ , ψ) так, что

$$\omega = \varphi(T , \pi) , \quad \pi = \psi (T , \omega). \qquad (I)$$

Второе правило – прием нормирующего преобразования таблицы можно сформулировать так:

Прежде, чем считать веса ω, π для столбцов x_j и строк S_i таблицы T , таблица T преобразуется с помощью вектор-функции ρ к виду U:

$$U = \rho (T). \qquad (2)$$

Третье правило – включения в подсчет оценок значений целевого признака, в явной форме, может быть записано так :

Для лучшего соответствия весов строк значениям целевого признака при подсчете употребляются такие вектор-функции α , β, что

$$\omega = \alpha(\bar{z} , T , \pi), \quad \pi = \beta(\bar{z} , T , \omega). \qquad (3)$$

Правила, соответствующие выражениям (2), (3), содержательно были сформулированы в [6], а конкретный вид функций ρ , α , β был реализован в программах метода "Цикл" [7] , а впоследствии в программе "Каскад – П" [3].

Уместно также сформулировать правило для образования таблицы целевых особенностей исходной таблицы T :

Для подсчета весов ω, π столбцов и строк таблицы T образуется таблица U с помощью вектор-функции σ :

$$U_c = \sigma (\mathcal{D} , T) , \qquad (4)$$

где $\mathcal{D}$ – перечень целевых условий к строкам таблицы T , который определяется в работе по подсчету тупиковых $\mathcal{D}$ – тестов [35].

Вышеописанные четыре правила играют большую роль при по –

с роении обобщений на основе МСО, но для построения самих обобщений целесообразно напомнить характеристику метода процедурно.

Метод согласованных оценок включает 2 вида процедур обработки таблиц – нецентрированная качельная процедура (НКП) и центрированная качельная процедура (ЦКП). Причем в рамках нецентрированной качельной процедуры производится оценка выраженности признаков на объектах, что могло бы соответствовать статистически оценке математического ожидания этой выраженности. В пределах же ЦКП вычисляются оценки отклонений выраженности признаков на объектах от средних величин, что соответствует в статистике оценке дисперсии.

Алгоритм НКП метода согласованных оценок [9] мало отличается от алгоритма ЦКП[28] и поэтому целесообразно сначала изложить математическую модель, которая обобщает оба эти алгоритма, а также включает, в виде частных случаев, известные алгоритмы "Цикл – I", "Цикл – 2" [6] и "Каскад – П" [3], в которых производится оценка соответствия между характеристической проявленностью строк и их целевой ценностью; затем, уже в рамках модели, легко указать место каждого алгоритма.

3. Математическая модель

Имеется таблица T из элементов t_{ij}, стоящих в пересечении строк S_i со столбцами x_j, отражающих естественную проявленность признака, обозначенного через x_j на объекте, обозначенном символом S_i. Число строк в таблице T равно m, а число столбцов – n

Имеется столбец z из m числовых элементов z_i, отражающих целевую ценность строк S_i, $i = I,2, \dots, m$. Элементы t_{ij} таблицы T и элементы z_i столбца z могут принадлежать различным числовым множествам, например, $0 \leq t_{ij} < \infty$, $0 \leq z_i \leq \infty$, или $t_{ij} \in \{0,I\}$, $z_i \in \{0,I\}$, или $t_{ij} \in \{0,I,2, \dots, k-I\}$, $z_i \in \{0,I,2, \dots, k-I\}$, $k > 0$ и т.п.

Образуются разности $t_{ij} - t_{i_1 j}$, $z_i - z_{i_1}$, $i, i_1 = I, 2, \dots, m$; $j = I,2, \dots, n$, причем, если паре (i, i_1) соответствует номер ν, при образовании разностей, то обозначается $\Delta t_{\nu j} = t_{ij} - t_{i_1 j}$, $\Delta z_\nu = z_i - z_{i_1}$.

Если M – число различных пар (i, i_1), для которых берутся разности, то элементы Δt_{vj} образуют таблицу ΔT размерности $M \times n$, а элементы Δz_v – столбец Δz из M компонент.

Укажем на содержательный смысл значений t_{ij}, z_i и разностей $t_{ij} - t_{i_1 j}$, $z_i - z_{i_1}$.

Элемент t_{ij} отражает собой независимую проявленность признака x_j на объекте S_i, не зависящую от его проявленности на других объектах.

Разность $t_{ij} - t_{i_1 j}$ отражает собой зависимую проявленность признака x_j на объекте S_i, зависящую от его проявленности на объекте S_{i_1}.

Аналогично величина z_i отражает собой независимое значение целевой ценности строки S_i, а величина $z_i - z_{i_1}$ – зависимое значение целевой ценности строки, зависящее от ценности строки S_{i_1}.

Продолжим изложение процедурной части.

Строится таблица F из строк таблиц T, ΔT путем помещения всех строк таблицы ΔT под всеми строками таблицы T, т. е. $F = \begin{bmatrix} T \\ \Delta T \end{bmatrix}$.

Образуется таблица U путем удаления из F некоторых строк и модификации других.

Вычисляются оценки строк π_i и столбцов ω_j для таблицы U. Вычисление корректируется величиной J соответствия между оценками строк таблицы U и отвечающими им значениями элементов пары $(z, \Delta z)$.

В зависимости от способов получения набора π оценок строк и набора ω – оценок столбцов, а также в зависимости от способа выбора таблицы U и характера соответствия J могут применяться различные конкретные вычислительные процедуры. Поэтому четверка (U, π, ω, J) определяет вид таких процедур. Введем следующее определение.

Определение I. Моделью обработки по методу согласованных оценок называется четверка (U, π, ω, J).

Целесообразно рассмотреть конкретные проявления изложенной модели с указанием, каковы их составляющие элементы, что выбирается в качестве U, π, ω, J, в каждом конкретном случае.

4. Нецентрированная качельная процедура

В рамках изложенной модели основная нецентрированная качельная процедура [33] метода согласованных оценок характеризуется тем, что:

I) в качестве таблицы U берется сама таблица T, т.е. имеет место равенство $U = T$;

2) в качестве оценок строк S_i таблицы T берутся предельные значения последовательностей $\pi_i^{(k+1)}$ вида:

$$\pi_i^{(k+1)} = \frac{\sum_{j=1}^{n} t_{ij} \cdot \omega_j^{(k)}}{\sum_{\alpha=1}^{m}\sum_{j=1}^{n} t_{\alpha j}\, \omega_j^{(k)}}, \quad i = I,2,\ldots,m; \quad \omega_j^{(0)} = I, \quad j = I,2,\ldots,n;$$

3) в качестве оценок столбцов x_j таблицы T употребляются значения пределов последовательностей $\omega_j^{(k+1)}$, получаемых по формулам:

$$\omega_j^{(k+1)} = \frac{\sum_{i=1}^{m} t_{ij}\, \pi_i^{(k)}}{\sum_{\beta=1}^{n}\sum_{i=1}^{m} t_{i\beta}\cdot \pi_i^{(k)}}, \quad j = I,2,\ldots,n; \quad \pi_i^{(0)} = I, \quad i = I,2,\ldots,m;$$

сделаем к 2), 3) следующее важное добавление: значения $\omega_j^{(k)}$, при $k \geq I$, требующиеся в 2) для вычисления $\pi_i^{(k+1)}$, вычисляются в 3), а значения $\pi_i^{(k)}$ для 3), при $k \geq I$, вычисляются в 2);

4) в качестве меры соответствия оценок $\pi_i^{(k+1)}$ значениям z_i целевого признака употребляется величина

$$J^{(k+1)} = \min_{1 \leq i \leq m} \left\{ \frac{\pi_i^{(k+1)}}{\pi_i^{(k)}}, \frac{\pi_i^{(k)}}{\pi_i^{(k+1)}} \right\}.$$

Обратим внимание на то, что в данном случае оценка $J^{(k+1)}$ соответствия между значениями $\pi_i^{(k+1)}$ и z_i является вырожденной, так как $J^{(k+1)}$ явно от z не зависит. Тем не менее, каким-то неявным образом $J^{(k+1)}$ зависит и от z, ибо окончательные предельные оценки π_i, как показано в [9,10], в ряде случаев, сильно коррелируют со значениями z_i.

5. Центрированная качельная процедура

В рамках модели центрированная качельная процедура, изложенная выше (§2) и в [24], характеризуется следующими элементами:

1) в качестве таблицы U употребляется сама таблица T, $U = T$;

2) в качестве оценок строк S_i таблицы T принимаются предельные значения последовательностей $\pi_i^{(k+1)}$ такого вида:

$$\pi_i^{(k+1)} = \left(\frac{\sum_{j=1}^{n} \left[\omega_j^{(k)} \cdot (t_{ij} - \bar{t}_{ij}^{\,q}) \right]^2}{\sum_{\alpha=1}^{m} \sum_{j=1}^{n} \left[\omega_j^{(k)} (t_{\alpha j} - \bar{t}_{\alpha j}^{\,q}) \right]^2} \right)^{1/2}, \quad i = 1,2,\dots,m,$$

где $q \in \{0,1\}$, $\bar{t}_{ij}^{\,0} = \frac{1}{n} \cdot \sum_{j=1}^{n}$, $\bar{t}_{ij}^{\,1} = \frac{1}{m} \sum_{i=1}^{m} t_{ij}$,

$\omega_j^{(0)} = 1$, $j = 1, \dots, n$;

3) в качестве оценок столбцов x_j таблицы T служат предельные значения последовательностей $\omega_j^{(k+1)}$, образованных по формуле

$$\omega_j^{(k+1)} = \left(\frac{\sum_{i=1}^{m} \left[\pi_i^{(k)} (t_{ij} - \bar{t}_{ij}^{\,1-q}) \right]^2}{\sum_{\beta=1}^{n} \sum_{i=1}^{m} \left[\pi_i^{(k)} (t_{i\beta} - \bar{t}_{i\beta}^{\,1-q}) \right]^2} \right)^{1/2}, \quad j = 1,2, \dots, n,$$

где q – то же самое значение, что и в 2), $\pi_i^{(0)} = 1$,

$i = 1,2, \ldots, m$; при $q = 1$ получается один вариант ЦКП, а при $q = 0$ – другой; 2), 3) согласованы, так как значения $\omega_j^{(k)}$ для 2) при $k \geq 1$ получаются из 3), а значения $\pi_i^{(k)}$ для 3) при $k \geq 1$ – из 2);

4) в качестве меры соответствия оценок $\pi_i^{(k+1)}$ значениям z_i целевого признака z применяется величина

$$J^{(k+1)} = \min_{1 \leq i \leq m} \left\{ \pi_i^{(k+1)} / \pi_i^{(k)}, \ \pi_i^{(k)} / \pi_i^{(k+1)} \right\}$$

6. Процедуры метода "целевой итерационной классификации" (Цикл – 1)

В рамках метода целевой итерационной классификации [7] процедура "Цикл – 1" характеризуется следующими составляющими:

1) в качестве таблицы U употребляется таблица ΔT , т.е. $U = \Delta T$;

2) за оценки строк G_ν таблицы ΔT принимаются предельные значения последовательностей $\pi_\nu^{(k+1)}$

$$\pi_\nu^{(k+1)} = \sum_{j=1}^{n} \Delta \tilde{t}_{\nu j} \cdot \omega_j^{(k+1)}, \quad \nu = 1,2, \ldots, M = C_m^2,$$

где $\Delta \tilde{t}_{\nu j} (\Delta t_{\nu j} - \min_\nu \{\Delta t_{\nu j}\}) / (\max_\nu \{\Delta t_{\nu j}\} - \min_\nu \{\Delta t_{\nu j}\})$, а $\omega_j^{(0)}$ таковы, что $0 \leq \omega_j^{(0)} \leq 1$, причем $\sum_{j=1}^{n} \omega_j^{(0)} = 1$;

3) в качестве оценок столбцов Δx_j таблицы ΔT берутся предельные значения последовательностей $\omega_j^{(k+1)}$

$$\omega_j^{(k+1)} = \omega_j^{(k)} \frac{A_j^{(k)}}{\sum_{j=1}^{n} A_j^{(k)} \cdot \omega_j^{(k)}}, \quad j = 1,2, \ldots, n,$$

где $A_j^{(k)} = \sum_{\nu=1}^{M} \left[A_{\nu j}^{(k)} + \bar{A}_{\nu j}^{(k)} \right]$, причем

$$A_{\nu j}^{(k)} = \min(\Delta \tilde{t}_{\nu j}, \Delta \tilde{z}_\nu) \cdot \max (1, \Delta \tilde{z}_\nu / \pi_\nu^{(k)}),$$

$$\bar{A}_{\nu j}^{(k)} = \min(1 - \Delta \tilde{t}_{\nu j}, 1 - \Delta \tilde{z}_\nu) \cdot \max(1, (1 - \Delta \tilde{z}_\nu)/(1 - \pi_\nu^{(k)})),$$

$$\Delta\tilde{z}_y = \frac{\Delta z_y - \min_y \{\Delta z_y\}}{\max_y \{\Delta z_y\} - \min_y \{\Delta z_y\}} ;$$

4) в качестве меры соответствия оценок $\pi_y^{(k+1)}$ значениям Δz_y целевого признака принимается величина

$$J^{(k+1)} = \frac{2M}{\sum_{y=1}^{M} (\mathcal{D}_y^{(k+1)} + \bar{\mathcal{D}}_y^{(k+1)})} ,$$

где $\mathcal{D}_y^{(k+1)} = max(I, \Delta\tilde{z}_y/\pi_y^{(k+1)})$, $\bar{\mathcal{D}}_y^{(k+1)} = max(I, (I-\Delta\tilde{z}_y)/(I-\pi_y^{(k+1)}))$.

Очевидно, что в данном случае оценка $J^{(k+1)}$ соответствия между значениями $\pi_y^{(k+1)}$ и Δz_y – невырожденная, т.к. явно зависит от $\pi^{(k+1)}$ и от Δz.

7. Процедура метода целевого классифицирования объектов (Каскад -П)

В рамках метода целевого классифицирования объектов процедуры "Каскад -П" характеризуются следующими особенностями:

1) в качестве таблицы U употребляется таблица T, $U = T$;

2) за оценки строк S_i таблицы T принимаются предельные значения последовательностей $\pi_i^{(k+1)}$, равных

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \hat{t}_{ij} \cdot \omega_j^{(k+1)} , \quad i = 1,2, \ldots, m ,$$

где $$\hat{t}_{ij} = \begin{cases} \tilde{t}_{ij} , & \text{при } \omega_j^{(k+1)} \geq 0, \\ -1, & \text{при } \omega_j^{(k+1)} < 0 , \end{cases}$$

причем $$\tilde{t}_{ij} = \frac{t_{ij} - \min_i \{t_{ij}\}}{\max_i \{t_{ij}\} - \min_i \{t_{ij}\}} ;$$

3) в качестве оценок столбцов x_j таблицы T берутся предельные значения последовательностей $\omega_j^{(k+1)}$

$$\omega_j^{(k+1)} = \omega_j^{(k)} \cdot \frac{\Lambda_j^{(k)}}{\sum_{j=1}^{n} \Lambda_j^{(k)} \cdot \omega_j^{(k)}},$$

где $\Lambda_j^{(k)} = \dfrac{\sum_{i=1}^{m} |\hat{t}_{ij}| \dfrac{\tilde{z}_i}{\pi_i^{(k)}}}{\sum_{i=1}^{m} |\hat{t}_{ij}|}$, $\tilde{z}_i = \dfrac{z_i - \min\limits_{\alpha} \{z_\alpha\}}{\max\limits_{\alpha} \{z_\alpha\} - \min\limits_{\alpha} \{z_\alpha\}}$;

4) в качестве меры соответствия оценок $\pi_i^{(k+1)}$ значениям z_i целевого признака применяется величина

$$J^{(k+1)} = I - \max_{1 \le i \le m} \left\{ | \tilde{z}_i - \pi_i^{(k+1)} | \right\},$$

которая явно зависит как от $\pi^{(k+1)}$, так и от z, а следовательно представляет собой невырожденную оценку соответствия между значениями $\pi_i^{(k+1)}$ и z_i.

Рассмотренные выше 4 варианта процедур имеют программное обеспечение и представляют собой идеологическое и процедурное развитие метода согласованных оценок. Тот факт, что эти процедуры описываются в рамках одной и той же модели, показывает ее широту. Возникает вопрос и о том, исчерпывают ли рассмотренные 4 процедуры все возможности модели. Отметим, что максимальные изменения в рассмотренных вариантах метода претерпевали элементы π, ω, J модели, а в качестве таблицы U выступали только T и ΔT. Можно, согласно модели, в качестве U взять и другие модификации таблиц T, ΔT. Поэтому разработка модификаций вида, таблицы U представляет дальнейший интерес.

Естественно, что в выборе таблицы U существенную роль должны играть, в сочетании, обе характеристики строк: как их выраженность, характеризуемая таблицей T, так и их взаимоотношение, описываемое таблицей ΔT.

Далее, в связи с тем, что в задачах фигурируют целевой признак z и его производное Δz, то желательно согласовывать выраженность таблицы T со значениями z, а взаимоотношение по ΔT – со значениями Δz. Кроме того, желательно также табли-

цу U составлять в соответствии с целью, что можно было бы назвать четвертым правилом - целевого препарирования таблицы исходных данных перед оценкой ее строк и столбцов. Изложим теперь один способ построения таблицы U, опирающийся на возможности тестового подхода [46,16].

8. Целевой способ выбора таблицы U_c

Произведем нормировку значений таблиц T, ΔT вида:

$$t^*_{ij} = \frac{t_{ij} - \min\limits_{\alpha}\{t_{\alpha j}\}}{\max\limits_{\alpha}\{t_{\alpha j}\} - \min\limits_{\alpha}\{t_{\alpha j}\}} \qquad \Delta t^*_{\nu j} = \frac{\Delta t_{\nu j} - \min\limits_{1\le k\le M}\{\Delta t_{kj}\}}{\max\limits_{1\le k\le M}\{\Delta t_{kj}\} - \min\limits_{k}\{\Delta t_{kj}\}}.$$

Полученные таблицы значений t^*_{ij} и $\Delta t^*_{\nu j}$ обозначим T^* и ΔT^* и назовем последнюю таблицей различий.

Такой же нормировкой преобразуются целевой признак z и производный от него Δz в столбцы z^* и Δz^* по формулам:

$$z^*_i = \frac{z_i - \min\limits_{\alpha}\{z_\alpha\}}{\max\limits_{\alpha}\{z_\alpha\} - \min\limits_{\alpha}\{z_\alpha\}}, \quad \Delta z^*_\nu = \frac{\Delta z_\nu - \min\limits_{k}\{\Delta z_k\}}{\max\limits_{k}\{\Delta z_k\} - \min\limits_{k}\{\Delta z_k\}}.$$

Образуем таблицы из элементов

$$\bar{t}^*_{ij} = I - t^*_{ij}, \qquad \Delta\bar{t}^*_{\nu j} = I - \Delta t^*_{\nu j},$$

которые обозначим $\bar{T}^*$ и $\Delta\bar{T}^*$, причем последнюю назовем таблицей сходств.

Из строк таблиц T^*, $\bar{T}^*$, ΔT^*, $\Delta\bar{T}^*$ составим таблицу U_c, сравнивая компоненты столбцов z^* и Δz^* со значением 0.5, по следующим шести правилам:

1) если z^*_i неизвестно, то в таблицу U_c не помещаются ни i - я строка из T^*, ни i - я строка из $\bar{T}^*$;
2) если неизвестно Δz^*_ν, то в таблицу U_c не заносятся ни ν - я строка таблицы ΔT^*, ни ν - я строка таблицы $\Delta\bar{T}^*$;
3) если $z^*_i \geqslant 0{,}5$, то в U_c записывается i - я строка

таблицы T^*;

4) если $z^* \leq 0{,}5$, то в U_c заносится i – я строка таблицы $\bar{T}^*$;

5) если $\Delta z^*_\nu \geq 0{,}5$, то в U_c включается ν – я строка из ΔT^*;

6) если $\Delta z^*_\nu \leq 0{,}5$, то в U_c помещается ν – я строка таблицы $\Delta \bar{T}^*$.

Изложенный способ построения таблицы U_c близок по идее способу построения целевого перечня $\mathcal{D}$ [35]. И там и здесь акцентируются сходства и различия между объектами. Но здесь еще и акцентируются проявленность и невыраженность объектов. Кроме того, здесь и сходства, и различия, и проявленность, и невыраженность функционируют вместе и их желательно согласовывать.

Такое согласование и производится тем, что строки таблиц T^*, $\bar{T}^*$, ΔT^*, $\Delta \bar{T}^*$ включаются на равных в одну и ту же таблицу U и подвержены одним и тем же оценочным процедурам поиска величин π, ω.

Не останавливаясь более на целевом способе выбора U, разберем модификации элементов π, ω [29], построенные на "поощрении" и "наказании" тех или иных столбцов и строк таблицы, обладающих нужными качествами.

9. Способы поощрения весовых коэффициентов в методе согласованных оценок

Заметим, что обычно таблица T представляет собой либо сводку экспериментальных данных, либо – экспертных оценок (в последнем случае "признак" есть мнение эксперта).

"Поощрению" или "наказанию" могут подлежать либо проявленность, либо невыраженность самих объектов или признаков, либо их отклонение от среднего значения объекта или среднего значения признака, либо другие характеристики объектов и признаков. Причем веса строк и столбцов должны быть согласованными между собой.

В силу последнего предложения, нам будет достаточно написать формулы для выражения самих весов друг через друга, т.к формулы для получения последующих приближений через предыдущие

получаются из них очевидным образом : левой части равенств приписывается сверху индекс ($k+1$), а правой части – индекс (k).

В работе [29] предлагаются следующие ряды формул для построения весов строк π_i и весов столбцов ω_j :

$$(\pi)\quad \begin{cases} \mu\cdot\pi_i = \sum_{j=1}^{n} \omega_j \cdot t_{ij} + h\ , & (\pi_1)\\ \mu\cdot\pi_i = \left(\sum_{j=1}^{n} \omega_j \cdot w_{ij}^{2}\right)^{\tau}\ , & (\pi_2)\\ \mu\cdot\pi_i = \exp\left\{-\gamma \sum_{j=1}^{n} \omega_j t_{ij} + h\right\}, & (\pi_3)\\ \mu\cdot\pi_i = \exp\left\{-\gamma \left(\sum_{j=1}^{n} \omega_j \cdot w_{ij}^{2}\right)^{\tau}\right\}, & (\pi_4) \end{cases}$$

$$(\omega)\quad \begin{cases} \nu\cdot\omega_j = \sum_{i=1}^{m} \pi_i \cdot t_{ij} + h\ , & (\omega_1)\\ \nu\cdot\omega_j = \left(\sum_{i=1}^{m} \pi_i \cdot w_{ij}^{2}\right)^{\tau}\ , & (\omega_2)\\ \nu\cdot\omega_j = \exp\left\{-\gamma\cdot \sum_{i=1}^{m} \pi_i \cdot t_{ij} + h\right\}, & (\omega_3)\\ \nu\cdot\omega_j = \exp\left\{-\gamma\cdot \sum_{i=1}^{m} (\pi_i \cdot w_{ij}^{2})^{\tau}\right\}, & (\omega_4), \end{cases}$$

где w_{ij} равно либо $t_{ij} - \frac{1}{m}\sum_{i=1}^{m} t_{ij}$, либо $t_{ij} - \frac{1}{n}\sum_{j=1}^{n} t_{ij}$.

Там также предлагается, комбинируя формулы (π_i) и (ω_j), $i, j = 1,2,3,4$, получать системы уравнений для согласованных весовых векторов.

Заметим, что формулы для нецентрированной и центрированной качельной процедур могут быть получены, как частные случаи та-

ких комбинаций, но уже формулы для процедур "Цикл" и "Каскад-П" так не выводятся. Для выведения последних необходимо в формулы поощрения весов вводить значения целевого признака.

В работах [6,3] для поощрения весов столбцов употребляется такая формула:

$$\omega_j^{(k+1)} = \omega_j^{(k)} \cdot \frac{\Lambda_j^{(k)}}{\sum_{j=1}^{n} \Lambda_j^{(k)} \cdot \omega_j^{(k)}} , \qquad (5)$$

где $\Lambda_j^{(k)} = \dfrac{\sum_{i=1}^{m} \hat{t}_{ij} \dfrac{\tilde{z}_i}{\pi_i^{(k)}}}{\sum_{i=1}^{m} \hat{t}_{ij}}$, $\tilde{z}_i = \dfrac{z_i - \min\limits_{1 \le \alpha \le m} \{z_\alpha\}}{\max\limits_{\alpha} \{z_\alpha\} - \min\limits_{\alpha} \{z_\alpha\}}$.

С учетом формулы (ω 5) могут быть получены формулы для всех вышеописанных процедур, включая "Цикл" и "Каскад - П", а также и другие формулы, еще не введенные в практику обработки таблиц. Они, однако, не исчерпывают, на наш взгляд, всего разнообразия возможных процедур поощрения весов столбцов и строк, что нуждается в дополнительном исследовании.

Здесь следует еще рассмотреть меры соответствия оценок $\pi_i^{(k+1)}$ значениям z_i целевого признака.

IO. Меры соответствия оценок цели

В вышеописанных методах в качестве меры соответствия употреблялись следующие 3 характеристики :

I) $\min\limits_{1 \le i \le m} \left\{ \dfrac{\pi_i^{(k+1)}}{\pi_i^{(k)}} , \dfrac{\pi_i^{(k)}}{\pi_i^{(k+1)}} \right\}$;

2) $\dfrac{2M}{\sum_{\nu=1}^{M} (\mathcal{D}_\nu^{(k+1)} + \bar{\mathcal{D}}_\nu^{(k+1)})}$;

$$3) \qquad I - \max\left\{\left|\tilde{z}_i - \pi_i^{(k+1)}\right|\right\}.$$

Здесь, в первом случае, мера отражает лишь близость оценок $\pi_i^{(k)}$, $\pi_i^{(k+1)}$ к предельным, т.е. целью является приближение величины $\pi_i^{(k)}$ к предельной.

В двух последних случаях меры отражают близость значений $\pi_i^{(k)}$ к значениям z_i, а целью является приближение величин $\pi_i^{(k)}$ к значениям z_i.

Мера 2) отражает среднюю близость значений $\pi_i^{(k)}$ к значениям z_i, а мера 3) – близость "наихудшего" из значений $\pi_i^{(k)}$ к z_i.

В мере 2) качество приближения определяется отношениями вида $\frac{\pi_i^{(k)}}{z_i}$, а в мере 3) – разностями вида $|\pi_i^{(k)} - z_i|$, что и придает им разные значения; эти различия делают меры 2) и 3) взаимно дополняющими.

Вопрос о полном наборе мер соответствия, исчерпывающем все возможные соответствия оценок строк значениям целевого признака, остается пока открытым. Помимо мер I) – 3) можно употреблять и другие, например, меру вида:

$$4) \qquad \frac{I}{m}\sum_{i=1}^{m}\left(I - \left|\tilde{z}_i - \pi_i^{(k+1)}\right|\right).$$

Вышеуказанные меры соответствия принимают значение в отрезке $[0,I]$. Чем ближе значение J к I, тем выше соответствие оценок $\pi_i^{(k)}$ значениям z_i для мер 2) – 4). Для меры I) близость J к I означает близость значений $\pi_i^{(k)}$, $\pi_i^{(k+1)}$ к предельным значениям π_i для всех $i = 1,2, \ldots, m$, и из формул не видно того, чтобы $\pi_i^{(k)}$, $\pi_i^{(k+1)}$ соответствовали значениям z_i целевого признака z. Однако, для нецентрированной качельной процедуры такое соответствие существует, и, хотя в явном формализованном виде оно и не усматривается, но всё же оценки строк, подсчитанные по НКП, зачастую коррелируют с тестовыми информационными весами строк, подсчитанными для этой же бинарной таблицы. Такая корреляция и показывает, что имеется хорошее соответствие между $\pi_i^{(k)}$, $\pi_i^{(k+1)}$ и z_i, так как обычно тестовые инфор-

мационные веса строк имеют высокую корреляцию со значениями x_i. Представляется целесообразным изложить это полнее.

II. Взаимосвязи с тестовым подходом

Вышеизложенные в 4. соотношения I) – 4) можно трактовать как модель: с одной стороны как модель оценивания объектов по их признакам и с другой стороны как содержательную модель. Разумеется, в последнем случае необходимо, чтобы даваемые моделью оценки были содержательно интерпретируемыми.

Факты о математических связях представляют собой указания на нетривиальные взаимоотношения между результатами тестового подхода и методом согласованных оценок. Это взаимоотношение выполняется довольно отчетливо на различном эмпирическом материале и может служить некоторым дополнительным доводом за существенность этих подходов в целом.

Отметим ряд особенностей введенной модели, которые полезно иметь ввиду при ее практическом применении:

а) Чувствительность нагрузок к количеству единиц в соответствующих строках или столбцах таблицы T. Если число единиц в какой-либо строке (или в каком-либо столбце) велико по сравнению с остальными строками (или столбцами), то соответствующая нагрузка будет большой просто благодаря этому превосходству в числе единиц. Поэтому нагрузки объектов и признаков зависят от того, как выбрано кодирование. "Эффект перекрытий" проявляется более отчетливо, когда хотя бы в столбцах таблицы T число единиц и нулей примерно одинаково. Такие таблицы дают и более хорошие оценки с точки зрения интерпретируемости их. Желательно, чтобы при кодировании, т.е. при построении таблиц, указанное условие по возможности соблюдалось.

б) Различия в чувствительности нагрузок к изъятию из таблицы различных строк или столбцов. Наблюдений в этом направлении сделано немного, но выраженность эффекта позволяет предположить, что он может иметь довольно общую природу. Эффект состоит в том, что изъятие из геологической таблицы строк, имеющих большие нагрузки, ведет к резкому изменению величин нагрузок оставшихся строк и всех столбцов; при изъятии же строк с малой нагрузкой

соответствующее изменение невелико. В первом случае строки играют роль как бы "корневых".

в) Связь нагрузок с тестовыми весами. Эта связь является эмпирическим фактом, проявляющимся довольно отчетливо на всех таблицах (около 30), для которых были найдены тестовые веса и нагрузки. Связь оказалась нетривиальной и содержание ее легче всего усмотреть из схематического изображения на рис. II.

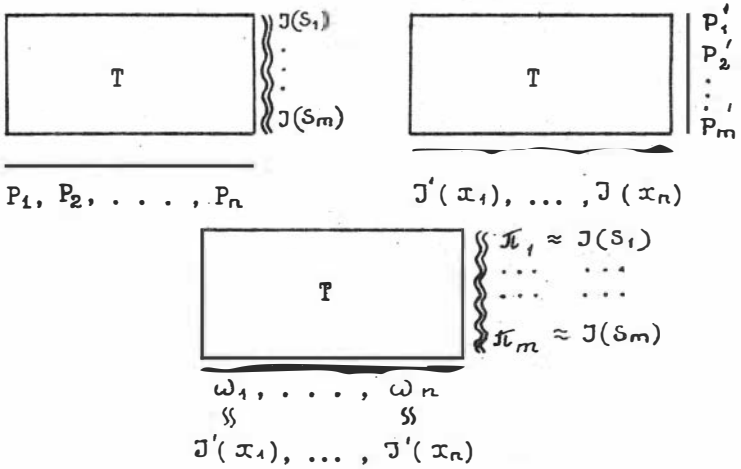


Рис. II

На рис. II слева вверху представлен подсчет тестовых весов столбцов P_j и строк $J(s_i)$; справа вверху представлен аналогичный подсчет для "перевернутой" таблицы T, т.е. для случая, когда тесты строятся из строк таблицы T, подсчитываются веса P'_i для строк, а затем - веса столбцов (как линейные комбинации весов строк); внизу - подсчет нагрузок для T. Волнистыми линиями отмечены наборы оценок, которые с точностью до множителя примерно одинаковы (это изображено знаком $\approx$).

Рис. II указывает, что в прямой тестовой процедуре [I6] имеет мес-

то близость между тестовыми весами строк и нагрузками строк, а при "перевернутой" тестовой процедуре аналогичное соотношение имеет место для оценок столбцов. Следует отметить, что в первом случае имеет также место резкое различие между тестовыми весами столбцов и нагрузками для столбцов, а во втором – резкое различие между оценками строк.

Указанные различия не случайны и являются отражением существенного различия тестовых столбцов и нагрузок столбцов по их смыслу – первые выражают для столбцов таблицы их "различающую" способность в этой таблице, а вторые – способность, которую можно охарактеризовать скорее как "связывающую". Обнаруженная связь не является общим правилом – в примере этой связи нет. Вместе с тем, эта связь довольно отчетливо наблюдается на "реальных" таблицах, пример же не отвечает какой-либо содержательной задаче. Эта связь означает, что имеет место уравнивание противоположностей, однако ссылка на такое уравнивание сама по себе ничего не объясняет, так как те же самые противоположности имеются в таблицах, где такой связи почему-то нет.

Если вернуться к определению модели (U, π, ω, J), то заметим, что в качестве U, т.е. исходной таблицы для подсчета оценок π, ω, выбиралась таблица T. В тестовом подходе этому соответствует подсчет количества всех тупиковых пакетов для таблицы T. Подсчету же количества всех тупиковых тестов для таблицы T соответствует такая качельная процедура, в которой в качестве таблицы U берется T^* – так называемая "таблица сравнений" для таблицы T, получаемая следующим образом: в таблице T каждой паре строк таблицы T ставится в соответствие строка из нулей и единиц, в которой единицы соответствуют различиям, а нули – совпадениям компонент пары, если m – число строк таблицы T, то в T^* число строк равно C_m^2. Если в качестве U взять таблицу T, получаемую из T заменой всюду нулей на I, а единиц на 0, то качельная процедура на такой таблице U соответствует подсчету всех тупиковых Ω-тестов таблицы T. Если же, согласно перечню $\mathfrak{D}$ целевых условий, выбрать в качестве U соответствующую таблицу $T_{\mathfrak{D}}$, то НКП для такой таблицы U соответствует подсчету количества всех тупиковых $\mathfrak{D}$ – тестов таблицы T. $T_{\mathfrak{D}}$ строится так: каждой паре строк из T, на

которую в $\mathcal{D}$ наложено условие согласования, в таблице $T_{\mathcal{D}}$ ставится в соответствие две строки: одна из таблицы T^*, другая - из $\overline{T}^*$; каждой паре строк из T, на которые в $\mathcal{D}$ налагается требование сходства, отвечает одна строка из $\overline{T}^*$; каждой паре строк из T, подлежащей различению, отвечает строка из T^*; все остальные строки таблиц T^*, $\overline{T}^*$ удаляются, а оставшиеся после удаления строки таблиц T^*, $\overline{T}^*$ образуют таблицу $T_{\mathcal{D}}$.

I2. Теорема Экарта-Янга и метод согласованных оценок

Рассматривая "метод согласованных оценок", предназначенный для упорядочивания и классификации многомерных объектов, следует подчеркнуть связь этого метода с одним классическим результатом, основанным на теореме Экарта-Янга о базисной структуре.

Пусть задано конечное множество объектов $a_i = x_{i\cdot}$ $(i = 1,\dots,m)$, характеризуемых признаками $b_\kappa = x_{\cdot\kappa}$ $(\kappa = 1,\dots,n)$, и каждой паре (a_i, b_κ) сопоставлено число $x_{i\kappa}$ (значение признака b_κ на объекте a_i).

Матрица $X = \| x_{i\kappa} \|$ называется таблицей "объект-признак". Её строки суть $x_{i\cdot}$ (объекты), столбцы - $x_{\cdot\kappa}$ (признаки).

Примем (хотя это не очень существенно), что $m \geq n$. Пусть r - ранг матрицы X^TX; $r \leq n$. Обозначения: λ_i - ненулевые собственные числа (с.ч.) матрицы X^TX, занумерованные в убывающем порядке: $\lambda_1,\dots,\lambda_r$. Вектор u_i (v_i) есть собственный вектор (с.в.) матрицы X^TX (XX^T), соответствующий с.ч. λ_i.

По определению с.в. имеем:

$$X^TX\,u_i = \lambda_i u_i, \qquad (1)$$

$$XX^T v_i = \lambda_i v_i. \qquad (2)$$

Из (1), (2) следуют соотношения между u_i и v_i:

$$v_i = \lambda^{-1/2} X u_i, \qquad (3)$$

$$u_i = \lambda^{-1/2} X^T v_i. \qquad (4)$$

В прикладных исследованиях часто возникает задача о нахождении сжатого представления таблицы данных. Так например, иногда требуется аппроксимировать таблицу X матрицей меньшего ранга [2I].

Задача: Дана прямоугольная матрица X ранга r. Найти матрицу X_k ранга $k \le r$, на которой обращается в минимум евклидова норма погрешности:

$$R = \|X - X_k\|^2 = tr[(X - X_k)^T(X - X_k)]. \qquad (5)$$

Решение: (см. [50],[21]):

$$X_k = \lambda_1^{1/2} v_1 u_1^T + \ldots + \lambda_k^{1/2} v_k u_k^T . \qquad (6)$$

При $k = r$ (если положить $X_2 = X$) соотношение (6) превращается в сингулярное разложение матрицы X, гарантируемое теоремой Экарта-Янга о базисной структуре [50], (см. также [21]).

Рассмотрим частный случай: $k = 1$. Найдем приближенное представление таблицы данных в виде произведения столбца на строку, с матрицей погрешностей R:

$$X = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} (\beta_1, \ldots, \beta_n) + R = \alpha\beta^T + R \qquad (7)$$

так, чтобы обратилась в минимум норма погрешности:

$$\|P\|^2 = tr[(X - \alpha\beta^T)^T (X - \alpha\beta^T)] \longrightarrow \min \qquad (8)$$

при условии $\beta^T\beta = 1$. Легко показать, что искомые векторы $\alpha = (\alpha_1, \ldots, \alpha_m)^T$, $\beta = (\beta_1, \ldots, \beta_n)^T$, удовлетворяют системе уравнений:

$$\alpha = X\beta / \lambda, \quad \lambda\beta = X^T\alpha, \qquad (9),(10)$$

где λ – главное с.ч. матрицы X^TX. (Отметим, что $\alpha^T\alpha = \lambda$). Отсюда следует, что векторы α, β совпадают, с точностью до нормировки, с решением уравнений (3),(4), или, что то же самое, уравнений (1),(2), при $\lambda = \lambda_1$.

Задача в такой постановке описана в работах [49], [53], в связи с проблемой оценивания пропущенных данных.

Система вида (9),(10) хорошо изучена в математической статистике. Если данные x_{ik} – центрированные и нормированные,

уравнения (9),(10) выражают преобразование, на котором основан известный метод главных компонент (МГК). При $\lambda = \lambda_1$ вектор α есть первая главная компонента выборки X; вектор β (главный с.в. ковариационной матрицы X^TX) задает направление наибольшего рассеивания выборки в пространстве признаков. Но следует заметить, что в работах [50],[49]данные не предполагаются центрированными .

Переходя к характеристике некоторых особенностей метода согласованных оценок, дадим краткую справку о возникновении метода.

Принцип упорядочивания объектов и признаков, состоящий в приписывании объектам α_i (и одновременно - признакам β_κ) числовых оценок α_i (β_κ), связанных уравнениями (9),(10),был предложен А.Н.Дмитриевым. Этот принцип, вместе с результатами его применения, неоднократно, начиная с 1967 года, обсуждался на семинарах в Институте Математики СО АН СССР. Возникший на его основе метод классификации был назван впоследствии "Методом согласованных оценок" (МСО); это название подчеркивает своеобразную согласованность оценок α_i, β_κ, которая выражается симметричными уравнениями (9),(10). По существу, МСО есть нецентрированный вариант МГК, причем в МСО используется только первая составляющая разложения (6), соответствующая главному с.ч.λ.

При обработке по МСО геологических данных было обнаружено интересное явление, состоящее в наличии значимой положительной корреляции между оценками объектов и некоторым целевым признаком, не включенным в таблицу X. (Результаты обработок см., например, в [9,10]). Таким образом, в некоторых случаях МСО помогает предсказывать значение целевого признака, которое трудно получить непосредственным измерением. Это позволяет считать МСО новым методом обработки экспериментальных данных, нуждающимся в теоретическом обосновании.

Если таблица содержит только неотрицательные данные ($x_{i\kappa} \geq 0$), то, в силу специфических свойств неотрицательных матриц, решением системы (9),(10) являются векторы с неотрицательными компонентами, которые можно интерпретировать как весовые коэффициенты объектов и признаков. (Некоторые другие способы нахождения согласованных весовых коэффициентов рассмотрены в [29]).

Вычислительные методы, позволяющие решить систему (9),(10),

можно найти, например, в [45].

Чтобы МСО перестал быть чисто эвристическим приемом, необходимо выяснить условия, при которых возникает вышеупомянутая значимая корреляция с целевым признаком.

В работе [9] (где данный метод описан впервые) вопрос о корреляции оценок с целевым признаком не рассматривается, и поэтому она не может считаться теоретическим обоснованием метода. Работа [9] не содержит новых математических результатов. Основное содержание этой статьи составляет итерационная процедура вычисления оценок, являющаяся разновидностью известного итерационного метода вычисления главного с.в. (см. [45]). Принятое в [9] довольно жесткое предположение о том, что таблица составлена только из нулей и единиц, является излишним. Не отмечена формальная связь предложенного способа оценивания с другими методами обработки данных, например, с МГК, с теоремой о базисной структуре, и т.д. В целом, сущность проблемы остается нераскрытой, и положительное значение статьи [9] - только в том, что она привлекла внимание к проблеме.

По-видимому, значимая корреляция оценок с целевым признаком может иметь место лишь при удачном выборе и кодировании признаков. Но соответствующее исследование пока никем не выполнено.

На основании вышеизложенного можно сделать два вывода:

I. МСО, как метод классификации, пока что теоретически полностью не обоснован, хотя его эффективность подтверждена рядом вычислительных экспериментов.

2. Оценки, полученные по МСО, являются решением классической оптимизационной задачи об аппроксимации таблицы данных матрицей меньшего ранга (в частности, ранга I) (Теорема Экарта-Янга, I936). Это позволяет надеяться, что МСО (в усовершенствованном варианте) может стать эффективным средством анализа данных.

§9 ПРИМЕРЫ КОНКРЕТНЫХ РЕШЕНИЙ

Рассмотрим кратко некоторые конкретные примеры применения МСО для решения геологических задач прогнозно-поискового профиля.

Пример I. Выяснение меры сходств упорядоченности месторождений нефти по значениям целевого признака и по числовым мерам нагрузок строк

Решались таблицы, в которых строками были заданы месторождения нефти Аравийской платформы, а столбцами-характеристические признаки этих месторождений [39]. Расчет нагрузок (по нецентрированному варианту решения) строк и столбцов произведен для учтенного объема информации в задаче сравнительного изучения нефтяных месторождений мира. В соответствии с целями обработки были взяты 22 таблицы (объемом, в среднем, $m \times n = 11 \times 25$). Вся совокупность характеристических признаков была подразделена на пять групп (см. табл. I).

Исследуемая совокупность объектов предварительно упорядочивается в соответствии с критерием значимости месторождений (запасами). В поисках совокупности нефтепроизводящих и нефтесохраняющих признаков важно выяснить вопрос: "Какая из выделенных групп признаков упорядочивает месторождения в наибольшей близости к порядку месторождений по запасам?"

Для сравнения меры сходства порядка строк таблицы использованы результаты тестового подхода и МСО. Меры сходства устанавливались по соотношению Спирмена и парной корреляции. Результаты оценки приведены в таблице I , где мера порядка месторождений по запасам и числовым значениям π и $J(S)$ приводится для группы месторождений на Аравийской платформе.

Меры сходств строк таблицы, поскольку они хорошо коррелируют с разведанными запасами нефти, представляют важный критерий при прогнозировании поисков нефтяных месторождений. Например, они могут применяться для упорядочения малоизученных площадей по их перспективе на обнаружение нефти. Следует отметить совпадение основных результатов по тестовому подходу и по МСО.

Таблица I

Группа признаков	Обозначение	по Спирмену		по коэффициентам корреляции	
		по нагрузкам π	по J(S)	по нагрузкам π	по J(S)
Нефтеносная свита	T_I	+0,I6	-0,02	+0,2I8	+0,I66
Поднефтеносная свита	T_2	+0,I5	-0,I6	-0,I58	-0,304
Наднефтеносная свита	T_3	-0,29	-0,34	-0,423	-0,347
Геотектоническая обстановка	T_4	+0,80	+0,5I	+0,554	+0,392
Структурная ловушка	T_5	+0,74	+0,79	+0,864	+0,93I

пороги: 0,0I-0,735; 0,02-0,685
0,05-0,602; 0,I -0,52I

Меры сходства четко выражены по группам признаков, характеризующих структурные ловушки и геотектонические обстановки, и практически не выражены по другим группам признаков. Это вполне согласуется с хорошо обоснованными генетическими представлениями. Ведь ловушка определяет возможный размер и характер залежи. В признаки геотектонической обстановки включены данные по объемам осадочных пород, принимавших участие в нефтеобразовании. Относительно нефтеносной свиты можно утверждать, что ее характер одинаков и на больших и на малых площадях. А данные по объемам пород нефтеносной свиты приведены в комплекс свойств геотектонической обстановки. Попутным результатом решения этой задачи были получены совокупности признаков, которые наиболее тесно увязаны с причинами упорядочения месторождений по важности.

Пример 2. Сравнительное изучение совокупности оловорудных месторождений Приморья с целью выяснения перспективы ряда рудоносных районов

В развитие работ [26,I4] с помощью МСО проведено расследование информации по регионам, перспективным на обнаружение

месторождений. Информационный массив данных был подготовлен группой геологов Приморья под руководством Константинова Р.М. (ИГЕМ АН СССР).

Объекты исследования представлены несколькими группами. Выделение групп осуществлено с учетом масштабов оруденения. Первую группу составляют наиболее крупные месторождения (m = 4); вторую группу образуют месторождения, средние по масштабу (m = 6); в третью группу включены мелкие месторождения (m = I8). Четвертую группу образуют объекты, подлежащие сортировке по трем предыдущим группам, поскольку убедительной принадлежности каждого представителя этой группы к трем промышленным группам месторождений не установлено. В дальнейшем этим четырем группам присваивается ранг класса: I–Ш классы-объекты-эталоны и IУ класс объединяет в себе объекты-пробы (подлежащие распознаванию). Отметим, что согласно экспертной оценке геологов часть из экзаменуемых объектов IУ класса более перспективна на обнаружение промышленных руд, а часть менее перспективна. Но по более перспективным объектам нет убедительной экспертизы на предмет того, какой из объектов подлежит первоочередному опоискованию.

Характеристические признаки исследуемых объектов представлены списком (n = I67). В общем список подразделен на две части:

а) признаки, освещающие тектонику региона и отдельных объектов;

б) признаки, характеризующие вещественный состав исследуемых объектов.

Более детальные подразделения пространства признаков по свойствам их родства представлены решением . Примеры формулировок значений признаков тектонического характера:

№5 – угол падения крыла складки до 45^{o};

№I4 – наличие флексур;

.
.
.

№77 – падение южное, угол $\geqslant 75^{o}$.

Примеры формулировок значений признаков вещественного состава:

№88 – биотитовые граниты в удалении до 5 км;

№I00 – дайки кислого состава;

№167 - карбонатная минерализация и минеральная ассоциация.

Отметим, что детальность и количество признаков, характеризующих тектоническую обстановку, выше (n = 98), чем совокупность признаков, характеризующих вещественный состав.

Геологическая постановка задачи

Постановка задачи подразделяется на производственный и теоретический уклоны.

По производственному руслу решения задачи важно ответить на такие вопросы:

1. Подразделение пространства признаков на группы, узко характеризующие каждый из выделенных продуктивных классов (классы эталонов);

2. Выделение группы признаков максимально различающих классы эталонных объектов;

3. Выделение группы признаков, наиболее тесно коррелирующих с запасами месторождений;

4. Указание группы объектов в качестве первоочередных для поиска и разведки из IV класса объектов проб.

Именно в связи с вопросом 4) возникает сложная теоретическая проблема о механизме рудообразования и генетических процессах. Предусматривая получение результатов теоретического плана, общую совокупность признаков можно подразделить на:

- тектоническую, магматическую группы и группу признаков ведущих минеральных ассоциаций. Для более детального рассмотрения геологических процессов, имевших место на изучаемой территории, требуется выяснение значимости каждой из родственно связанных подгрупп признаков. Выявление количества и структуры взаимосвязей подгрупп признаков позволило внести последовательность шагов решения задачи на алгоритмическом этапе. Всего выделено 13 подгрупп признаков, число которых колеблется от 7 до 18. В процессе постановки геологической задачи была проведена предварительная обработка информации по уточнению значений признаков для 8-ми объектов исследования.

Формализованная постановка задачи

Обработка таблиц данных, общим объемом в 63 строки и I67 столбцов, проводилась в соответствии с основными целями этого изучения, сформулированными в геологической постановке задачи, по следующей схеме (рис. I2).

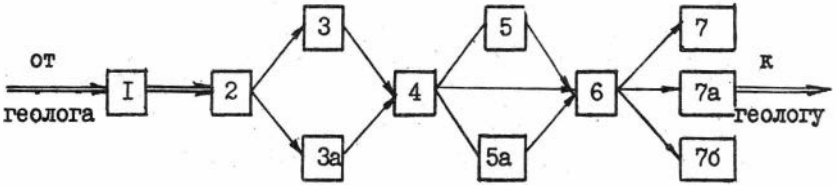


Рис. I2

I – мобилизация исходной информации и ее предварительная обработка (уточнение, кодирование, составление таблиц); 2 – геологическая постановка задачи; 3 – производственный уклон задачи; 3а – теоретический уклон задачи; 4 – формализованная постановка (подразделение на подгруппы, выделение таблиц, выявление шагов решения); 5 – общее решение таблицы; 5а – решение по частям, согласно подразделению признаков по весам; 6 – процедуры сортировки объектов-проб; 7 – результаты по признакам; 7а – результаты по группам признаков; 7б – результаты сортировки объектов; 7в – выделение первоочередных объектов-проб.

Конкретные шаги решения задачи проводились в два этапа:

а) этап решения общей таблицы исходных данных в полном объеме для целей ранжировки объектов и целевой минимизации признаков (ранжирования по содержательным подгруппам и выделения наиболее информативного подмножества);

б) этап решения подтаблиц, организованных в соответствии с заданными классами для сортировки объектов проб.

Отметим, что теоретическая совокупность вопросов для своего ответа нуждается в результатах решения обоих этапов.

Решение задачи и обсуждение некоторых результатов

Обсуждение результатов производится в соответствии с конкретными этапами и шагами решения.

а) Решение таблицы в полном объеме

По программе [22] таблица (состоящая из нулей и единиц) оловорудных объектов объемом 63×137 была подвергнута счету на БЭСМ-6. Результатами этого счета явились количественные оценки каждого столбца (признака) и каждой строки (объекта) таблицы. Эти количественные оценки представили не только интерес в плане интерпретации признакового пространства, но и легли в основу организации подтаблиц для второго этапа-б).

Расположение строк таблицы по значениям нагрузок показало, что в первый ранг значимости объектов вошли месторождения первого и второго классов (из II класса один объект выпал во второй ранг значимости). В 27 объектов второго ранга значимости вошло 5 объектов третьего класса. На 10 объектов третьего ранга значимости пришлось 2 объекта III класса, остальные-из IУ класса. Попадание наиболее существенных месторождений в первый ранг значимости объектов по оценкам нагрузок строк МСО иллюстрирует как правильность организации информации, так и пригодность итерационных процедур для ранжирования объектов в корреляции с целевым признаком.

Расположение столбцов по нагрузкам, согласно убыванию их значений, обнаруживает также ранжирование признаков на группы Выявлено таким образом 6 рангов, на которые разделилось пространство характеристических признаков. Характерно, что меньшему рангу значимости соответствует большее число составляющих его признаков. В данной таблице отмечается для одних столбцов высокий уровень корреляции (по коэффициенту Спирмена) значений нагрузок с числом единиц в столбце ($\rho_i = 0,92$), в других же случаях эти корреляции весьма низки, по существу отсутствуют ($\rho_i = 0,16$). С интерпретационных позиций следует отметить, что признакам бо-

лее высоких рангов соответствуют наиболее существенные характеристики месторождений. Минимизационной процедурой (вычленение признаков высших рангов) для диагностических целей на втором этапе решения было отобрано 44 признака.

В теоретическом отношении, в связи с вопросом выделения рудоконтролирующих комплексов признаков (поисковых), оказалось необходимым построить дерево численно оцененных признаков-столбцов с задачей установить важность тех или иных ветвей родственно сжатых признаков. "Взвешивание" ветвей дерева производилось усреднениями значений нагрузок по группам и подгруппам признаков. Наиболее общие единицы профессионального подразделения признакового пространства на тектонические и вещественные не обнаружили взаимного превосходства, поскольку группа "Тектоника" имеет усредненный вес вершины 0,325; а группа "вещественный состав" - 0,324.

Содержательно важным для интерпретации фактов явилось обнаружение значительного информационного превосходства подгруппы признаков "пликативные структуры" (усредненная нагрузка вершины 0,407) над подгруппой признаков "дизъюнктивные структуры" (усредненная нагрузка вершины 0,244). Этот факт полезен не только в смысле ориентации поиска и организации комплекса поисковых признаков, но и наталкивает на дополнительные гипотезы относительно рудогенеза исследуемых месторождений.

При оценке группы признаков "вещественный состав" обнаруживается высокое информационное значение подгруппы "комплекс вмещающих пород" (усредненная нагрузка вершины 0,401) и невысокое (0,266) для подгруппы признаков "характер" магматизма. При переходе к более дробному подразделению признакового пространства обнаруживается такая последовательность значимости подгрупп (см. табл. 2).

Таблица 2

№ п/п	Название подгрупп	Усреднение значения нагрузок вершины подгруппы
I	Вмещающие породы	0,513
2	Высшие порядки пликативных структур	0,432

3	Вторичные изменения	0,419
4	Дайки кислого состава	0,418
5	Пликативные структуры П порядка	0,381
6	Характеристики руды	0,348
7	С-З крупные дизъюнктивные структуры	0,335
8	СВ крупные дизъюнктивные структуры	0,276
9	СВ крупные дизъюнктивные структуры	0,270
I0	Минеральные ассоциации в рудных полях	0,264
II	Субмеридиональные крупные дизъюнктивные структуры	0,237
I2	Субширотные крупные дизъюнктивные структуры	0,I27
I3	Штоки, штокообразные интрузии	0,II4

Результаты этого этапа решения дали возможность в минимальное число шагов реализовать цели второго этапа решения.

б) Сортировка объектов-проб.

Исследование месторождений процедурами сравнительного изучения их описаний с помощью МСО на втором этапе началось составлением подтаблиц решения в соответствии с заданными классами. Три выделенных класса месторождений неравнозначны по числу объектов и неравноценны по масштабу запасов. Причем масштаб запасов объектов I класса гораздо выше в I классе, чем во П. Этот межклассовый разрыв в запасах между П и Ш классами намного меньше, так что объекты этих классов оказываются сильно сближенными по значениям целевого признака. Этот факт важен в том отношении, что поисковые комплексы признаков для объектов П и Ш классов по существу одинаковы.

На шаге распознавания при сортировке объектов-проб классы месторождений были переорганизованы в связи с их количественной и качественной несопоставимостью. Для того, чтобы в первом приближении выявить принадлежность проб к промышленным месторождениям необходима проверка диагностической способности всех и каждого признака. По мере выявления размера этой способности признаки

урезались до поисковой группы, наиболее важной для исследования сходств и различий организованных классов.

Процедура сортировки объектов-проб производилась на следующим образом организованных классах. Из 4-х объектов первого класса был изъят недоохарактеризованный объект, а три оставшиеся объекта были включены в состав второго класса. Таким образом, диагностическая таблица T_I состояла из 9 объектов, наиболее существенных в промышленном отношении. В связи с тем, что объекты II класса сближены с объектами III класса, от группы поисковых признаков требовалась высокая сортировочная способность. В связи с тем, что процедура сортировки фиксирована, основное внимание по улучшению распознавания было перенесено на модификацию состава признаков в поисковом комплексе. Выбирались те признаки, которые хорошо различали таблицы T_I и T_2 (таблица T_2 состояла из 10 месторождений III класса); 8 месторождений из III класса принимались в качестве обучения и внутреннего экзамена [14].

Сортировка объектов-проб проводилась в три шага.

На первом шаге 8 месторождений третьего класса распознавались по всем 167 признакам общего пространства признаков. Результат сортировки - из 8 объектов к своему классу было отнесено только 3 объекта (к T_2), пять объектов было ошибочно отнесено к T_I. На этом же шаге были проведены сортировочные операции и для всех объектов-проб; путем вычисления коэффициентов принадлежности к классу все пробы были разнесены по двум классам, заданным T_I и T_2.

На втором шаге, на подпространстве признаков с максимальными нагрузками (признаки высших рангов, $n = 44$). результаты сортировки объектов экзамена (8 месторождений III класса) значительно улучшились. Один объект попал в "чужой" класс и еще один попал в зону отказа, т.е. объект на заданных классах не распознавался.

На третьем шаге решения были выделены поисковые признаки, по которым таблицы T_I и T_2 наиболее различались. Результаты сортировки улучшились, из 8 объектов 7 попали в "свой" класс и один объект получил отказ (тот, который попал в "чужой" класс на предыдущем шаге). Объекты-пробы ($n = 35$) были все подвергнуты сортировке (по этому поисковому комплексу признаков) на принадлежность к T_I и T_2. Из совокупности прогнозируемых ру-

допроявлений к объектам промышленного значения было отнесено всего 4 объекта, I9 объектов отнесено ко П классу (T_2) и I2 получили отказ.

После дополнительной проверки информации и решений 3 объекта (из 4-х попавших в класс промышленных объектов) были рекомендованы к опоискованию и оценочному бурению.

Таким образом, в задаче изучения совокупности оловорудных месторождений Приморья и разбраковки площадей, перспективных по косвенным данным на обнаружение промышленных объектов, с помощью МСО были получены результаты практического и теоретического значения.

I. Выделены существенные группы признаков в поисковом отношении: вмещающие породы ⟶ высшие порядки пликативных структур ⟶ вторичные изменения ⟶ дайки кислого состава;

2. Выделена группа из 36 признаков, имеющая поисковое и диагностическое значение;

3. Указаны 3 объекта, перспективных в промышленном отношении.

Тесный контакт с геологами осуществляет коррекцию хода решения задач, эта коррекция в процессе данного прогноза осуществлялась дважды. Характерно, что коррекция взаимно дополнительна: один раз заказчику была указана некорректность во включении одного из объектов в обучающую выборку и один раз геологи дали дополнительную информацию по существенному улучшению пространства признаков (после одного из шагов решения).

Пример 3. Оценка перспективности площадей Норильского и сопредельных районов на медно-никелевое оруденение (этап обучения)

В последние годы норильскими геологами опытно-методической экспедиции НПО "Севморгео" разрабатывается направление по оценке перспективности площадей северо-запада Сибирской платформы на медно-никелевое оруденение с использованием статистических методов распознавания [37,38]. В результате собран и систематизирован огромный материал по обширной площади и выданы конкретные рекомендации.

В развитие этих работ, в рамках договора о сотрудничестве, проводятся логико-математические исследования по оценке перспек-

тив площадей Норильского и сопредельных районов. Здесь приведены положительные результаты решения задачи разделения известных (эталонных) площадей по структурно-тектонической, метаморфо-метасоматической и минералогической группам признаков с использованием метода согласованных оценок [34,24] .

Объекты исследования представлены элементарными участками (8-10км^2), за которые были приняты листы крупномасштабной карты. В свою очередь участки объединены в классы. Первый класс составляют площади, содержащие медно-никелевые месторождения (десять эталонных участков), второй класс - участки с неперспективными рудопроявлениями и без рудопроявлений (пятьдесят шесть объектов).

Эталонные объекты охарактеризованы тремя группами признаков. Группа структурно-тектонических признаков подразделена на три части: восемь характеристик пликативных структур, десять - дизъюнктивных и пять - инъективных дислокаций. Группа метаморфо-метасоматических признаков состоит из двенадцати характеристик преобразования пород и трех параметров формы и масштабности этих преобразований. Группа минералогических признаков представлена девятью характеристиками состава породообразующих минералов и девятью рудными минералами. Список приведен в работе Л.Г.Сухова и др. [38].

Геологическая задача заключается в выявлении информативности признаков в каждой из трех перечисленных групп с целью разделения двух классов эталонных участков. Таким образом, геологическая задача решалась в трех постановках: в первой - обрабатывалась таблица из 66 строк и из 23 столбцов, во второй, соответственно, 66 × I5 и в третьей - 66 × I8. Таблицы исходных данных приведены в работе Л.Г.Сухова и др. [38]. Рассмотрим результаты решения по постановкам.

Первая задача. В результате решения получены количественные оценки для каждого признака и для каждого объекта таблицы. Достигнуто полное разделение объектов первого класса от второго по структурно-тектоническим признакам (рис. I6).

Расположение строк таблицы по значениям нагрузок (рис. I3) показало, что первый класс достаточно однороден, т.е. его объекты имеют близкие структурно-тектонические характеристики. В

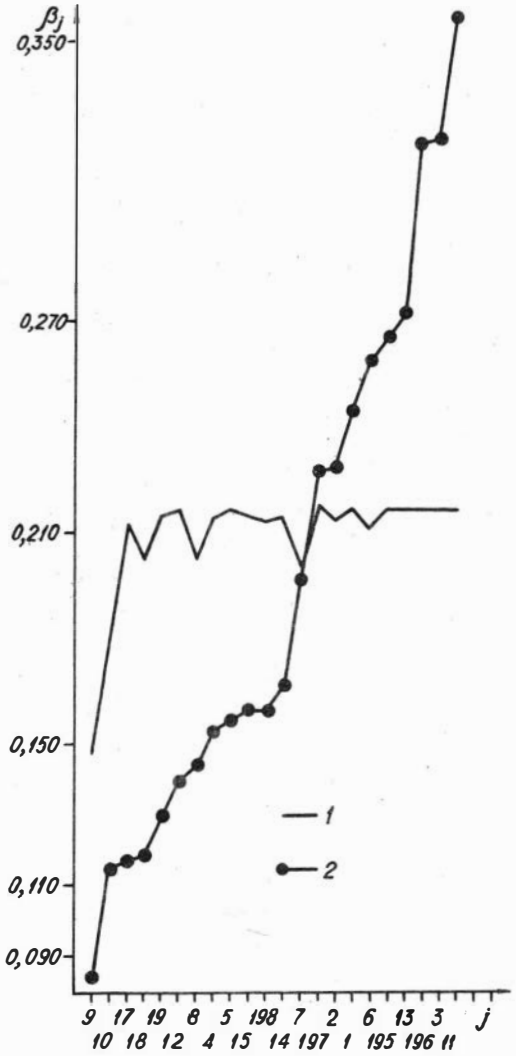


Рис. I3. Ранжирование структурно-тектонических признаков по классам: I – продуктивных объектов; 2 – непродуктивных объектов

связи с этим признаки в классе по информативности мало отличаются друг от друга. Исключение составляет седьмой объект (участок южнее Норильск-I), который отличается от других площадей отсутствием флексур и инъекций вещества осадочного чехла. Что же касается второго класса, то он является неоднородным по структурно-тектоническим характеристикам. В связи с чем, признаки в классе по информативности более дифференцированные (рис. I3). Объекты второго класса по значениям нагрузок строк условно можно разбить на три ранга значимости. Такое ранжирование объектов второго класса в отличие от объектов первого можно объяснить не только разницей в структурно-тектонической обстановке. В первом классе объекты имеют определенную "специализацию" и территориально близко расположены друг около друга. Во второй класс объекты выбирались с помощью таблиц случайных чисел из множества участков, отвечающих указанным условиям [37] , и территориально сильно разобщены между собой.

По характеру распределения нагрузок столбцов (рис. I3) структурно-тектонические признаки условно можно ранжировать на три группы. В первую группу попадают признаки, информативность которых уменьшается при переходе от объектов первого класса к объектам второго. Приведем примеры наиболее ярких представителей этой группы. Характеристики пликативных структур - расположение в пределах областей замыкания положительных структур четвертого порядка, площадью 10^2 - 10^4 кв.км, поперечником более 10 км (пятый признак), и наличие флексур (восьмой признак). Эти признаки выполнены на площадях, содержащих месторождения, на объектах второго класса они, как правило, отсутствуют. Характеристики инъективных дислокаций - средняя площадь интрузивов, т.е. суммарная площадь интрузивов в кв.км на их количество (I7-й признак) и наличие инъекций вещества осадочного чехла (I8-й признак). Во втором классе по сравнению с первым увеличивается средняя площадь интрузивов и отсутствуют инъекции вещества осадочного чехла.

Во вторую группу попадают признаки, имеющие увеличение информативности при переходе от объектов первого класса к объектам второго класса. Из характеристик пликативных структур это расположение в пределах крыльев структур четвертого порядка (третий признак). По дизъюнктивным структурам - наличие систем криволинейно-концентрических разрывов (2I-й признак). Данные характерис-

тики выполнены на всех объектах первого класса, во втором классе они встречаются только в половине случаев.

Остальные признаки образуют третью группу с неизменяющимися информационными весами.

Вторая задача. Здесь также получены численные оценки для каждого признака и для каждого объекта таблицы и достигнуто полное разделение объектов первого класса с объектами второго, но уже по метаморфическим и метасоматическим признакам. При упорядочивании строк таблицы по значениям нагрузок первый класс отчетливо подразделяется на две группы объектов. В первую входят три объекта: северо-западная ветвь Верхне-Талнахской интрузии, г.Зуб и участок южнее Норильск-I. Остальные объекты образуют однород - ную группу, т.е. имеют близкие метаморфо- метасоматические ха - рактеристики. Отличие первых трех объектов от остальных заклю - чается в следующем: на северо-западном участке отсутствует изме- ненность базальтов (I07-й признак); на площади г. Зуб отсутст - вует измененность базальтов и комплекс низкотемпературных гидро- термальных околотрещинных изменений (I06-й признак); на участке южнее Норильск-I отсутствуют кремнещелочные новообразования (98-й признак) и магнезиальные скарны (99-й признак)(рис. I4,I6) .

Для второго класса также как и в первой задаче характерна большая неоднородность. Здесь есть группа объектов (30% от обще- го числа в классе), у которых отсутствуют все метаморфо- метасо- матические признаки, кроме одного - метаморфизма углей. Связано это с отсутствием интрузивных тел на данных площадях. Есть груп- па объектов, содержащая около половины признаков и группа - про- межуточная между первыми двумя. По двум количественным признакам оба класса также отличаются достаточно контрастно. Во втором классе наблюдается больший метаморфизм углей и меньшая мощность зон метаморфо- метасоматических преобразований.

Третья задача. В результате решения получено полное разде- ление объектов первого и второго классов по минералогическим при- знакам на основе вычисления оценок каждого объекта и каждого при- знака таблицы (рис. I5,I6).

При расположении строк таблицы по их нагрузкам заметно под - разделение объектов первого класса на 3 группы. В первую группу попадают интрузии: Норильск-I, II, г. Зуб, г. Черная, Нижне-Тал- нахская. Они имеют одинаковые минералогические характеристики в

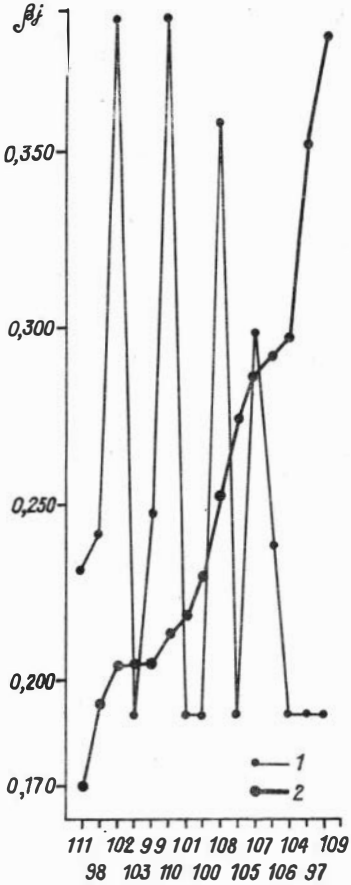


Рис.I4. Ранжирование метаморфо-метасоматических признаков по классам: I-продуктивных объектов;2-непродуктивных объектов

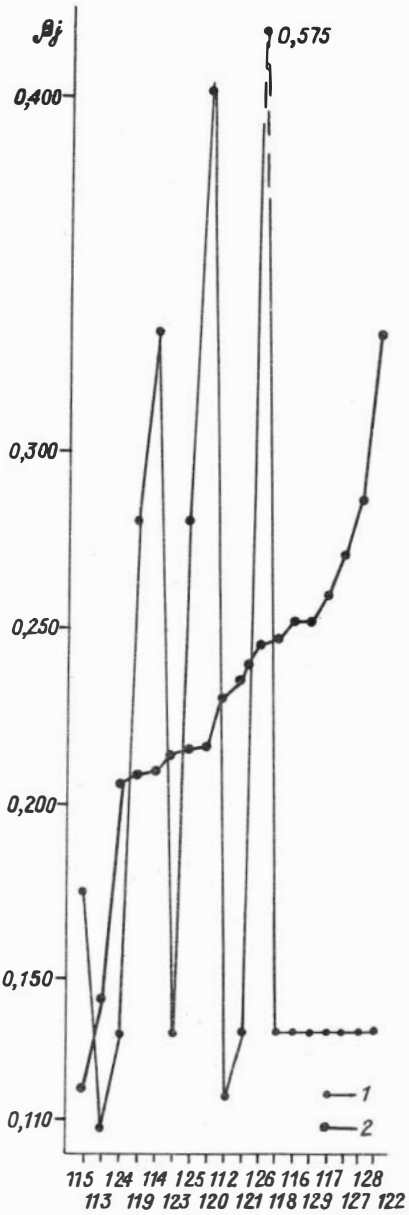


Рис.I5. Ранжирование минералогических признаков по классам: I-продуктивных объектов; 2 - непродуктивных объектов

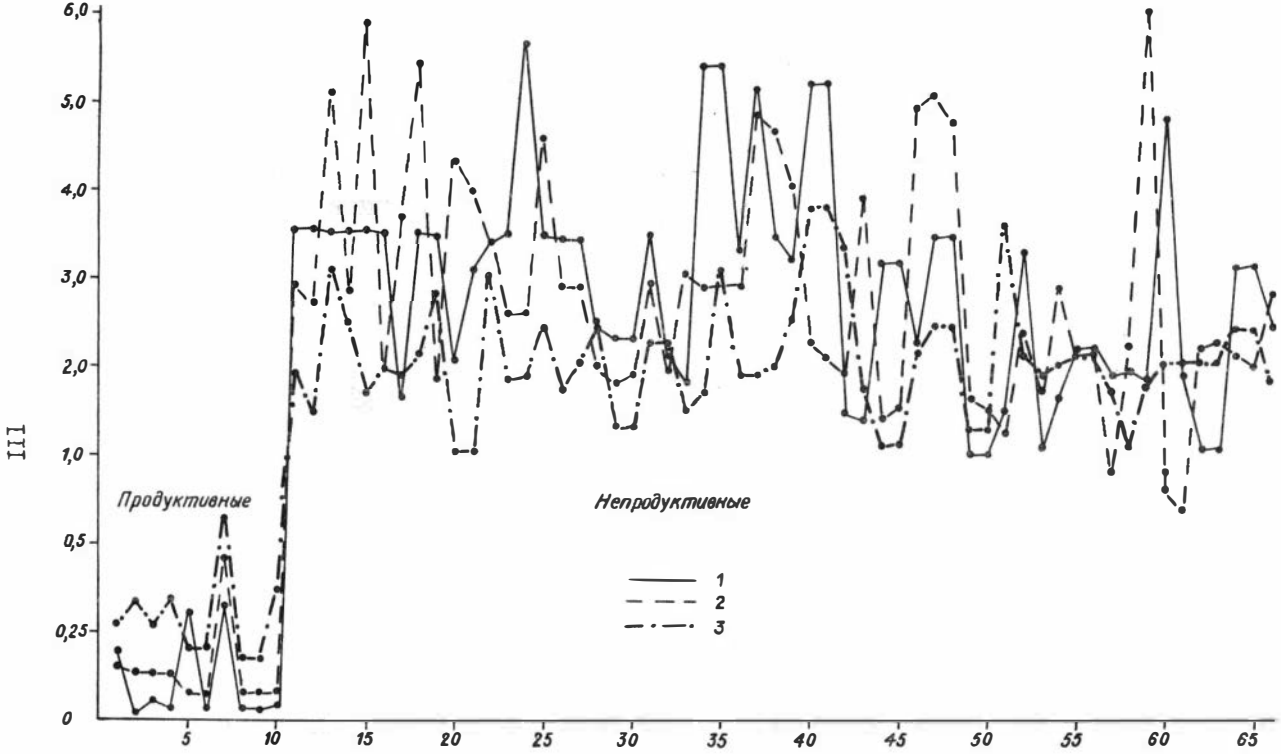


Рис.16. Разделение продуктивных объектов от непродуктивных по признакам:
I - структурно-тектоническим; 2 - метаморфо-соматическим; 3 - минералогическим

пятнадцати случаях из восемнадцати. Во вторую группу входят все четыре ветви Верхне-Талнахской интрузии, имея сходство почти по всем минералам. Верхне-Талнахская интрузия отличается от первой группы меньшей основностью плагиоклаза долеритов, большей магнезиальностью оливинов и наличием бор- или фторсодержащих минералов. Особняком от первых двух групп стоит объект (7) - южнее Норильск-I. Основное его отличие - отсутствие хромита и халькозина в коренных породах или шлихах.

В связи с такой неоднородностью класса признаки имеют широкий диапазон информативности от 0,108 до 0,575.

Объекты второго класса по нагрузкам таковы, что сам класс представляется однородным. В связи с этим и признаки по информативности мало отличаются друг от друга.

Характер поведения информативности столбцов при переходе от одного класса к другому резко меняется почти по всем минералам: либо наблюдается увеличение информативности, либо ее уменьшение. Такое изменение происходит из-за значительных различий в минералогической структуре классов.

В заключение отметим ряд моментов, сравнивая результаты решения рассматриваемой задачи по методу согласованных оценок и методам математической статистики [37,38].

Результаты проведенных исследований показали возможность и целесообразность использования центрированного варианта метода согласованных оценок для разделения перспективных от неперспективных на медно-никелевое оруденение площадей по отдельным группам признаков. Полученные результаты на этапе обучения могут в дальнейшем использоваться для распознавания неизвестных участков-проб.

В отличии от решения методами математической статистики [37,38] , где оно осуществлялось в два алгоритмически несвязанных этапа, здесь решение задачи разделения проводилось сразу без предварительного свертывания информации методом главных компонент, что значительно упростило и сам ход решения, и возможность понимания полученных результатов.

Особенность центрированного варианта метода согласованных оценок заключается в том, что решение задачи разделения двух классов площадей проводится на основе полученной групповой информативности признаков. По метаморфо-сомотической и минералогической группам результаты совпали с работой [37,38] .

§IO МЕТОД ГЛАВНЫХ КОМПОНЕНТ

(Краткий очерк)

Дается изложение метода главных компонент (МГК) с простыми доказательствами известных результатов. Предлагаемый алгоритм представляет собой последовательное решение однотипных задач на отыскание главных собственных векторов, что соответствует методу исчерпывания в линейной алгебре. Показана связь МГК с факторным разложением таблицы данных.

I. Исходные предпосылки

Пусть X – таблица экспериментальных данных размера $m \times n$; строки X суть объекты, столбцы-признаки этих объектов. Таблице X соответствует выборка X объема m из некоторой совокупности, характеризуемой n признаками. Элемент x_{ik} есть значение k-го признака на i-м объекте. Обычно предполагается, что данные X – стандартизованные (центрированные и нормированные): средние по столбцам равны нулю, средние квадратические – единице. Но большинство приводимых ниже рассуждений сохраняет силу также и для нестандартизованных данных.

Метод главных компонент (МГК) описан во многих руководствах по математической статистике (см., например [I]). Вычисляемые по МГК величины обладают многими интересными свойствами, полезными для приложений ([3I, 52]).

Наиболее распространенная схема изложения МГК состоит в следующем. Сначала ищется вектор c^1, такой, что проекция на него $y^{(1)} = Xc^1$ обладает максимальным разбросом. Затем ищется другой вектор c^2, такой, что проекция $y^{(2)} = Xc^2$ имеет максимальный разброс и при этом величина $y^{(2)}$ некоррелирована с $y^{(1)}$. И так далее, требуя каждый раз, чтобы величина $y^{(i)} = Xc^{(i)}$ была некоррелирована со всеми величинами $y^{(1)}, \ldots, y^{(i-1)}$, полученными на предыдущих этапах. Таким образом, на каждом следующем этапе добавляется новое ограничительное условие. В случае

стандартизованных данных векторы $c^{(1)}, \ldots, c^{(n)}$ задают направление главных осей эллипсоида инерции выборки X.

Дир (*Dear*) показал связь МГК с факторным разложением простейшего вида, когда X представляется как произведение столбца на строку (см. ниже); выкладки Дира приведены в работе[53]. Дарроч (*Darroch*) доказал теорему, вследствие которой МГК можно рассматривать как средство получения факторного разложения таблицы данных с минимальной нормой погрешности ([48]). Результат Дира следует из теоремы Дарроча как частный случай.

Чтобы проследить связь между МГК и факторными разложениями более детально, мы дадим иную схему МГК, отыскивая на каждом шаге направление, в котором достигается максимальный разброс выборки остатков, получаемой из исходной выборки вычитанием наилучшей линейной оценки, основанной на результатах предыдущих этапов, или, что то же самое, минимизируя на каждом шаге норму остаточной погрешности факторного разложения простейшего вида. Алгебраические предпосылки:

Определение: $W = X^T X$ есть ковариационная матрица (к.м.) выборки X. (Заметим, что это определение не вполне совпадает с принятым в математической статистике, особенно если учесть, что мы рассматриваем не только центрированные данные X). Будем предполагать, что собственные числа (с.ч.) матрицы W попарно различные; ввиду симметричности W все с.ч. вещественные. Примем также, что ранг W равен n. Занумеруем с.ч. в убывающем порядке: $\lambda_1, \ldots, \lambda_n$; в том же порядке занумеруем собственные векторы (с.в.) матрицы W: $c^1, \ldots, c^n$. Пусть $L = \mathrm{diag}\,\lambda$ есть диагональная матрица с.ч.; C – матрица, столбцами которой являются с.в. c^i ($i = 1, \ldots, n$). Не теряя общности, можно считать с.в. нормированными: $c^T c = I$; тогда справедливо соотношение: $C^T C = C C^T = I$.

Основная спектральная задача МГК:

$$X^T X C = C L, \qquad C^T C = I. \tag{I}$$

Полное решение задачи (I) есть набор "собственных пар" (λ, c), или, что то же самое, (L, C). Собственную пару (с.п.) (λ_1, c^1) будем называть главным решением задачи (I). Вообще же, с.п. (λ_i, c^i) есть решение (I). Определение: Матрица $Y = X C$ есть

матрица главных компонент выборки X.

Сопряженная спектральная задача МГК:

$$X X^{T} Y = Y L \quad . \tag{2}$$

Имеет место соотношение ортогональности: $Y^{T} Y = L$, аналогичное вышеприведенному ($C^{T} C = I$).

Столбцы y^{i} матрицы Y называются главными компонентами (г. к.) выборки X. Связь между исходными признаками (столбцами матрицы X) и г.к. дается формулами (прямое и обратное преобразование МГК):

$$Y = X C \quad , \qquad X = Y C^{T} \quad . \tag{3),(4}$$

Отметим также соотношения:

$$W = C L C^{T} \quad , \qquad C^{T} W C = L \tag{5),(6}$$

Представление к.м. в виде (5) позволяет легко найти обратную матрицу: $W^{-1} = C L^{-1} C^{T}$.

2. МГК и факторное разложение первого порядка

Перейдем теперь к выводу и обоснованию алгоритма МГК.

Рассмотрим преобразование:

$$y = X c \tag{7}$$

Здесь $c = (c_1, \ldots, c_n)$ есть n-мерный нормированный вектор: $c^{T} c = I$; вектор y (размерности m) есть проекция выборки X на c. Образуем квадратичный функционал:

$$Q = c^{T} X^{T} X c \tag{8}$$

Если данные стандартизованы, то Q характеризует рассеяние выборки X в направлении c; в общем случае Q определяет момент второго порядка (проекции $X c$) относительно начала координат. Будем искать вектор c, на котором Q обращается в максимум. Учитывая нормировку $c^{T} c = I$, приходим к задаче на безусловный экстремум функционала:

$$F = c^{T} X^{T} X c - \lambda \left(\sum c_i^{2} - I \right) , \tag{9}$$

где λ – множитель Лагранжа. Необходимое условие экстремума F :

$$X^{T} X c = \lambda c \qquad (10)$$

При сделанных предположениях относительно w уравнение (10) имеет n решений (λ_i , c), где λ – с.ч., c^i – с.в. матрицы w ($i = 1, \ldots, n$). Таким образом, мы получили основную спектральную задачу (1). Домножив (10) на c^{T} (слева), найдем, что максимальное значение функционала (8) равно с.ч. λ :

$$c^{T} w c = \lambda c^{T} c = \lambda \quad , \qquad (11)$$

при этом, очевидно, следует взять наибольшее с.ч. λ_1 и с.в. c – т.е. главное решение задачи (10). Итак,

$$Q_{max} = \lambda_1 \qquad (12)$$

Подставляя в (7) $c = c^1$, получаем первую г.к. выборки X :

$$y^1 = X c^1 \qquad (13)$$

И вообще, i –я г.к. определяется по формуле: ($i = 1, \ldots, n$)

$$y^i = X c^i \qquad (14)$$

Из (14) следует:

$$(y^i)^{T} y^i = (c^i)^{T} w c^i = \lambda_i \; ; \qquad (15)$$

$$(y^i)^{T} y^j = (c^i)^{T} X^{T} X c^j = \lambda_j (c^i)^{T} \cdot c^j = 0 \quad (i \neq j) \qquad (16)$$

(свойство некоррелированности г.к. для центрированных данных).

Из (7) и (10) вытекает сопряженная спектральная задача:

$$X X^{T} y = \lambda y \qquad (17)$$

$$(X^{T} y = \lambda c \Longrightarrow X X y = \lambda X c \Longrightarrow (17)).$$

Таким образом, г.к. y есть с.в. матрицы $X X^{T}$. Подставив (7) в (10), получим:

$$X^{T} y = \lambda c \qquad (18)$$

Это означает, что (с точностью до скалярного множителя λ) c есть г.к. выборки X^{T} (признаки X рассматриваются как объекты таблицы X). Векторы c , y , соответствующие одному и тому же с.ч. λ , образуют сопряженную пару (c, λ). Объединим уравнения (7) и (8):

$$\left.\begin{aligned} y &= X c \\ \lambda c &= X^{T} y \end{aligned}\right\} \qquad (19)$$

Легко видеть, что из (I9) вытекают взаимно сопряженные задачи (I0),(I7). Известно, что матрицы X^TX и XX^T имеют одинаковый ранг и одинаковые ненулевые с.ч. Если $m > n$ и ранг X^TX равен n, то спектр XX^T содержит нулевое с.ч. кратности $m-n$ ([I3]).

Полагая $u = \lambda^{-\frac{1}{2}} y$, можно придать уравнениям (I9) более симметричную форму:

$$\begin{aligned} \lambda^{1/2} u &= X c , \\ \lambda^{1/2} c &= X^T u \end{aligned} \qquad (20)$$

Уравнения (I9),(20) возникают при нахождении простейшего факторного разложения таблицы данных с минимальной нормой погрешности:

$$X = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} (\beta_1 , \dots , \beta_n) + R = \alpha \beta^T + R \qquad (2I)$$

Такое представление назовем <u>факторным разложением</u> (таблицы X) первого порядка.

Здесь $\alpha = (\alpha_1 , \dots , \alpha_m)^T$, $\beta^T = (\beta_1 , \dots , \beta_n)$ – искомые векторы, один из которых, например β, не теряя общности, можно считать нормированным: $\beta^T\beta = I$, R есть матрица погрешностей приближенного представления:

$$X \stackrel{\wedge}{=} \alpha \beta^T \qquad (22)$$

(знак $\stackrel{\wedge}{=}$ обозначает приближенное равенство). За меру качества примем след матрицы $R^T R$:

$$\mathcal{L} = tr (R^T R) = \sum_i \sum_k (x_{ik} - \alpha_i \beta_k)^2 \qquad (23)$$

Минимум $\mathcal{L}$ по α_i, β_k достигается на векторах α, β, удовлетворяющих системе уравнений:

$$\begin{aligned} \alpha &= \frac{I}{\beta^T\beta} X\beta , \\ \beta &= \frac{I}{\alpha^T\alpha} X^T\alpha \end{aligned} \qquad (24)$$

Положив $\alpha^T\alpha = \lambda$ и учитывая, что $\beta^T\beta = I$, получим

$$\alpha = X\beta$$
$$\lambda\beta = X^T\alpha \qquad (25)$$

Сравнивая (25) и (I9), видим, что β есть с.в. матрицы X^TX, α есть с.в. матрицы $X X^T$, или, что то же самое, г.к. выборки X : $\beta = c$, $\alpha = y$. Докажем, что наименьшее значение достигается на главном решении задачи (I0). Имеем :

$$R^TR = (X - y c^T)^T(X - y c^T) = X^TX - cy^TX - X y c^T + c y^T y c^T; \qquad (26)$$

$$\mathcal{L} = tr(R^TP) = tr(X^TX) - 2\, tr(X^T y c^T) + \lambda \qquad (27)$$

Из (I9) следует:

$$tr(X^T y c^T) = \lambda tr(cc^T) = \lambda ; \qquad (28)$$

$$\mathcal{L} = tr(X^TX) - \lambda . \qquad (29)$$

Так как

$$tr(X^TX) = \sum_i \lambda_i , \qquad (30)$$

то, в силу (29), для получения наименьшего значения $\mathcal{L}$ надо положить $\lambda = \lambda_{max} = \lambda_1$. Будем иметь:

$$\mathcal{L}_{min} = \sum_{i=2}^{n} \lambda_i . \qquad (3I)$$

Следовательно, наилучшее представление вида (22) дается формулой:

$$X \hat{=} y^1 (c^1)^T \qquad (32)$$

3. Факторное разложение МГК

Для упрощения записи будем опускать индексы (номера) с.в. и г.к., если это не приводит к недоразумению.

Найдя сопряженную пару (c , y), которая максимизирует (8) и минимизирует (23), образуем матрицу первых остатков:

$$X_I = X - y c^T \quad ; \quad (X_I = R) \qquad (33)$$

Введем квадратичный функционал:

$$Q_I^I = c^T w_1 c \quad , \quad w_1 = X_1^T X_1 \qquad (34)$$

и будем искать нормированный вектор c ($c^T c = I$), на котором Q_1 обращается в максимум. Как и прежде, приходим к спектральной

задаче:

$$W_1 c = \lambda c \qquad (35)$$

и к формуле для Q^{I}_{max}

$$Q^{1}_{max} = \lambda \qquad (36)$$

Покажем, что главное решение задачи (35) есть вторая с.п. задачи (I0) ($\lambda = \lambda_2$, $c = c^2$).

Лемма. Пусть (λ_i, c^i) ($i = I, \ldots, n$) есть набор с.п. задачи (I0) с к.м. X^TX и $W_1 = (X - X c c^T)^T (X - X c c^T)$, где c – один из с.в. X^TX ; для определенности положим $c = c^K$ (индекс фиксирован; $I \le K \le n$). Тогда :

$$W_1 c^K = 0, \qquad W_1 c^i = \lambda_i c^i \qquad (i \neq K) \qquad (37)$$

Доказательство:

$$W_1 = X^TX - cc^TX^TX - X^TX\, cc^T + cc^TX^TXcc^T \qquad (38)$$

Домножим (38) справа на $c = c^K$; учитывая, что $X^TXc = \lambda c$, $c^Tc = I$, получим:

$$W_1 c = 0.$$

Домножим (38) на $c^i \neq c^K = c$; учитывая взаимную ортогональность векторов c^i и c, будем иметь:

$$W_1 c^i = \lambda c^i,$$

что и требовалось.

Таким образом, ненулевая часть спектра W_1 состоит из с.ч. W, за исключением λ_K.

Поэтому в формуле (36) $\lambda = \lambda_2$:

$$Q^{1}_{max} = \lambda_2 \qquad (39)$$

Матрицу S -х остатков X_S определим индуктивно:

$$X_S = X_{S-1} - X_{S-1}\, cc^T ; \quad (S = I, \ldots, n); \; X_0 = X \qquad (40)$$

причем вектор c есть главный с.в. задачи:

$$X^T_{S-1} X_{S-1} c = \lambda c \qquad (4I)$$

По доказанной лемме, c есть s-й с.в. матрицы X^TX, а вектор $y^s = X_{s-1} c^s$ есть s-я г.к. выборки X. Сопряженная пара (c^s, y^s) обеспечивает наилучшее представление (22) для выборки X_{s-1}. При этом:

$$Q^{s}_{max} = c^T X^T_{s-1} \cdot X_{s-1} c = \lambda_s \qquad (42)$$

Остаточная погрешность:

$$\mathcal{L}^{s}_{min} = \sum_{s=1}^{n} \lambda_i \quad ; \quad \mathcal{L}^{n}_{min} = 0 \tag{43}$$

Рассмотрим последовательность задач:

$$\begin{aligned} X^{T}_{s-1} X_{s-1} c &= \lambda c \\ y &= X_{s-1} c \end{aligned} \tag{44}$$

Взяв при каждом s решение, отвечающее наибольшему с.ч., получим набор сопряженных пар (c^s, y^s), определяющих преобразование МГК (3): векторы c^s, y^s суть столбцы матриц C, Y соответственно. Таким образом, все г.к. выборки X получаются как решение однотипных задач, каждая из которых в основном состоит в нахождении главного с.в. симметричной матрицы. Этот вычислительный процесс аналогичен известному "методу исчерпывания" ([45]).

Решив ρ задач вида (44) при $s = I, \ldots, \rho$, $(\rho \leq n)$, получим ρ сопряженных пар (c^s, y^s), определяющих представление выборки X в виде

$$X \hat{=} Y_\rho C_\rho^{T} \quad , \tag{45}$$

где Y_ρ, C_ρ – матрицы, столбцами которых являются г.к. y^s и с.в. c^s соответственно ($s = I, \ldots, \rho$). При $\rho = n$ имеем: $Y_n = Y$, $C_n = C$, и равенство (45) становится точным. Частный случай $= I$ впервые рассмотрен Диром (Dear) в связи с оцениванием пропущенных данных (см. [53]).

Формула (45) представляет собой факторное разложение МГК ρ-го порядка.

Если $\rho < n$, мы имеем "сжатие информации", поскольку Y_ρ представляет собой приближенное описание массива X с меньшим количеством параметров-признаков. Отбрасывание последних г.к. (при использовании Y_ρ, C_ρ вместо Y, C) реализует, кроме того, неявную фильтрацию массива данных, ибо г.к. с большими номерами более подвержены влиянию помех.

Рассмотрим факторное разложение общего вида:

$$X = A_\rho B_\rho + R \; ; \quad X \hat{=} A_\rho B_\rho \quad , \tag{46}$$

где A_ρ и B_ρ имеют размеры $m \times \rho$ и $\rho \times n$ $(\rho \leq n)$.

Специальные ограничения, накладываемые на A_ρ, B_ρ, R, приводят к различным моделям факторного анализа ([I]). Столбцы

A_ρ – признаки в новой системе координат; элементы B_ρ суть "факторные нагрузки". Если $A_\rho = Y_\rho$, $B_\rho = C^T_\rho$, мы получаем представление МГК (45) с погрешностью (43).

Последовательная процедура синтеза представления МГК (45) на основе главных решений задач вида (44) оптимальна, по критерию минимума $\mathcal{L}$ (23), на каждом шаге. Возникает вопрос: приводит ли этот процесс к глобальному оптимуму, т.е. к наилучшей по всем A_ρ, B_ρ аппроксимации (46) исходной таблицы данных? Положительный ответ дается теоремой, полученной, в несколько различных вариантах, Рао ([52]) и Даррочем ([48]). Приводимая ниже формулировка ближе к варианту Дарроча.

Пусть $\mathcal{L} = tr(R^T R)$. (Заметим, что $tr(R^T R) = \|R\|^2$, где $\|R\|$ – эвклидова норма R).

Теорема. Минимум $\mathcal{L}$ достигается, если $A_\rho = Y$, $B_\rho = C^T$; $\mathcal{L}_{min} = \sum_{i=\rho+1}^{n} \lambda_i$ (λ_s, y^s, c^s ($s = I, \ldots, \rho$)) соответствуют спектральной задаче (I) с матрицей X^TX; ρ фиксировано). Индекс ρ опустим, для упрощения записи.

Доказательство. Сделаем предварительное замечание. Любое представление вида (46) можно заменить эквивалентным (т.е. имеющим такую же матрицу погрешностей R): $X = \tilde{A}\tilde{B} + R$, $\tilde{A} = AD^{-1}$, $\tilde{B} = BD$, где D – произвольная невырожденная матрица порядка $\rho \times \rho$. Как известно, D можно выбрать так, что строки $\tilde{B}$ будут взаимно ортогональными. Поэтому, не теряя общности, можно с самого начала принять, что $BB^T = I$.

Дифференцирование $\mathcal{L}$ по элементам A (B) при фиксированных элементах B (A) и приравнивание производных нулю дает систему:

$$XB^T = ABB^T ; \qquad (47)$$

$$X^TA = B^TA^TA \qquad (48)$$

В силу (48), $X^TAB = B^TA^TAB$; поэтому.

$$\mathcal{L} = tr(X^TX - 2X^TAB + B^TA^TAB) = tr(X^TX) - tr(B^TA^TAB) \qquad (49)$$

Учитывая сделанное замечание, приведем (49) к более удобному виду. Воспользуемся также тем, что для произвольной матрицы D : $tr(DD^T) = tr(D^TD)$.

Положим $BB^T = I$; тогда (47) примет вид $A = XB^T$. Имеем:

$$tr(B^TA^TAB) = tr(ABB^TA^T) = tr(AA^T) = tr(A^TA) = tr(BX^TXB^T)$$

$$\mathcal{L} = tr(X^TX) - tr(BX^TXB^T) \qquad (50)$$

Обозначение: $g = tr(BX^TXB^T)$. Легко проверить, что

$$g = \|X\beta^1\|^2 + \ldots + \|X\beta^\rho\|^2 \qquad (51)$$

($(\beta^s)^T$ – строки матрицы B). Проделав элементарные выкладки (с учетом нормировки $\beta\beta^T = I$), можно убедиться, что g обращается в максимум на векторах β^s, удовлетворяющих уравнениям $X^TX\beta^s = \lambda_s\beta^s$, $s = 1, \ldots, \rho$, что согласуется с ограничением $BB^T = I$. Следовательно, $g_{max} = \sum_{s=1}^{\rho}\lambda_s$, и

$$\mathcal{L}_{min} = tr(X^TX) - g_{max} = \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{\rho}\lambda_i = \sum_{i=\rho+1}^{n}\lambda_i \qquad (52)$$

и этот минимум достигается на матрицах $B^T = C_\rho$, $A = XB^T = XC_\rho = Y_\rho$. Теорема доказана. Она может быть получена как простое следствие теоремы Дарроча ([48]), имеющей более сложное доказательство.

Таким образом, Y_ρ есть наилучшее приближение к таблице X среди всех матриц ранга ρ в том смысле, что элементы X аппроксимируются ρ-линейными комбинациями исходных данных (столбцами $Y_\rho C_\rho^T$) с наименьшей среднеквадратической ошибкой.

4. Оптимальный линейный предиктор

Рао получил аналогичный результат, применив несколько иной подход, представляющий самостоятельный интерес, ибо при этом подходе обнаруживается нетривиальная связь между МГК и линейной регрессией. Чтобы кратко повторить выкладки Рао, надо ввести "наилучший линейный предиктор" $\hat{X} = VB$, представляющий собой аппроксимацию "в среднем" столбцов (признаков) таблицы X столбцами V по методу наименьших квадратов. Вернемся к факторному разложению (46) и предположим, что матрица A задана: $A = V$. Тогда, в силу (48), норма погрешности $\mathcal{L} = tr(R^tR)$ обращается в минимум на матрице $B^T = X^TV(V^TV)^{-1}$, а наилучший линейный предиктор X по V определяется выражением:

$$\hat{X} = V (V^T V)^{-1} V^T X , \qquad (53)$$

что представляет собой формулу линейной регрессии X на V .

Остаточная к.м. $\hat{W}$, получаемая после вычитания $\hat{X}$ из X:

$$\hat{W} = X^T X - X^T V (V^T V)^{-1} V^T X \qquad (54)$$

Рассмотрим "сжатие информации" $V = XH$, где H – произвольная матрица ($n \times \rho$) ранга ρ . Наилучший линейный предиктор X по V :

$$\hat{X} = XH (H^T W H) H^T W \qquad (55)$$

Остаточная к.м. (после вычитания $\hat{X}$ из X):

$$\hat{W} = W - W (H^T W H) H^T W \qquad (56)$$

Рао показал, что $tr \hat{W}$ и евклидова норма $\| W \|$ обращаются в минимум, если $H = C_\rho$ и, следовательно, $V = V_\rho$. Остаточные погрешности:

$$\min tr \hat{W} = \sum_{\rho+1}^{n} \lambda_i \quad , \quad \min \| \hat{W} \| = \left(\sum_{\rho+1}^{n} \lambda_i^2 \right)^{1/2} \qquad (57),(58)$$

Подставим $H = C_\rho$. $V = V_\rho = XC_\rho$ в (55):

$$\hat{X} = XC_\rho (C_\rho^T X^T X C_\rho)^{-1} C^T X^T X \qquad (59)$$

По теореме (см. выше) оптимальный предиктор $\tilde{X} = Y_\rho C_\rho^T$. Приравняем $\tilde{X}$ и $\hat{X}$ (59):

$$XC_\rho C_\rho^T = XC_\rho (C_\rho^T X^T X C_\rho)^{-1} C_\rho^T X^T X \qquad (60)$$

Домножая (60) справа на C_ρ, получаем тождество, доказывающее эквивалентность (по норме $\mathcal{L}$) этих предикторов.

5. Двойная ортогональность

Оптимальное (по минимуму $\mathcal{L}$) факторное разложение (46) обладает свойством "двойной ортогональности":

$$A_\rho^T A_\rho = Y_\rho^T Y = L_\rho \quad ; \quad BB^T = C_\rho^T C_\rho = I \qquad (61)$$

(L_ρ – диагональная матрица первых ρ с.ч.). Возьмем произвольное разложение (46), не обладающее в общем случае свойством (61).

Как уже отмечалось, можно заменить (46) на эквивалентное разложение (не меняя R), введя промежуточные взаимно-обратные матричные множители:

$$X = \tilde{A}\tilde{B} + R; \quad \tilde{A} = AZ, \quad \tilde{B} = Z^{-1}B \text{ (индекс } \rho \text{ опущен)}. \quad (62)$$

(Матрица Z - квадратная, порядка ρ). Зададимся целью подобрать такую матрицу Z, чтобы сомножители $\tilde{A}$, $\tilde{B}$ обладали свойством (61);

$$\tilde{A}^T\tilde{A} = Z^T A^T A Z = L \quad (63)$$

$$\tilde{B}\tilde{B}^T = Z^{-1}BB^T (Z^{-1})^T = I \quad (64)$$

Обозначим $\mathcal{A} = A^TA$, $\mathcal{B} = (BB^T)$ и перепишем (63),(64) в виде:

$$Z^T\mathcal{A}Z = L, \quad Z^T\mathcal{B}Z = I \quad (65),(66)$$

Из (65),(66) следует спектральная задача относительно Z:

$$\mathcal{B}^{-1}\mathcal{A}Z = ZL, \quad Z^T\mathcal{B}Z = I \quad (67),(68)$$

Таким образом, надо найти с.в. матрицы $\mathcal{B}^{-1}\mathcal{A}$, ортонормированные в $\mathcal{B}$ - метрике, Если матрица $\mathcal{B}$ положительно-определенная, то эта задача имеет решение, причем Z есть главная матрица регулярного пучка квадратичных форм: $x^T\mathcal{A}x - \lambda x^T\mathcal{B}x$ (x есть ρ-мерный вектор). Преобразование $y = Zx$ приводит обе формы ($x^T\mathcal{A}x$ и $x^T\mathcal{B}x$) к каноническому виду (см. [13]).

Матрица $\mathcal{B}^{-1}\mathcal{A}$, вообще говоря, несимметричная, но, как нетрудно показать, она подобна симметричной матрице $S = \mathcal{B}^{-\frac{1}{2}}\mathcal{A}\mathcal{B}^{-\frac{1}{2}}$ и поэтому имеет простую структуру и вещественные с.ч..Используя преобразование $U = \mathcal{B}^{\frac{1}{2}}Z$, можно перейти от (67),(68) к спектральной задаче $SU = UL$ с обычной нормировкой: $U^TU = I$.

Рассмотрим один частный случай (46): $\rho = n$, $R = 0$, $B = W = X^TX$. Имеем:

$$X = VW \quad (69)$$

Таким образом, в качестве "факторных нагрузок" здесь взяты ковариации $(w_{ik}) = W$. Матрица $V = XW^{-1}$ обладает свойствами:

$$VX^T = XV^T = I; \quad V^TV = W^{-1}; \quad V^TVC = CL^{-1} \quad (70)$$

(L, C - решение основной задачи (I)). Ортогонализация представления (69) обеспечивается преобразованием

$$Z = WC = X^TY :$$

$$X = (V W C)(C^{-1} W^{-1} W) = (V X^T Y) C^T = Y C^T \quad (7I)$$

Нетрудно проверить, что $Z = WC$ удовлетворяет системе :

$$W Z = Z L ; \qquad Z^T W^{-2} Z = I , \qquad (72)$$

являющейся частным случаем системы (67), (68).

Аналогичная задача на одновременное приведение двух матриц к диагональному виду возникает в дискриминантном анализе (см., например, [43]).

Соотношение между МГК и факторным анализом изучалось многими авторами (см. [I] , где дана обширная библиография по этой проблеме).

В заключение сформулируем теорему, доказанную Окамото и Каназавой (Okamoto , Kanasawa) ([5I]), которая охватывает вышеприведенные результаты Рао и Дарроча. Рассмотрим факторное разложение (46) ($p \le n$).

Теорема. Пусть f - вещественная функция, определенная на множестве неотрицательных матриц порядка n , причем f строго возрастающая (если ее рассматривать как функцию скалярных аргументов: $f = g(\lambda_1, \dots, \lambda_n)$ и, кроме того, f инвариантна относительно ортогональных преобразований матричного аргумента: $f(C^T D C^T) = f(D)$ (C - ортогональная матрица). Тогда $F = f[(X - AB)(X - AB)]$ обращается в минимум, тогда и только тогда, когда $AB = Y_p C_p^T$, и $F_{min} = g(\lambda_{p+1}, \dots, \lambda_n, 0 \dots 0)$ (λ_i - с.ч. матрицы $X^T X$).

В частности, можно положить $f(D) = tr D$ или $f(D) = \|D\|$.

Другие сведения по МГК см. в [I , 3I , 52] , а также в обзоре [2] .

6. Теорема Рао-Дарроча и восстановление пропусков

Рассмотрим представление таблицы X опытных данных в виде:

$$X = AB + R \qquad (I)$$

Матрицы X, A, B, P имеют размеры соответственно $(m \times n), (m \times p)$,

$(p \times n)$, $(m \times n)$. Строки X суть объекты, столбцы – признаки; таким образом, x_{ik} есть значение k-го признака на i-м объекте. Столбцы A ("факторы") суть признаки в новой системе координат, задаваемой матрицей B "факторных нагрузок". Наконец, R есть матрица погрешностей приближенного представления $X \approx AB$. В зависимости от ограничений, накладываемых на A, B, R, возникают различные наборы факторов.

Метод главных компонент (МГК) иногда позволяет получить разложение вида (I) с небольшим количеством некоррелированных факторов при относительно малой норме погрешности (см. [I, 3I]).

Основное уравнение МГК:

$$X^T X C = C L \qquad (2)$$

Здесь L – диагональная матрица собственных чисел матрицы $X^T X$, а C есть ортогональная матрица соответствующих собственных векторов. Будем считать, что собственные числа λ_i занумерованы в убывающем порядке: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

Прямое и обратное преобразование МГК:

$$Y = XC; \quad X = YC^T. \qquad (3),(4)$$

Введем матрицы Y_p, C_p, образованные первыми p-столбцами матриц Y, C. Факторное разложение МГК:

$$X = Y_p C_p^T + R \qquad (5)$$

Столбцы матрицы Y_p называются главными компонентами.

Рассмотрим факторное разложение общего вида (I). В качестве нормы погрешности примем след матрицы $R^T R$:

$$Q = tr\,[(X - AB)^T (X - AB)] \qquad (6)$$

Фиксируем количество факторов p. Спрашивается: при каких A, B норма Q обращается в минимум? (Многозначность, вызванная очевидным соотношением $AB = AD^{-1} DB$, устраняется, если, не теряя общности, потребовать, чтобы выполнялось условие ортогональности: $BB^T = I$). Ответ на поставленный вопрос дает следующая теорема (см. [3I, 48]):

Теорема Рао-Дарроча: Минимум Q достигается при $A = Y_p$, $B = C_p^T$. При этом: $Q_{min} = \lambda_{p+1} + \lambda_{p+2} + \ldots + \lambda_n$.

Рао получил серию аналогичных результатов, характеризующих

экстремальные свойства преобразований МГК, в частности, что при том же способе задания A,B достигается также минимум эвклидовой нормы матрицы $R^T R$. Вышеприведенная формулировка наиболее соответствует теореме, доказанной Даррочем ([48]).

Пусть $\rho = I$. Имеем простейшее факторное разложение:

$$X = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} (\beta_1 , \dots , \beta_n) + R \qquad (7)$$

По теореме Рао-Даррочa, для минимизации нормы погрешности можно принять, что вектор $\beta^T = (\beta_1 , \dots , \beta_n)^T$ есть главный собственный вектор матрицы $X^T X$, а вектор $\alpha = (\alpha_1 , \dots , \alpha_m)^T$ - первая главная компонента. Из условия $Q = min$ следует, что искомые α, β удовлетворяют системе:

$$\begin{aligned} X^T \alpha &= \mu \beta \\ X \beta &= \nu \alpha \end{aligned} \qquad (8)$$

(Скаляры μ, ν - нормирующие множители: $\mu\nu = \lambda_{max}$).

Допустим теперь, что таблица X содержит пропуски $\tilde{x}_{ik}$ (символ $\tilde{x}_{ik}$ означает, что значение x_{ik} неизвестно). "Восстановление пропусков" есть оценивание пропущенных значений по имеющимся данным, за счет предполагаемых в таблице X скрытых корреляционных связей.

Дир (Dear) предложил методику восстановления пропусков, взяв за основу факторное разложение (7); (см. [49, 53]).

Алгоритм Дира (схематическое описание):

1. Все пропуски заменяются средними (по столбцам) значениями;
2. Из уравнений (8) находятся векторы α, β;
3. В соответствии с формулой (7) вычисляются оценки пропусков:

$$\tilde{x}_{ik} = \alpha_i \beta_k \qquad (9)$$

Несовершенство данной методики проявляется следующим образом. Пусть имеется полный набор собственных векторов и главных компонент, вычисленных после заполнения пропусков средними значениями:

$$\beta^\rho = (\beta_1^\rho , \dots , \beta_n^\rho)^T ; \quad \alpha^\rho = (\alpha_1^\rho , \dots , \alpha_m^\rho)^T .$$

Рассмотрим семейство оценок:

$$x_{ik} = \alpha_i^1 \beta_k^1 + \ldots + \alpha_i^\rho \beta_k^\rho \quad , \quad (\ \rho = 1, \ldots, n\ ) \qquad (10)$$

Эти оценки можно вычислить не только для пропусков, но и для известных элементов матрицы данных. С увеличением ρ они должны становиться более правдоподобными, ибо (10) представляет собой, по существу, разложение МГК (5), с погрешностью, норма которой с возрастанием ρ убывает. При максимальном значении $\rho = n$, в силу преобразований МГК (3),(4), оценки пропусков (10) совпадают с элементами, заместившими пропуски на первом этапе алгоритма, т.е. со средними значениями. Таким образом, естественное обобщение алгоритма Дира приводит к тривиальному результату.

Однако более тщательный анализ возможностей, предоставляемых теоремой Рао-Даррочa, показывает, что предварительное оценивание пропусков не является принципиально необходимым, если несколько обобщить постановку задачи. В случае $\rho = 1$ векторы α, β следует искать из условия минимума видоизмененного функционала $\tilde{Q}$, получаемого из Q удалением слагаемых, соответствующих пропускам $\tilde{x}_{ik}$. В случае $\rho > 1$ оптимальные векторы α^ρ, β^ρ можно найти последовательно, решая на каждом этапе задачу с параметром $\rho = 1$.

Таким путем, методика Дира может быть существенно улучшена.

ЛИТЕРАТУРА

1. Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений.- М., 1974.-240 с.
2. Андрукович П.Ф. Применение МГК в практических исследованиях.- М.: Изд. МГУ, 1973.-123 с.
3. Бабич В.В. Алгоритмическое описание итерационного метода классифицирования и упорядочения объектов ("Каскад П").- В кн.: Программные комплексы для целевой обработки информации. Новосибирск, 1977, с.27-38.
4. Бабич В.В., Федосеев Г.С. Прогнозная оценка магнитных анома-

лий Холзуно-Инского района.- В кн.: Логико-информационные исследования в геологии. Новосибирск, 1977, с.94-109.
5. Бишаев А.А. Итерационный способ нахождения информативной системы признаков для целевой классификации объектов.- В кн.: Ш Всесоюзная конференция по проблемам теоретической кибернетики. Новосибирск, 1974, с.185-187.
6. Бишаев А.А. Метод "Целевая итерационная классификация" ("Цикл").- В кн.: Логико-математическая обработка геологической информации (теория и математический аппарат). Новосибирск, 1976, с.70-92.
7. Бишаев А.А. Комплекс программ к методу "Целевая итерационная классификация".- В кн.: Программные комплексы для целевой обработки информации. Новосибирск, 1977, с.57-77.
8. Бугаец А.Н., Дуденко Л.Н. Математические методы при прогнозировании месторождений полезных ископаемых.- Л.: Недра, 1976.-256 с.
9. Васильев Ю.Л., Дмитриев А.Н. Спектральный подход к сравнению объектов, охарактеризованных набором признаков.- Докл. АН СССР, 1972, т.206, №6, с.1309-1312.
10. Васильев Ю.Л., Дмитриев А.Н. Простой способ сравнения объектов охарактеризованных набором признаков.- В кн.: Применение математических методов и ЭВМ для решения прогнозных задач нефтяной геологии. Новосибирск, 1973, с.60-63.
11. Васильев Ю.Р., Дмитриев А.Н., Золотухин В.В. Оценка существенности основных признаков дифференцированных трапповых интрузий с медно-никелевым оруденением логико-математическими средствами анализа для поисковых целей.- В кн.: Состояние и направление исследований по металлогении траппов. Красноярск, 1974, с.115-117.
12. Вышемирский В.С., Дмитриев А.Н., Трофимук А.А. Поисковые признаки гигантских нефтяных месторождений.- М., 1971.-16 с. (Спец. докл. к УШ мировому нефтяному конгрессу).
13. Гантмахер Ф.Р. Теория матриц.- М., 1966.-576 с.
14. Гуваков А.И., Дмитриев А.Н., Кандыба В.Н. Оценка перспективности оловорудных районов Приморья.- В кн.: Логико-информационные исследования в геологии. Новосибирск, 1975, с.68-94.
15. Дмитриев А.Н., Красавчиков В.О. Программы метода согласованных оценок.- В кн.: Логико-математическая обработка геологи-

ческой информации. Новосибирск, 1975, с.6-13.
16. Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П. О математических принципах классификации предметов и явлений.- В кн.: Дискретный анализ. Вып.7. Новосибирск, Наука, 1966, с.3-15.
17. Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П. Об одном принципе классификации и прогноза геологических объектов и явлений.- Геол. и геофиз., 1968, №5, с.50-64.
18. Дмитриев А.Н. Вопросы формализованных постановок геологических задач прогнозно-поискового профиля.- В кн.: Логико-математическая обработка геологической информации. Новосибирск, 1976, с.3-21.
19. Дмитриев А.Н., Красавчиков В.О. Процедуры математической обработки описаний нефтяных месторождений.- Геол. и геофиз., 1976, №11, с.86-96.
20. Запивалов Н.П., Каштанов В.А., Соколов А.Д. и др. Прогноз продуктивности локальных поднятий юга Западно-Сибирской плиты.- В кн.: Математические методы решения прогнозных задач нефтяной геологии. Новосибирск, 1978, с.36-77.
21. Йереског К.П., Кловсен Д.И., Реймент Р.А. Геологический факторный анализ.- Л.: Недра, 1980.-224 с.
22. Кандыба В.Н. Программа П1 "Качели для бинарных таблиц".- В кн.: Логико-математическая обработка геологической информации. Новосибирск, 1975, с.14-19.
23. Кандыба В.Н. Программа П2 "Расчет коэффициентов".- В кн.: Логико-математическая обработка геологической информации. Новосибирск, 1975, с.20-26.
24. Кандыба В.Н. Программа "Вычисление согласованной системы информационных весов по ЦКП" (П1).- В кн.: Программные комплексы для целевой обработки информации. Новосибирск, 1977, с.12-15.
25. Константинов Р.М., Сиротинская С.В., Бортников Н.С. О формационной классификации гидротермальных оловянных месторождений на статистической основе.- В кн.: Локальное прогнозирование в рудных районах Востока СССР. М.: Наука, 1972,с.147-159.
26. Константинов Р.М., Дмитриев А.Н. Использование математических методов для анализов геологических факторов, влияющих на масштабы оруденения.- Геология рудных месторождений, 1970, №2, с.56-64.

27. Кренделев Ф.П., Дмитриев А.Н., Журавлев Ю.И. Сравнение геологического строения зарубежных месторождений докембрийских конгломератов с помощью дискретной математики.- ДАН СССР, 1967, т.173, №5, с.1149-1152.
28. Макаров С.В., Смертин Е.А. Центрированная качельная процедура для нахождения согласованной системы информационных весов.- В кн.: Логико-математическая обработка геологической информации. Новосибирск, 1976, с.92-100.
29. Макаров С.В. Задача о нахождении весовых коэффициентов и согласованных оценок.- В кн.: Методы моделирования сложных производственных систем и непрерывных тектонических процессов. Томск, 1978, с.103-110.
30. Макаров С.В. Теорема Рао-Дарроча и восстановление процусков.- В кн.: Математические вопросы анализа данных. Новосибирск, 1980, с.58-61.
31. Рао С.Р. Линейные статистические методы и их применения.- М., 1968.
32. Распознавание образов в задачах качественного прогноза рудных месторождений. /Федосеев Г.С., Бабич В.В., Зайков В.В. и др.- Новосибирск: Наука, 1980.-208 с.
33. Смертин Е.А. Программа ПЗ "Симметрия".- В кн.: Логико-математическая обработка геологической информации. Новосибирск, 1975, с.27-35.
34. Смертин Е.А., Дмитриев А.Н., Макаров С.В. Комплекс программ по методу согласованных оценок.- В кн.: Программные комплексы для целевой обработки информации. Новосибирск, 1977, с.6-11.
35. Смертин Е.А. Вопросы теории и алгоритма на базе построения - тестов.- В кн.: Логико-информационные исследования в геологии. Новосибирск, 1977, с.48-67.
36. Суппес П., Зинес Дж. Основы теории измерений.- В кн.: Психологические измерения. М.: Мир, 1967, с.110-196.
37. Сухов Л.Г., Дуденко Л.Н., Городнянский А.А. Автоматизированная оценка перспектив трапповых провинций на Си - руды.- Геол. и геофиз., 1981, №1, с.82-99.
38. Сухов Л.Г., Дуденко Л.Н., Наторхин И.А. Количественные методы прогнозирования эндогенных рудных месторождений.- Л.: Недра, 1981.-176 с.

39. Трофимук А.А., Васильев Ю.Л., Вышемирский В.С., Дмитриев А.Н. Сравнительное изучение месторождений нефти спектральным методом.- В кн.: Применение математических методов и ЭВМ для решения задач нефтяной геологии. Новосибирск, 1977, с.3-6.
40. Трофимук А.А., Вышемирский В.С., Дмитриев А.Н., Карогодин Ю.Н. Геолого-геохимические критерии нефтегазоносности.- Новосибирск: Наука, 1976.- 135 с.
41. Трофимук А.А., Дмитриев А.Н. Нефтепрогноз как информационная проблема.- В кн.: Математические методы решения прогнозных задач в нефтяной геологии. Новосибирск, 1978, с.3-35.
42. Трофимук А.А., Вышемирский В.С., Дмитриев А.Н. Прогнозирование месторождений нефти математическими методами.- В кн.: Критерии поисков зон нефтегазонакопления. М.: Наука, 1979, с.30-35.
43. Уилкс С.С. Математическая статистика.- М., 1967.- 632 с.
44. Усманов Ф.А. Основы математического анализа геологических структур.- Ташкент: ФАН, 1977.- 206 с.
45. Фаддеев Д.К., Фаддеева В.Н. Вычислительные методы линейной алгебры.- М., 1963.- 736 с.
46. Чегис И.А., Яблонский С.В. Логические способы контроля работы электрических схем.- Труды Матем. ин-та им. В.А.Стеклова. М., 1958, т.51, с.270-339.
47. Шакин В.В. Уравновешивание матрицы данных.- В кн.: Опознование и описание линий. М., 1972.
48. Darroch J.N. An Optimal property of principal components. - The Ann. of Math.Stat., v.36, N5, Okt., 1965, p.p1579-1582.
49. Dear R.E. A principle-component missing data method for multiple regression model. System Dev.Corp.Tech.rep.,1959, 86p.
50. Eckart C., Young G. The approximation of one mftrix by another of lower ronk.-Psychometrika, 1936, v.1, p.p.211-218.
51. Okamoto M., Kanazawa M. Minimization of eigenvalues of a matrix and optimality of principal components. - Ann. Math. Stat., 1968, v.39, N 3.
52. Rac C.R. The use and interpretation of principal component analysis in applied research. Sankhya, 1965, ser.A,26,part 4.
53. Timm N.H. The Estimation of variance-covariance and correlation matrices from incomplete data.-Psychomet.,1970,v.35, N 4.

С О Д Е Р Ж А Н И Е

Метод согласованных оценок

Методические рекомендации

Ответственный за выпуск Т.И.Штатнова

Утверждено к печати
Институтом геологии и геофизики СО АН СССР

Технический редактор Н.Н. Александрова

Подписано к печати 25.03.82. МН 15391.
Бумага 60x84/16. Печ.л. 8,25. Уч.-изд.л. 7,6.
Тираж 400. Заказ 180. Цена 55 коп

Институт геологии и геофизики СО АН СССР
Новосибирск, 90. Ротапринт.