ТАРТУСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ТРУДЫ

ВЫЧИСЛИТЕЛЬНОГО ЦЕНТРА

36

ТАРТУ
1976

ТАРТУСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ТРУДЫ ВЫЧИСЛИТЕЛЬНОГО ЦЕНТРА

ВЫПУСК 36

ТАРТУ 1976

К ТЕОРИИ ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ
(АСПЕКТ КВАНТОВОЙ ФИЗИКИ)

Т.Э. Мелс

Как известно ([1], [2], [3]), вероятностный формализм квантовой физики существенно отличается от формализма обычной теории вероятностей. В частности, класс всех событий в физике не является булевой алгеброй из-за недистрибутивности ([3]). Были предприняты попытки включить оба формализма в более общую схему как частные случаи ([4], [7]). Однако, на этом пути возникает ряд трудностей, например, теряется естественная и наглядная интерпретация, свойственная классической теории вероятностей. Все это вызывает чувство неуверенности и заставляет сомневаться в целесообразности предлагаемых обобщений. Тем не менее, обобщения такого типа необходимы.

В настоящей статье описывается в общих чертах новая возможность для строения мостика между классическим и квантовым формализмами вероятностей. О полезности предлагаемой идеи судить рано, однако некоторым физикам (например тем, кто продолжают верить в существование скрытых параметров у квантовых систем или их среды) наша концепция может представлять определенный интерес.

Автор выражает здесь искреннюю благодарность Л.-Р. Харку за полезное обсуждение некоторых вопросов, затронутых в настоящей статье.

§1. Общее описание подхода

В колмогоровской теории вероятностей [8] случайность моделируется тем, что элементарное событие (=случай, результат испытания) рассматривается как случайная точка ω в некотором множестве Ω, а все случайные величины случайны именно потому, что являются функциями ω. Для формальной теории совсем не важно, какие механизмы определяют конкретное значение ω в отдельном испытании. Требуется лишь, чтобы действие этих механизмов могло быть в конце концов описано в терминах вероятностного распределения ω на некоторой выделенной σ-алгебре подмножеств (=событий) в Ω. Нам кажется, что для получения интерпретируемой квантово-вероятностной теории важно уточнение именно процесса генерирования случайного значения ω, т.е. случая.

Рассмотрим это подробнее. Всякая случайная величина принимает числовое значение исключительно в результате испытания (=эксперимента), которое вызывает случай (=элементарное событие), и для каждой случайной величины существует испытание, определяющее значение этой случайной величины. В колмогоровской теории, например, постулируется существование испытания, определяющего значения сразу всех случайных величин. Иными словами, все случайные величины в классической теории зависят от одного и того же случая ω, вызываемого испытанием.

Испытание, каждый возможный результат которого точно определяет значение случайной величины x, называем x-испытанием. Формально могут существовать еще и такие испытания, никакие результаты которых не определяют значения x, но которые могут быть в принципе дополнены или усовершенствованы таким образом, что превращаются в x-испытание. Эти испытания называем x-нейтральными. Далее мыслимы ситуации, где выполнение некоторого испытания принципиально лишает испытателя возможности наложить на это испытание или на его расширение функции x-испытания. Такие испытания естественно называть x-несовместимыми. Любое x-несовместимое испытание вызывает случаи, при которых значение случайной величины x остается принципиально неопределенным. Если x-испытание y-несовместимо или наоборот, то называем случайные величины x и y несовместимыми. У несовместимых случайных величин не может быть совместного распределения и вообще их совместное вероятностное поведение лишено всякого смысла. Однако, в колмогоровской теории все случайные величины совместимы. Приведем теперь несколько поясняющих примеров.

Пример 1. Бросают игральную кость. В качестве значений ω рассматриваем разные положения брошенной кости. Некоторое испытание - это выделение положения ω игральной кости в определенный момент времени, в определенной части пространства или же по некоторому другому признаку. Например, количество очков при обычной его трактовке имеет, как случайная величина x, в качестве x-испытания испытание, которое устанавливает положение кости после того, как кость лежит на одной из граней. Таким образом, здесь x-испытание (=экспери-

мент) вызывает (=генерирует) элементарное событие ω, которое однозначно определяет значение случайной величины x.

Если игральная кость имеет отсеченные вершины, то при x-испытании, в случае, когда после бросания кость покоится на одной из отсеченных вершин, следует повторять бросание до тех пор, пока кость не будет лежать на одной из основных граней, и лишь тогда испытание заканчивается. Если отсеченные вершины занумеровать числами от 1 до 8, а через y обозначить соответствующую этим числам случайную величину, то y-испытание x-несовместимо, а x-испытание y-несовместимо. Другими словами, x и y несовместимы.

Пример 2. В непрозрачной стене имеется 2 щели, через которые электроны могут пролететь с катода до экрана. Поток электронов настолько слаб, что практически с вероятностью 1 в любой момент времени он состоит не более чем из одного электрона. Рассматриваем случайные величины x, y, z и u, где x (соответственно y) есть координата первого попавшего на экран электрона, если открыта только верхняя (соответственно только нижняя) щель, а z (соответственно u) есть координата электрона на экране, если открыты обе щели, причем не наблюдают (соответственно наблюдают), какую из них электрон проходит. Тогда элементарные события, определяющие значения этих случайных величин, следующие. При x-испытании - это траектории электрона (вообще говоря не как частицы!), определенные с точностью до точек "катод - верхняя щель - точка попадания на экране"; при z-испытании - это траектории электрона, определенные с точностью до своих точек "катод - пара щелей - место попадания на экране"; при u-испытании - это

траектории, фиксированные с точностью до точек "катод - любая (но только одна) из щелей - место попадания на экране". Каждое из рассмотренных элементарных событий получается в специальном (x-, y-, z-, u-) испытании, которое вынуждает систему (эмиттированного электрона) перейти в некоторое состояние ω во вполне определенном множестве элементарных событий, а если это не удается, то испытание следует продолжать. Неудачные попытки генерировать требуемое элементарное событие в испытании никакого значения не имеют и ими можно пренебречь.

Как показывает подробный анализ физической стороны приведенного примера (см., например, [7], стр. 251-258), множества элементарных событий, генерируемых в x-, y- и z-испытаниях, не пересекаются.

Пример 3. Случайной величиной x является координата элементарной частицы. Тогда, если частица не находится в (физическом) состоянии с определенной координатой (например, после измерения импульса), то x-испытание (состоящее, например, в установлении фотопластинки на путь электрона) случайным образом возмущает состояние частицы, пока частица не переходит в состояние с определенной координатой. Полученное координатное состояние (регистрируемое на пластинке) и является тем элементарным событием, которое генерируется в испытании.

Из анализа приведенных примеров видно, что мы предполагаем существование взаимосвязи между испытанием и элементарными событиями, которые могут появляться в результате этого испытания (и рассматриваются как случайные). Каждому испытанию соответствует вполне определенное множество возможных

элементарных событий (=множество допустимых элементарных событий). Кроме того, каждой случайной величине x также соответствует некоторое множество элементарных событий - область определения случайной величины. Если области определения случайных величин x и y не совпадают, то эти случайные величины несовместимы.

Мы будем предполагать, что окончательное (случайное, но допустимое) значение ω вырабатывается в испытании путем случайного преобразования определенного исходного состояния ω_0, которое при этом может быть или фиксированным (тогда говорим, что система находится в чистом состоянии), или же полученным как случайное. В последнем случае предполагается, что вероятностное поведение ω_0 подчиняется классической теории вероятностей, т.е. ω_0 имеет обычное вероятностное распределение на некоторых выделенных измеримых подмножествах элементарных событий. Заимствуя терминологию у физиков, будем в этом случае говорить, что система находится в смешанном состоянии (см., например, [6], гл. 5 и 6). Смешанных состояний дальше не рассматриваем.

Можно считать, что испытание заканчивается, как только случайное элементарное событие ω, полученное в испытании, оказывается допустимым. Пока этого нет, испытание продолжается и заключается в продолжении случайных преобразований над исходным элементарным событием ω_0. Если ω_0 само является допустимым для испытания, то никаких дальнейших случайных преобразований $\omega_0 \to \omega$ не будет и испытание заканчивается результатом ω_0.

Для математического формализма совершенно не важно, ка-

кой конкретный механизм возмущает исходное состояние ω_0 и требуется ли для этого конечный промежутск времени (как, например, в примере 1), или нет. Важно только вероятностное поведение (=распределение) элементарного события. Нашим основным допущением будет, что вероятностное поведение ω не зависит от испытания. Роль испытания заключается только в запуске процесса возмущения и в выборе первого (например, по упорядочению) допустимого элементарного состояния, возникающего в процессе возмущения.

Имея в виду ситуацию в квантовой механике, вероятностное поведение ω нельзя описывать средствами классической теории вероятностей. Здесь требуется вместо обычного вероятностного распределения более общее понятие мультираспределения. Математическое описание равномерного мультираспределения на гильбертовом пространстве будет дано в другой статье, а ниже приведены необходимые сведения о нем, опуская доказательства.

§2. Вероятностные распределения в квантовой физике

Согласно общеизвестному формализму квантовой механики (см. [5] или [9]), чистые состояния квантованной физической системы представляются одномерными подпространствами в сепарабельном (бесконечномерном) комплексном гильбертовом пространстве H_c. Измеримые физические величины (=случайные величины, наблюдаемые) представляются линейными самосопряженными операторами в H_c.

Пусть x - некоторая измеримая величина, L_x - соответст-

вующий ей самосопряженный оператор, ω_0 – базисный вектор одномерного подпространства (начального чистого) состояния, $\omega_0 \in H_c$. Пусть еще $F_x(\lambda)$, $\lambda \in \mathbb{R}$, обозначает разложение единицы, принадлежащее оператору L_x, в частности, $F_x(\lambda)$ – ортопроектор в H_c, $F_x(-\infty) = 0$ (нулевой оператор), $F_x(\infty) = 1$ (единичный оператор); при $\lambda_1 \leq \lambda_2$ имеем $F_x(\lambda_1) \leq F_x(\lambda_2)$ и проекторнозначная функция $F_x(\lambda)$ непрерывна слева по λ (см. [10]). Тогда вещественная функция

$$\Phi_x(\lambda) = (F_x(\lambda)\omega_0, \omega_0) \,/\, \|\omega_0\|^2 ,$$

где $(\ ,\ )$ обозначает скалярное произведение в H_0 и $\|\omega_0\|^2 = (\omega_0, \omega_0)$, обладает всеми свойствами обычной функции распределения (в смысле теории вероятностей). Согласно стандартной интерпретации формализма, $\Phi_x(\lambda)$ соответствует вероятностному распределению величины x при системе, находящейся в состоянии ω_0. Например, вероятность получить при измерении x значение в промежутке $[a,b)$ равна

$$P\{x \in [a,b)\} = \|(F_x(b) - F_x(a))\omega_0\|^2 \,/\, \|\omega_0\|^2 . \qquad (*)$$

Если оператор L_x имеет чисто дискретный спектр (т.е. x – дискретная величина), то x может быть в принципе измерена абсолютно точно. Если же спектр оператора L_x не чисто дискретен, то не существует испытания, которое с достоверностью генерировало бы элементарные события, точно определяющие значение x (ср. [5], стр. 165). Действительно, если такое испытание существовало бы, то после этого испытания система должна находиться в состоянии ω_1, где x имеет точное значение, скажем a. Но из (*) видно, что если $F_x(a+0) - F_x(a) = 0$

(т.е. a не входит в дискретную часть спектра L_x), то должно быть

$$1 = P\{x = a\} = \|(F_x(a+0) - F_x(a))\omega_1\|^2 / \|\omega_1\|^2 = 0,$$

что невозможно. Это на первый взгляд неестественная ситуация, в самом деле, даже лучше соответствует действительности, чем положение в классической теории вероятностей, где каждая случайная величина может быть измерена абсолютно точно.

Усложнения формализма в недискретном случае настолько серьезные, что здесь мы должны ограничиваться чисто дискретным случаем.

Итак, если x - дискретная величина, то допустимые для x-испытания элементарные события (одномерные подпространства в H_c) составляют множество

$$D = \bigcup_{j=1}^{\infty} \{c_j e_j \mid c_j \in \mathbb{C},\ c_j \neq 0\},$$

где $\mathbb{C}$ есть поле комплексных чисел, а $\{e_1, e_2, \ldots\}$ - ортонормированный базис в H_c, составленный из собственных векторов оператора L_x. Если оператор L_x имеет простой спектр (т.е. кратность всех собственных чисел 1), то этот базис определяется с точностью до умножения векторов e_j на числа модуля 1. Если же спектр не простой, то можно рассматривать x как функцию некоторой другой дискретной величины y, $x = f(y)$, причем y уже имеет простой спектр. В этом случае всякое y-испытание является и x-испытанием и мы считаем, что измерение x - это фактически измерение y с последующим преобразованием функцией f полученного числового результата (поскольк-

ку у не определяется по x однозначно, существует бесконечно много принципиально разных x-испытаний). Учитывая сказанное, можем всегда ограничиться случаем, где оператор величины x имеет дискретный спектр.

Допустим теперь в духе §1, что в процессе измерения измерительная аппаратура производит над ω_0 некоторые случайные преобразования $\omega_0 \to \omega$ до тех пор, пока ω не будет допустимым для x-испытания, т.е. пока $\omega \in D$. Уточним вид преобразования $\omega_0 \to \omega$, считая $\omega = \omega_0 + \xi$, где $\|\xi\| < \|\omega_0\|$ и ξ – случайный (в определенном далее смысле) вектор в H_C. Вероятностное поведение ξ опишем специальной системой распределений (=мультираспределением). Определим для этого при каждом $n=1,2,\ldots$ пространство V_n линейных комбинаций $\sum \alpha_i 1_{S_i}$ индикаторных функций 1_{S_i} симплексов S_i с вершинами, количество которых $\leq n$, в бесконечномерном <u>вещественном</u> гильбертовом пространстве H_R, причем S_i могут быть и вырожденными (считаем, что нормы в H_C и H_R совпадают и вообще H_R является вещественной формой H_C). Тогда, очевидно, V_n является линейной решеткой.

Определим, далее, $P_n 1_S$ (P_n – функционал) как геометрический объем симплекса S в H_R и пусть $P_n(\sum \alpha_i 1_{S_i}) = \sum \alpha_i P_n 1$. Тогда P_n является на V_n линейным, неотрицательным (если $f_n \geq 0$, то $P_n f_n \geq 0$) и непрерывным в том смысле, что при (поточечной) сходимости $f_i \downarrow 0$ в V_n имеем $P_n f_i \downarrow 0$. Оказывается, что с пространства V_n функционал P_n однозначно продолжается (с сохранением указанных трех свойств) на некоторую максимальную линейную решетку K_n вещевтвенных функций на H_R ($V_n \subset K_n$). При этом $K_n \subset K_{n+1}$ для всех n, и если $f \in K_n$, то $P_{n+1} f = 0$. Иными словами, если f входит в некоторую из ре-

шеток K_n (в этом случае говорим, что f входит в область определения мультираспределения), то максимально для одного m значение функционала P_m от f ($f \in K_m$) может отличаться от нуля. Последовательность пар $\{<P_n, K_n> \mid n=1,2,\dots\}$ мы и называем в настоящей работе мультираспределением.

Если $|f|$ входит в область определения мультираспределения и при некотором n будет $P_n|f| > 0$, то называем f существенной. Эту терминологию переносим также к множествам: множество существенное, если его индикаторная функция существенная. Если f существенная, то имеется ровно одно n так, что $P_n|f| > 0$, $P_{n+1}|f| = 0$ и $f \notin K_{n-1}$ или $n=1$. Однозначно определенное число n называем степенью существенности функции f (или множества).

Пусть Ω_X – область определения случайной величины (наблюдаемой) x и Ω_{X,ω_0} – подмножество в Ω_X, состоящее из тех элементарных событий ω, которые могут появляться в результате случайного преобразования $\omega_0 \to \omega$ в x-испытании. Допустим, что множество Ω_{X,ω_0} имеет степень существенности n и пусть еще множество $\Omega_{X,\omega_0} \cap B$ входит в область определения мультираспределения. Тогда

$$1_{B\cap\Omega_{X,\omega_0}} \in K_n,\ P_n 1_{\Omega_{X,\omega_0}} > 0,\ P_n 1_{\Omega_{X,\omega_0}} \geq P_n 1_{B\cap\Omega_{X,\omega_0}} \geq 0$$

и, значит,

$$0 \leq (P_n 1_{B\cap\Omega_{X,\omega_0}}) / (P_n 1_{\Omega_{X,\omega_0}}) \leq 1.$$

Мы интерпретируем отношение $P_n 1_{B\cap\Omega_{X,\omega_0}} / P_n 1_{\Omega_{X,\omega_0}}$ как вероятность события $\{X \in B\}$ в x-испытании, если система находится в

состоянии ω_0. Интерпретация оправдывается тем, что указанное отношение ведет себя как вероятность на σ-алгебре всех событий вида $\{x \in B\}$ (x и ω_0 фиксированы).

Возвращаясь к рассмотрению квантовых систем, имеем $\Omega_x = D$ и $\Omega_{x,\omega_0} = G \cap D$, где

$$G = \{\omega_0 + \xi \mid \|\xi\| < \|\omega_0\|, \xi \in H_0\}.$$

Покажем, что Ω_{x,ω_0} – существенное множество в H_R со степенью существенности 3. Для этого отметим, что множество $B_j = D_j \cap G$, где $D_j = \{ce_j \mid c \in \mathbb{C}\}$ представляет окружность радиуса $|(e_j,\omega_0)|$ в плоскости векторов e_j и ie_j, ортогональных в H_R (здесь i – мнимая единица; напомним, что $(\ ,\ )$ есть скалярное произведение в H_0). Поэтому $1_{B_j} = \lim\limits_n \uparrow 1_{S_{jk}}$, где S_{jk} – некоторые симплексы с 3 вершинами на этой плоскости. Увидим, что $1_{B_j} \in K_3$ и $P_3 1_{B_j}$ равно площади окружности B_j, т.е. $P_3 1_{B_j} = \pi |(e_j,\omega_0)|^2$. Но

$$1_{\Omega_{x,\omega_0}} = \lim_n \uparrow \sum_{j=1}^{n} 1_{B_j},$$

так что

$$0 < P_3 1_{\Omega_{x,\omega_0}} = \pi \lim_n \uparrow \sum_{j=1}^{n} |(e_j,\omega_0)|^2 = \pi \|\omega_0\|^2 < \infty,$$

что и доказывает сказанное.

Согласно квантовой механике, возможными состояниями дискретной величины x являются собственные значения $\lambda_1, \lambda_2, \ldots$ ее оператора L_x, и если $\omega = ce_j$ ($c \neq 0$), то $x(\omega) = \lambda_j$. У нас равенство $\omega = ce_j$ равносильно включению $\omega \in D_j$. Поэтому выше-

изложенные соображения и формулы приведут к соотношениям

$$P\{x=\lambda_j\} = P\{\omega \in D_j\} = P_{\mathfrak{Z}^1 D_j \cap \Omega_{x,\omega_0}} \,/\, P_{\mathfrak{Z}^1 \Omega_{x,\omega_0}} =$$

$$= P_{\mathfrak{Z}^1 B_j} \,/\, P_{\mathfrak{Z}^1 \Omega_{x,\omega_0}} = \frac{\pi |(e_j,\omega_0)|^2}{\pi \|\omega_0\|^2} = \left|(e_j,\frac{\omega_0}{\|\omega_0\|})\right|^2,$$

которые равносильны определению (*), поскольку

$$\|(F_x(\lambda_j+0) - F_x(\lambda_j))\omega_0\|^2 = |(e_j,\omega_0)|^2.$$

Таким образом, мы получим вывод распределения x, которое в квантовой механике просто постулируется. Дж. фон Нейман пишет в своей книге по квантовой механике ([5], стр. 163): "Мы ответили при сделанных предположениях относительно оператора L_x (дискретный и простой спектр у L_x - Т.М.) на вопрос о том, "что" происходит при измерении его величины x. Конечно, вопрос "как" остается пока невыясненным". Наша гипотеза о переходе $\omega_0 \to \omega$ касается именно вопроса "как?". Отметим, однако, что в нашем подходе нужны некоторые дополнительные физические предположения. Например, измерение x при непростом спектре L_x возможно только с помощью установки для измерения некоторой другой величины с простым спектром. Вопрос о согласии этих, по существу новых предположений, с экспериментом требует еще дополнительного исследования.

Литература

1. Cohen, L., Can Quantum mechanics be formulated as a classical probability theory? Philos. Soi., 1966, 33 (4).
2. Varadarajan, V.S., Probability in physics and a theorem on simultaneous observability. Commun. on P. a. Appl. Math., 1962, 15, 189-217.
3. Suppes, P., The probabilistic argument for a nonclassical logic of quantum mechanics. Philos. Soi., 1966, 33.
4. Gudder, S., Spectral methods for generalized probability theory. Trans. Am. Math. Soc., 1965, 119.
5. Нейман Дж. фон, Математические основы квантовой механики. Наука, 1964.
6. Кемпфер Ф., Основные положения квантовой механики. Мир, 1967.
7. Кац М., Вероятность и смежные вопросы в физике. Мир, 1965.
8. Колмогоров А.Н., Основные понятия теории вероятностей. Наука, 1974.
9. Макки Дж., Лекции по математическим основам квантовой механики. Мир, 1965.
10. Ахиезер Н.И., Глазман И.М., Теория линейных операторов в гильбертовом пространстве. Наука, 1966.

О РАСПРЕДЕЛЕНИЯХ, СВЯЗАННЫХ С ВЫБОРОЧНОЙ КОРРЕЛЯЦИОННОЙ МАТРИЦЕЙ

Т.Х.-А. Колло

1. Введение

Пусть нам дана выборка X объема m из нормальной p-мерной генеральной совокупности

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p1} & x_{p2} & \cdots & x_{pm} \end{pmatrix},$$

где столбцы X_j ($j=1,\ldots,m$) матрицы X соответствуют независимым наблюдениям и являются p-мерными случайными векторами с распределением $N(\vec{\mu},\Sigma)$, где

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}; \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}.$$

После применения обычного ортогонального преобразования для элиминирования математического ожидания $\vec{\mu}$ (см. [1], § 3.3)

получим из матрицы X $(p \times n)$-матрицу Z:

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ z_{p1} & z_{p2} & \cdots & z_{pn} \end{pmatrix}, \qquad (1)$$

где $n = m - 1$.

Столбцы

$$Z_j = \begin{pmatrix} z_{1j} \\ z_{2j} \\ \vdots \\ z_{pj} \end{pmatrix} \quad (j=1,\ldots,n) \qquad (2)$$

матрицы Z - независимые случайные векторы с распределением $N(0,\Sigma)$. Строки

$$Y_i = (z_{i1}, z_{i2}, \ldots, z_{in}) \quad (i=1,\ldots,p) \qquad (3)$$

матрицы Z соответствуют исследуемым признакам и являются n-мерными случайными векторами, координаты которых - независимые случайные величины с распределением $N(0,\sigma_{ii})$ $(i=1,\ldots,p)$. Вместо теоретической ковариационной матрицы Σ в практике используют ее несмещенную оценку[1] S_p

$$S_p = \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i'. \qquad (4)$$

---

[1] В дальнейшем индекс у квадратной матрицы для обозначения ее порядка может быть опущен там, где это не приводит к недоразумениям.

Для удобства обычно обозначают

$$A = nS ,$$

где элементы матрицы A определены формулой

$$a_{ij} = \sum_{k=1}^{n} z_{ik} z_{jk} \quad (i,j=1,\dots,p). \tag{5}$$

2. Распределение вероятности матрицы выборочных коэффициентов корреляции

Обозначим через P корреляционную матрицу генеральной совокупности

$$P = (\rho_{ij}) \quad (i,j=1,\dots,p),$$

где

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} .$$

В матричном виде

$$P = (\delta_{ij}\sigma_{ii})^{-\frac{1}{2}} \sum (\delta_{ij}\sigma_{ii})^{-\frac{1}{2}} ,$$

где δ_{ij} – символ Кронекера и $(\delta_{ij}\sigma_{ii})$ – диагональная матрица:

$$(\delta_{ij}\sigma_{ii}) = \begin{pmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_{pp} \end{pmatrix} .$$

В частном случае, когда матрица Σ диагональная ($\Sigma=(\delta_{ij}\sigma_{ii})$), матрица P равна единичной матрице I порядка p:

$$P = I.$$

Выборочной корреляционной матрицей называем матрицу $R = (r_{ij})$:

$$R = (\delta_{ij}a_{ii})^{-\frac{1}{2}}A(\delta_{ij}a_{ii})^{-\frac{1}{2}}. \qquad (6)$$

Для элементов матрицы R получим формулу

$$r_{ij} = \frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}} \quad (i,j=1,\ldots,p). \qquad (7)$$

Приведем теперь схематичное доказательство (нам потом будет нужен промежуточный результат доказательства) хорошо известной теоремы о плотности вероятности совместного распределения элементов выборочной корреляционной матрицы (см. [1], § 7.6).

ТЕОРЕМА 1. Если случайные p-мерные векторы $X_1,\ldots,X_m$ независимы и одинаково распределены с законом распределения $N(\vec{\mu},(\delta_{ij}\sigma_{ii}))$, то плотность вероятности выборочных коэффициентов корреляции дается формулой

$$\frac{\Gamma^p\left(\frac{n}{2}\right)|R|^{(n-p-1)/2}}{\prod_{i=1}^{p}\Gamma\left(\frac{n-i+1}{2}\right)\pi^{p(p-1)/4}}, \qquad (8)$$

где $n = m - 1$.

Доказательство. Исходим из распределения матрицы A, плот-

ность вероятности которой в нашем случае равна

$$\frac{|A|^{(n-p-1)/2}\exp(-\frac{1}{2}\sum_{i=1}^{p}\frac{a_{ii}}{\sigma_{ii}})}{2^{np/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\sigma_{ii}^{n/2}\prod_{i=1}^{p}\Gamma(\frac{n-i+1}{2})}.$$

Проведем в этой формуле замену переменных

$$\begin{cases} a_{ij} = \sqrt{a_{ii}}\sqrt{a_{jj}}\,r_{ij} & (i\neq j,\ i,j=1,\dots,p) \\ a_{ii} = a_{ii}\,. \end{cases}$$

Якобиан J этого преобразования равен

$$\prod_{i=1}^{p} a_{ii}^{(p-1)/2}\,.$$

Если подставить выражение для a_{ij} из формулы замены переменных в выражение плотности вероятности матрицы A и результат умножить на J, то получится следующая совместная плотность вероятности случайных величин r_{ij} $(i>j,\ i,j=1,\dots,p)$ и a_{ii} $(i=1,\dots,p)$:

$$\frac{|R|^{(n-p-1)/2}}{\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma(\frac{n-i+1}{2})}\prod_{i=1}^{p}\frac{a_{ii}^{n/2-1}\exp(-\frac{1}{2}\frac{a_{ii}}{\sigma_{ii}})}{2^{n/2}\sigma_{ii}^{n/2}}. \qquad (9)$$

Так как

$$\int_0^\infty \frac{a_{ii}^{n/2-1}\exp(-\frac{1}{2}\frac{a_{ii}}{\sigma_{ii}})}{2^{n/2}\sigma_{ii}^{n/2}}\,da_{ii} = \Gamma(\tfrac{n}{2}),$$

то из выражения (9) получим формулу (8), что и завершает доказательство.

Из теоремы 1 получим следующее простое следствие.

СЛЕДСТВИЕ 1. Если выполнены предположения теоремы 1, то множества случайных величин $\{r_{ij}\} = \{r_{ij} \mid i,j=1,\dots,p;\ i>j\}$ и $\{a_{ii}\} = \{a_{ii} \mid i=1,\dots,p\}$ независимы.

Доказательство. Следствие прямо вытекает из выражения (9), так как оттуда видно, что

$$f(\{a_{ii}\},\{r_{ij}\}) = f(\{a_{ii}\})f(\{r_{ij}\}),$$

где $f(\{a_{ii}\},\{r_{ij}\})$ - совместная плотность вероятности множеств случайных величин $\{r_{ij}\}$ и $\{a_{ii}\}$, а $f(\{r_{ij}\})$ и $f(\{a_{ii}\})$ - плотности вероятности множеств $\{r_{ij}\}$ и $\{a_{ii}\}$ соответственно.

Изучим теперь свойства элементов матрицы R.

ЛЕММА 1. Внедиагональные элементы одной строки (одного столбца) выборочной корреляционной матрицы R являются независимыми случайными величинами, если $P = I$.

Доказательство. По определению

$$r_{ij} = \frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}}.$$

Из равенств (2) - (5) получим

$$a_{ij} = Y_i Y_j'.$$

Отсюда видно, что при зафиксированном случайном векторе $Y_i = Y_i^o$

$$Y_i^o Y_j' = \sum_{k=1}^{n} z_{ik}^o z_{jk} = a_{ij}^o .$$

Поскольку случайные величины z_{jk} независимы при различных значениях индекса j, то и случайные величины a_{ij}^o независимы при разных значениях индекса j, если индекс i зафиксирован. Андерсон (см. [1], § 4.2) показывает, что условная плотность вероятности случайной величины r_{ij} при зафиксированном случайном векторе $Y_i = Y_i^o$ дана следующим выражением:

$$\frac{\Gamma(\frac{n}{2})(1-r_{ij}^2)^{(n-3)/2}}{\Gamma(\frac{n-1}{2})\sqrt{\pi}} . \qquad (10)$$

Но так как плотность вероятности (10) не зависит от вектора Y_i^o, то она является и безусловной плотностью вероятности величины r_{ij}. Это значит, что случайные величины r_{ij} и r_{ij}^o, где

$$r_{ij}^o = \frac{a_{ij}^o}{\sqrt{a_{ii}^o a_{jj}}} ,$$

одинаково распределены. По следствию 1 случайные величины r_{ij} $(i,j=1,\ldots,p,\ i>j)$ не зависят от случайных величин a_{ii} $(i=1,\ldots,p)$. Тогда величины r_{ij}^o не зависят от случайных величин a_{jj} $(j=1,\ldots,p)$ и независимость коэффициентов корреляции r_{ij} в i-той строке (в i-том столбце) вытекает из независимости случайных величин a_{ij}^o в i-той строке (в i-том столбце), но эту независимость мы уже установили. Так как i-тая

строка взята произвольно ($1 \leq i \leq p$), то лемма 1 доказана.

ЛЕММА 2. Не учитывая симметрии, недиагональные элементы выборочной корреляционной матрицы R попарно независимы, если $P = I$.

Доказательство. Возьмем два элемента r_{ij} и r_{kl} из матрицы R. Если r_{ij} и r_{kl} находятся в одной строке или в одном столбце ($i = k$ или $j = l$) матрицы R, то по лемме 1 r_{ij} и r_{kl} независимы.

Если r_{ij} и r_{kl} находятся в разных строках и столбцах, то по следствию 1 эти случайные величины не зависят от a_{ii} ($i=1,\ldots,p$), а поскольку a_{ij} и a_{kl} являются суммами различных независимых случайных величин (см. равенство (5)), то они независимы, откуда вытекает и независимость случайных величин r_{ij} и r_{kl}. Лемма доказана.

3. Собственные значения и векторы выборочной корреляционной матрицы

При исследовании случайных матриц обычно особый интерес представляют собственные значения и векторы исследуемой матрицы. Известно, что для любой симметричной матрицы B порядка p существует ортогональная матрица C такая, что

$$B = C \Lambda C', \qquad (11)$$

где Λ – матрица собственных значений матрицы B:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & \lambda_p \end{pmatrix} = (\delta_{ij}\lambda_i)$$

(см. например [1], приложение), притом, не ограничивая общности, можно предположить, что $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Столбцы матрицы C, определенной равенством (11), являются нормированными собственными векторами матрицы B.

В книге [6] § 7.4 автор, опираясь на работу [7], показывает, что любую ортогональную матрицу C можно в общем случае представить через $p(p-1)/2$ углы вращения θ_{ij} ($i=1,\ldots,p-1$; $j=i,i+1,\ldots,p-1$) следующим образом:

$$O = \prod_{i=1}^{p} \prod_{j=i}^{p} R_j(\theta_{ij}),$$

где

$$R_j(\theta) = \begin{pmatrix} I_{j-1} & 0 & 0 \\ 0 & \begin{matrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{matrix} & 0 \\ 0 & 0 & I_{p-j-1} \end{pmatrix}.$$

Пусть B положительно определенная случайная матрица и $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. Представляя случайную ортогональную матрицу C через углы вращения θ_{ij}, можем смотреть на равенство (11) как на замену переменных: от элементов матрицы B (матрица B имеет в общем случае $p(p+1)/2$ разных элементов) перейдем к собственным значениям λ_i ($i=1,\ldots,p$) и углам вращения θ_{ij} ($i \leq j$, $i=1,\ldots,p-1$). Тумура [7] показал, что якобиан этого преобразования равен модулю выражения

$$\prod_{i=1}^{p} \prod_{j>i} (\lambda_i - \lambda_j) \prod_{i=1}^{p-2} \prod_{j=i}^{p-2} \sin^{p-j-1}\theta_{ij}. \qquad (12)$$

Интеграл от этого якобиана по переменным θ_{ij} равен

$$\frac{\prod_{i=1}^{p}\prod_{j>i}(\lambda_i - \lambda_j)\pi^{p(p+1)/4}}{\prod_{i=1}^{p}\Gamma(\frac{p-i+1}{2})}. \qquad (13)$$

В книге [6], стр. 441 автор показывает, что если матрица C представлена через углы вращения θ_{ij}, то плотность вероятности ее элементов дана формулой

$$\frac{\prod_{i=1}^{p}\Gamma(\frac{p-i+1}{2})}{\pi^{p(p+1)/4}}\prod_{i=1}^{p-2}\prod_{j=i}^{p-2}\sin^{p-j-1}\theta_{ij}. \qquad (14)$$

Сформулируем теперь следующий результат.

ТЕОРЕМА 2. Пусть

1) плотность вероятности f(B) случайной симметричной матрицы B_p инвариантна относительно всех ортогональных преобразований C, т.е. $f(B) = f(C'BC)$;

2) элементы на главной диагонали матрицы B константны.

Тогда совместная плотность вероятности собственных значений $\lambda_i(\lambda_1 > \lambda_2 > \ldots > \lambda_p)$ матрицы B равна в области изменения собственных значений матрицы B следующему выражению:

$$\frac{\pi^{p(p-1)/4}}{\prod_{i=1}^{p-1}\Gamma(\frac{p-i}{2})} f(\Lambda)\prod_{i=1}^{p-1}\prod_{j>i}(\lambda_i - \lambda_j), \qquad (15)$$

где $\Lambda = (\delta_{ij}\lambda_i)$, и нулю в противном случае.

Через $f(\Lambda)$ обозначим выражение плотности f(B) для диаго-

нальной матрицы $B = \Lambda$.

Доказательство. Известно (см. например [3], стр. 330), что плотность вероятности совместного распределения собственных значений $\lambda_i (\lambda_1 > \ldots > \lambda_p)$ симметричной матрицы B при предположении 1) настоящей теоремы имеет вид

$$\frac{\pi^{p(p+1)/4}}{\prod_{i=1}^{p} \Gamma(\frac{p-i+1}{2})} f(\Lambda) \prod_{i=1}^{p} \prod_{j>i} (\lambda_i - \lambda_j).$$

Доказательство этого результата исходит из того факта, что матрицу B можно представить в виде (11). У нас по условиям теоремы среди элементов матрицы B имеются в общем случае $p(p-1)/2$ различных невырожденных случайных величин. И хотя мы можем матрицу B представить в виде (11), на новые случайные переменные c_{ij} и λ_i налагают p дополнительных условий:

$$\sum_{k=1}^{p} c_{ik}^2 \lambda_k = b_{ii} \quad (i=1,\ldots,p).$$

Из этих равенств получим выражение для элементов c_{ij} $(i=1,\ldots,p)$ j-того столбца матрицы C через остальные ее элементы:

$$c_{ij} = \sqrt{\frac{b_{ii} - \sum_{\substack{k=1 \\ k\neq j}}^{p} c_{ik}^2 \lambda_k}{spB - \sum_{\substack{k=1 \\ k\neq j}}^{p} \lambda_k}} \quad (i=1,\ldots,p),$$

где через spB обозначим след матрицы B.

Это значит, что матрица C определена ортогональной мат-

рицей порядка (p-1), притом матрица Λ тоже определена (p-1) собственными значениями. То есть, матрицы C и Λ фактически определены в (p-1)-мерном подпространстве p-мерного пространства, где это подпространство определено дополнительными условиями. Тогда якобиан преобразования дан формулой (12), если в ней заменить p на (p-1); матрица C имеет распределение, плотность вероятности которого дана формулой (14), если в ней заменить p на (p-1). Учитывая еще формулу (13), получим интересующий нас результат.

СЛЕДСТВИЕ 2. Плотность вероятности совместного распределения собственных значений $\lambda_1(\lambda_1 > \ldots > \lambda_p)$ выборочной корреляционной матрицы R выражается следующей формулой в области изменения собственных значений матрицы R:

$$\frac{\Gamma^p(\frac{n}{2})\prod_{i=1}^{p}\lambda^{(n-p-1)/2}\prod_{i=1}^{p-1}\prod_{j>i}(\lambda_i-\lambda_j)}{\prod_{i=1}^{p}\Gamma(\frac{n-i+1}{2})\prod_{i=1}^{p-1}\Gamma(\frac{p-i}{2})}, \qquad (16)$$

если $P = I$.

Доказательство. Возьмем в качестве матрицы B выборочную корреляционную матрицу R. Тогда она представится в виде

$$R = C\Lambda C'. \qquad (17)$$

Так как

$$\frac{\Gamma_p(\frac{n}{2})\,|C'RC|^{(n-p-1)/2}}{\prod_{i=1}^{p}\Gamma(\frac{n-i+1}{2})\,\pi^{p(p-1)/4}} = \frac{\Gamma_p(\frac{n}{2})(|C'||R||C|)^{(n-p-1)/2}}{\prod_{i=1}^{p}\Gamma(\frac{n-i+1}{2})\,\pi^{p(p-1)/4}} =$$

$$= \frac{\Gamma_p(\frac{n}{2})\,|R|^{(n-p-1)/2}}{\prod_{i=1}^{p}\Gamma(\frac{n-i+1}{2})\,\pi^{p(p-1)/4}},$$

то плотность вероятности (8) инвариантна относительно ортогональных преобразований, и теорема 2 применима к этому случаю. Тогда из выражений (8) и (15) сразу получим наше утверждение и этим следствие доказано.

Итак, мы установили формулу плотности вероятности собственных значений выборочной корреляционной матрицы при нормальной выборке, когда теоретическая ковариационная матрица диагональная. Из формул (8) и (12) с помощью доказательства теоремы 2 получим формулу для плотности вероятности совместного распределения элементов случайных матриц Λ и C, если предположим, что матрица C представлена через углы вращения θ_{ij} $(i,j=1,\ldots,p-1,\ i \geq j)$:

$$\frac{\Gamma_p(\frac{n}{2})\prod_{i=1}^{p}\lambda_i^{(n-p-1)/2}}{\prod_{i=1}^{p}\Gamma(\frac{n-i+1}{2})\,\pi^{p(p-1)/4}}\prod_{i=1}^{p-1}\prod_{j>i}(\lambda_i-\lambda_j)\prod_{i=1}^{p-3}\prod_{j=i}^{p-3}\sin^{p-j-2}\theta_{ij}.$$

Отсюда видно, что собственные значения λ_i матрицы R статистически не зависят от ее собственных векторов, плотность ве-

роятности которых дана формулой (14), если в ней заменить p на (p-1). В итоге можем сформулировать следующий результат.

СЛЕДСТВИЕ 3. Если C - матрица, столбцами которой являются собственные векторы выборочной корреляционной матрицы R_p, то плотность вероятности этой матрицы дана формулой

$$\frac{\pi^{p(p-1)/4} \prod_{i=1}^{p-3} \prod_{j=1}^{p-3} \sin^{p-j-2}\theta_{ij}}{\prod_{i=1}^{p-1} \Gamma(\frac{p-i}{2})} \qquad (18)$$

и элементы матрицы C статистически независимы от собственных значений матрицы R, если P = I.

4. Асимптотическое поведение собственных значений выборочной корреляционной матрицы

При применении статистических методов в практике часто порядки случайных матриц бывают сравнительно большими и непосредственное вероятностное исследование этих матриц - очень трудоемкая задача из-за громоздких вычислений. Поэтому, особый интерес представляют разные асимптотические распределения и оценки в предельном случае, когда порядок матрицы стремится к бесконечности. Для изложения асимптотической теории нам выгодно сначала ввести еще некоторые обозначения и определения.

Пусть B - (m x n)-матрица. Через vee B обозначим (mn x 1)-матрицу, которую получим из B, подставляя ее столбцы один под другим в естественном порядке. Кронекеровским произведе-

нием $(m \times n)$-матрицы B и $(p \times q)$-матрицы C называется блок-матрица

$$B \otimes C = [Bc_{ik}]$$

с mp-строками и nq-столбцами. Переставленной единичной матрицей $I_{(m,n)}$ называется $(mn \times mn)$-матрица, составленная из таких $(m \times n)$-подматриц, что в (ij)-той подматрице (ji)-тый элемент равен единице, а остальные все нули. Если нам дана $(m \times n)$-матрица B и $(p \times q)$-матрица C, то

$$I_{(p,m)}(B \otimes C) = (C \otimes B)I_{(q,n)}.$$

В дальнейшем через $\xrightarrow{\mathcal{L}}$ обозначим сходимость по распределению. Приведем теперь два известных результата из [4].

ЛЕММА 3. Пусть $A \sim W_p(n,\Sigma)$. Тогда, если $n \to \infty$, то

$$\sqrt{n}\ \mathrm{vec}(S - \Sigma) \xrightarrow{\mathcal{L}} N(0,U_{p^2}),$$

где

$$U_{p^2} = (I_{p^2} + I_{(p,p)})(\Sigma \otimes \Sigma). \qquad (19)$$

ЛЕММА 4. Пусть $\{\bar{X}_n\}$ - последовательность p-мерных случайных векторов. Предположим, что при $n \to \infty$

$$\sqrt{n}\ (\bar{X}_n - \vec{\mu}) \xrightarrow{\mathcal{L}} N(0,\Psi_p),$$

где Ψ_p - постоянная матрица и $\vec{\mu}$ - постоянный вектор. Пусть $g = g(X)$ - q-мерный случайный вектор, чьи координаты имеют непрерывную производную в окрестности точки $X = \vec{\mu}$. Тогда, если $n \to \infty$, то

$$\sqrt{n}\,[g(X_n) - g(\vec{\mu})] \xrightarrow{\mathcal{L}} N(0, \xi' \Psi \xi),$$

где

$$\xi = \left(\frac{\partial g(X)}{\partial X}\right)_{X=\vec{\mu}}$$

есть $(p \times q)$-матрица, составленная из частных производных первого порядка от $g(X)$ по X в точке $\vec{\mu}$.

Используя приведенные две леммы можно легко доказать следующее утверждение.

ЛЕММА 5. Если $P = I$, то

$$\sqrt{n}\ \mathrm{vec}(R - P) \xrightarrow{\mathcal{L}} N(0, \xi' U_{p^2} \xi),$$

где

$$U_{p^2} = (I_{p^2} + I_{(p,p)})(\Sigma \otimes \Sigma)$$

и $(p^2 \times p^2)$-матрица ξ определяется равенством

$$\xi = \left(\frac{\partial\ \mathrm{vec}\ R}{\partial\ \mathrm{vec}\ S}\right)_{\mathrm{vec}\ S = \mathrm{vec}\ \Sigma}.$$

Доказательство. Из формул (5) и (7) видно, что

$$R = (\delta_{ij} s_{ii})^{-\frac{1}{2}} S (\delta_{ij} s_{ii})^{-\frac{1}{2}}. \qquad (20)$$

По лемме 3 распределение матрицы $\sqrt{n}\,(S - \Sigma)$ асимптотически нормально, если $n \to \infty$. Но учитывая равенство (20) можем применить лемму 4, где случайным вектором X_n возьмем $\mathrm{vec}\,(S - \Sigma)$. Тогда $\Psi = U_{p^2}$ и функция g определена формулой (20). В таком случае по лемме 4 при $n \to \infty$,

$$\sqrt{n}\,(R - P) \xrightarrow{\mathcal{L}} N(0, \xi' U_{p^2} \xi),$$

где

$$\xi = \left(\frac{\partial \operatorname{vec} R}{\partial \operatorname{vec} S}\right)_{\operatorname{vec} S = \operatorname{vec} \Sigma},$$

а этим лемма доказана.

Приведем еще один вспомогательный результат, следуя Барра (см. [2], стр. 107).

ЛЕММА 6. Пусть $X_1, \ldots, X_k$ случайные попарно независимые векторы. Если вектор

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$$

нормальный, то векторы $X_1, \ldots, X_k$ независимы.

Если $n \to \infty$, то распределение вектора $\sqrt{n}\ \operatorname{vec}(R-P)$ приблизится к распределению случайного вектора $\operatorname{vec} V$, где

$$\operatorname{vec} V \sim N(0, \xi' U_{p^2} \xi).$$

Так как по лемме 2 элементы матрицы R попарно независимы, то, очевидно, попарно независимы и элементы матрицы V, если не учитывать симметричность. Но тогда по лемме 6 элементы матрицы V независимы. Обозначим собственные значения матрицы V через $\lambda_{1p} \geq \ldots \geq \lambda_{pp}$. Нормированной спектральной функцией $\mathcal{F}_p(x)$ матрицы V называется следующая случайная функция:

$$\mathcal{N}_p(x) = \frac{1}{p} \sum_{i=1}^{p} \theta(x - \lambda_{ip}),$$

где

$$\theta(y) = \begin{cases} 1, & y > 0; \\ 0, & y \leq 0. \end{cases}$$

При каждом фиксированном значении x функция $\mathcal{N}_p(x)$ является случайной величиной, сложным образом выражающейся через элементы матрицы V. Нас интересует поведение $\mathcal{N}_p(x)$, если $p \to \infty$. Исходя из матрицы $V = (v_{ij})$ составим последовательность матриц $\{V^{(p)}\}$, где

$$V^{(p)} = (v_{ij}^{(p)}) = \frac{V}{\sqrt{p}}.$$

Тогда

$$Ev_{ij}^{(p)} = 0$$

и, если $i \neq j$

$$Dv_{ij}^{(p)} = \frac{Dv_{ij}}{p} = \frac{1}{p},$$

поскольку

$$Dv_{ij} = \lim_{n\to\infty} D(\sqrt{n}\, r_{ij}) = \lim_{n\to\infty} (nDr_{ij}) = \lim_{n\to\infty} n \cdot \frac{1}{n} = 1$$

(в том, что $Dr_{ij} = \frac{1}{n}$ можно легко убедиться, вычислив второй момент случайной величины r_{ij} с помощью плотности вероятности (10)). При помощи формулы (10) можем проверить, что и все остальные моменты элементов матрицы $V^{(p)}$ конечны.

Теперь можем сформулировать полукруговый закон Вигнера для матриц $\{V^{(p)}\}$, так как все предположения этой теоремы

выполнены (о полукруговом законе Вигнера см. например [3], [4], [8]): достаточно требовать, что а) для каждого p (p=1,2,...) случайные величины $v_{ij}^{(p)}$ ($i \geq j$, $i,j=1,\dots,p$) независимы; б) $Ev_{ij}^{(p)} = 0$; в) $Dv_{ij}^{(p)} = \frac{1}{p}$, $i \neq j$; г) $E\left[v_{ij}^{(p)}\right]^4 < \infty$.

ТЕОРЕМА 3. Если $p \to \infty$, то нормированная спектральная функция $\mathcal{N}_p(x)$ матрицы $V^{(p)}$ сходится по вероятности к неслучайной функции распределения $G(x)$:

$$G(x) = \begin{cases} 0, & x \leq -2; \\ \frac{1}{\pi} \int\limits_{-2}^{x} \sqrt{1 - \frac{y^2}{4}}\, dy, & |x| < 2; \\ 1, & x \geq 2. \end{cases}$$

Сходимость по вероятности значит здесь, что последовательность случайных величин $\{\mathcal{N}_p(x)\}$ сходится по вероятности к числу $G(x)$ для каждого вещественного числа x.

Литература

1. Андерсон Т., Введение в многомерный статистический анализ. М., 1963.
2. Барра Ж.-Р., Основные понятия математической статистики. М., 1974.
3. Гирко В.Л., Случайные матрицы. Киев, 1975.
4. Arnold, L., On the Asymptotic Distribution of the Eigenvalues of Random Matrices. Journal of Mathematical Analysis and Applications, 1967, 20, 262-268.
5. Izenman, A.J., Reduced-Rank Regression for the Multivariate Linear Model. Journal of Multivariate Analysis, 1975, 5, 248-264.
6. Kshirsagar, A.M., Multivariate Analysis. N.-Y., 1972.
7. Tumura, Y., The Distribution of Latent Roots and Vectors. Tokyo Rica University, Mathematics, 1965, 1.
8. Wigner, E.P., On the Distribution of the Roots of Certain Symmetric Matrices. Annals of Mathematics, 1958, 67, 325-326.

ЗАДАЧА ДИСПЕРСИОННОГО АНАЛИЗА ДЛЯ ЭВМ В СЛУЧАЕ СЛУЧАЙНОЙ МОДЕЛИ

Т.А.Кельдер

1. Введение

Целью статьи является ознакомление метода дисперсионного анализа, реализованного в рамках системы статистической обработки данных на ЭВМ "Минск-32" в ВЦ ТГУ. Так как рассматриваемая система ориентирована на данные анкетного типа, то выбраны методы решения, пригодные для большого количества факторов и дающие результаты в случае несбалансированных данных.

Во втором пункте статьи описывается постановка задачи и основные ограничения для модели. В третьем дается метод получения несмещенных точечных оценок для компонентов дисперсии, который является основным методом в данной системе дисперсионного анализа. Далее излагается метод проверки гипотез о влиянии изучаемых факторов, который применим для данных, мало отличающихся от сбалансированных.

2. Постановка задачи

Предположим, что для изучения влияния факторов A_i, $i=1,\dots,k$, с уровнями соответственно $1,\dots,n_i$, $i=1,\dots,k$, сделано N наблюдений. Измеренные результаты определяют N-(дименсиональный) вектор

$$\vec{Y} = \begin{pmatrix} y_1 \\ \dots \\ y_N \end{pmatrix},$$

который мы в дальнейшем назовем вектором наблюдений. Основным предположением является линейная модель

$$\vec{Y} = \vec{1}\mu + \sum_{i=1}^{k} X_{A_i} \vec{\beta}_{A_i} + \vec{e}, \qquad (1)$$

где $\vec{1}$ – N-вектор, составленный из единиц, μ – неизвестная константа, случайные n_i-векторы $\vec{\beta}_{A_i}$ выражают влияние факторов A_i, $i=1,\dots,k$, случайный N-вектор $\vec{e}$ – вектор случайных ошибок, а X_{A_i} – известные $(N \times n_i)$-матрицы плана, которые определяют, на каком уровне каждого фактора получено соответственное наблюдение (компонент вектора $\vec{Y}$). Если в случае произвольного фактора A_i количество наблюдений на каждом его уровне одинаково, то данные называются сбалансированными. Заметим, что некоторые факторы A_i в модели (1) могут в действительности быть взаимодействиями факторов $A_{i_1},\dots,A_{i_m}$. В модели (1) предполагаются все факторы случайными, т.е. уровни этих факторов выбраны случайным образом из общей (бесконечной) совокупности уровней каждого фактора.

Для модели (1) предполагается дополнительно, что:

а) компоненты вектора $\vec{Y}$ распределены нормально с общим средним μ и дисперсией σ^2;

б) случайные векторы $\vec{\beta}_{A_i}$, $i=1,\dots,k$, и $\vec{e}$ взаимно независимы и распределены как

$$\begin{cases} \vec{\beta}_{A_i} \sim N(\vec{0};\sigma^2_{A_i} J_{n_i}), & i=1,\dots,k, \\ \vec{e} \sim N(\vec{0};\sigma^2_e J_N) , \end{cases} \qquad (2)$$

где J_m единичная матрица порядка m.

При предположениях (2) получим, что

$$\sigma^2 = \sum_{i=1}^{k} \sigma^2_{A_i} + \sigma^2_e,$$

поэтому величины $\sigma^2_{A_i}$, $i=1,\dots,k$, и σ^2_e называются компонентами дисперсии.

Для модели (1) с ограничениями (2) решаются следующие две задачи:

1. Находятся несмещенные точечные оценки для дисперсионных компонентов $\sigma^2_{A_i}$, $i=1,\dots,k$, и σ^2_e.

2. Проверяются гипотезы $\sigma^2_{A_i} = 0$, $i=1,\dots,k$, т.е. существенность факторов A_i.

3. Метод нахождения точечных оценок

Для получения точечных оценок для компонентов дисперсии обычно приравнивают некоторые квадратичные формы $\vec{Y}'Q\vec{Y} = SS$ от вектора наблюдений к их математическим ожиданиям, которые вычисляются из общей формулы

$$E(\vec{Y}'Q\vec{Y}) = tr(QV) + \vec{\mu}'Q\vec{\mu} , \qquad (3)$$

где $\vec{\mu}$ и V являются соответственно вектором математических ожиданий и ковариационной матрицей случайного вектора $\vec{Y}$, а tr(X) означает след матрицы X.

В предположениях (1) и (2) получаем

$$E(\vec{Y}'Q\vec{Y}) = \mu^2\vec{1}'Q\vec{1} + \sum_{i=1}^{k} \sigma^2_{A_i} tr(QX_{A_i}X'_{A_i}) + \sigma^2_e tr(Q). \qquad (4)$$

Выбор класса квадратичных форм для получения точечных оценок зависит от конкретного метода (см. [6]). В настоящей системе для дисперсионного анализа используется первый метод Хендерсона (см. [4], [7] или [6], стр. 34-46). Преимущества данного метода следующие.

1. Полученные оценки для дисперсионных компонентов являются несмещенными, независимо от сбалансированности данных.

2. В случае сбалансированных данных рассматриваемые квадратичные формы являются обычными "суммами квадратов" SS, которых вычисляют в дисперсионном анализе.

3. Требуемые квадратичные формы и их математические ожидания можно вычислять непосредственно, не пользуясь подпрограммами матричной алгебры. Это очень существенно, ибо порядок матрицы Q может быть очень большим.

Придержимся в дальнейшем следующего вида обозначений. Обозначим количество и сумму наблюдений на уровне m фактора A_i соответственно $n(A_i,m)$ и $y(A_i,m)$. Количество наблюдений на уровне m фактора A_i и на уровне r фактора A_j обозначим через $n(A_i,m;A_j,r)$. Пусть I_k является $(k \times k)$-матрицей, составленной из единиц, а J_k, как прежде, $(k \times k)$-единичной матрицей. Пусть сумма $\sum_{i=1}^{k}{}^{+} M_i$ обозначает прямую сумму матриц M_i,

$i=1,\dots,k$.

Для первого метода Хендерсона вычисляют квадратичные формы

$$\begin{cases} T_{A_i} = \sum\limits_{m=1}^{n_i} \dfrac{y(A_i,m)^2}{n(A_i,m)}, \quad i=1,\dots,k, \\ T_\mu = \dfrac{1}{N}\left(\sum\limits_{m=1}^{N} y_m\right)^2, \\ T_0 = \sum\limits_{m=1}^{N} y_m^2 . \end{cases} \tag{5}$$

Если фактор A_i является взаимодействием каких-то факторов $A_{i_1},\dots,A_{i_m}$, то квадратичная форма T_{A_i} иногда обозначается и через $T(A_{i_1},\dots,A_{i_m})$.

Для нахождения математических ожиданий квадратичных форм (5) выпишем их матрицы. Нетрудно видеть, что $T_0 = \vec{Y}'Q_Q\vec{Y}$, где $Q_0 = J_N$ и $T_\mu = \vec{Y}'Q_\mu\vec{Y}$, где $Q_\mu = \frac{1}{N}I_N$. Далее, если для произвольного фактора A_i упорядочить компоненты вектора $\vec{Y}$ по порядку уровней фактора A_i (от этого не изменяется значение квадратичной формы T_{A_i}), то $T_{A_i} = \vec{Y}'Q_{A_i}\vec{Y}$, где

$$Q_{A_i} = \sum_{m=1}^{n_i} \frac{1}{n(A_i,m)} I_{n(A_i,m)} .$$

Используя формулу (4) получим математические ожидания квадратичных форм (5)

$$
\begin{cases}
E(T_{A_i}) = N\mu^2 + \sum_{j=1}^{k}\left[\sum_{m=1}^{n_i}\dfrac{\sum_{r=1}^{n_j} n(A_i,m;A_j,r)^2}{n(A_i,m)}\sigma^2_{A_j}\right] + n_i\sigma^2_e \quad (i=1,\dots,k), \\
E(T_\mu) = N\mu^2 + \sum_{i=1}^{k}\left[\sum_{j=1}^{n_i} n(A_i,j)^2\right]\sigma^2_{A_i} \,/\, N + \sigma^2_e\,, \\
E(T_0) = N\mu^2 + N\sum_{i=1}^{k}\sigma^2_{A_i} + N\sigma^2_e\,.
\end{cases}
\tag{6}
$$

Пусть p - порядок наивысшего вычисляемого взаимодействия. Найдем линейные комбинации $SS(A_{i_1},\dots,A_{i_m})$ (или SS_{A_i}, если не имеет значения, что A_i - взаимодействие) из квадратичных форм (5)

$$
\begin{cases}
SS(A_{i_1},\dots,A_{i_m}) = \sum_{j=0}^{m-1}(-1)^j \sum_{\{A_{r_1},\dots,A_{r_{m-j}}\}\subset\{A_{i_1},\dots,A_{i_m}\}} T(A_{r_1},\dots,A_{r_{m-j}}) + (-1)^m T_\mu, \\
\text{где } \{A_{i_1},\dots,A_{i_m}\}\subset\{A_1,\dots,A_k\} \text{ и } m=1,\dots,p, \\
SS_e = T_0 - T_\mu - \sum_{m=1}^{p}\sum_{\{A_{i_1},\dots,A_{i_m}\}\subset\{A_1,\dots,A_k\}} SS(A_{i_1},\dots,A_{i_m}).
\end{cases}
\tag{7}
$$

Те самые линейные комбинации вычисляем и для математических ожиданий (6). Приравнивая квадратичные формы (7) к их математическим ожиданиям, получаем систему линейных уровнений относительно дисперсионных компонентов $\sigma^2_{A_i}$, $i=1,\dots,k$, и σ^2_e. Решения данной системы являются несмещенными оценками дис-

персионных компонентов.

В случае сбалансированных данных квадратичные формы (7) являются обыкновенными "суммами квадратов" для дисперсионного анализа, но для несбалансированных данных они не обязательно положительно определены, т.е. некоторые из них могут иметь отрицательные значения.

Может оказаться (даже для сбалансированных данных), что часть оценок дисперсионных компонентов отрицательна. Для интерпретации такого случая имеется несколько возможностей, например:

1) действительные значения данных компонентов равны 0;
2) не достает данных для анализа всех факторов;
3) не выполнены некоторые из основных предположений в (1) и (2).

4. Проверка гипотез

Для проверки гипотез о влиянии факторов A_i, $i=1,\ldots,k$, т.е. гипотез $\sigma^2_{A_i} = 0$, мы должны знать распределения квадратичных форм (7). Известно (см. [1], стр. 314–351), что в случае сбалансированных данных эти квадратичные формы SS_{A_i} распределены как $E(SS_{A_i})\chi^2(f_{A_i})$, а число степеней свободы f_{A_i} является коэффициентом σ^2_e в математическом ожидании данной квадратичной формы. Но в случае несбалансированных данных это, в общем, не так, и точные распределения квадратичных форм (7) не известны. Мы можем только предполагать, что если данные мало отличаются от сбалансированных, то распределения квадратичных форм (7) мало отличаются от распределе-

ния типа хи-квадрат.

Для проверки гипотез применяется приближенный F-метод Саттертвайта (см. [1], стр. 349-351 или [3]), сущность которого состоит в следующем.

Пусть для каких-то квадратичных форм SS_{A_i} и SS_{A_j} мы имеем $E(SS_{A_i}) = \lambda\sigma^2_{A_1} + \tau$, а $E(SS_{A_j}) = \tau$. Тогда для проверки гипотезы $\sigma^2_{A_1} = 0$ можно использовать статистику

$$\frac{SS_{A_i}}{f_{A_i}} \Big/ \frac{SS_{A_j}}{f_{A_j}},$$

которая при верности гипотезы $\sigma^2_{A_1} = 0$ имеет F-распределение $F(f_{A_i}, f_{A_j})$. Но подходящих квадратичных форм, особенно в случае несбалансированных данных, для каждого $\sigma^2_{A_i}$, $i=1,\ldots,k$, нет. Саттертвайт показал, что если $MS_i \sim 1/f_i E(MS_i)\,\chi^2(f_i)$ и $MS = \sum_{i=1}^{n} \alpha_i MS_i$, то MS приблизительно распределено как $1/fE(MS)\chi^2(f)$, где f оценивается по формуле

$$\hat{f} = \frac{MS^2}{\sum_{i=1}^{n} \frac{(\alpha_i MS_i)^2}{f_i}} \qquad (8)$$

и это приближение отличное, если все $\alpha_i \geq 0$.

Для проверки гипотез мы можем использовать подходящие линейные комбинации квадратичных форм SS_{A_i} / f_{A_i}, отношение которых будет иметь приблизительное F-распределение $F(\hat{f}_1, \hat{f}_2)$, где $\hat{f}_1$ и $\hat{f}_2$ вычисляется по формуле (8).

В случае, когда некоторые α_i отрицательны, формулу (8) следует использовать с ограничениями. Гейлор и Хоппер (см. [3]) исследовали пригодность формулу Саттертвайта в случае

разницы двух квадратичных форм $MS_1 - MS_2$ (подходящим образом перегруппируя слагаемые в линейной комбинации, мы можем всегда добиться такого положения, при котором необходимо выполнить только одно действие вычитания), со степенями свободы f_1 и f_2. Они показали, что формула Саттертвайта пригодна для аппроксимации распределения разницы двух квадратичных форм с χ^2-распределением, если

$$\rho = \frac{E(MS_1)}{E(MS_2)} \geq F^{\alpha}(f_2, f_1).$$

Здесь $F^{\alpha}(f_2,f_1)$ – верхняя α-точка распределения F с f_2 и f_1 степенями свободы, а максимальная величина коэффициента α выбирается, исходя из следующих условий.

1. Если $f_1 \leq 10$, то $\alpha = 0.05$.
2. Если $10 < f_1 < 100$ и $f_2 > f_1/2$, то $\alpha = 0.025$.
3. Если $10 < f_1 \leq 20$, а f_2 произвольное или $20 < f_1 < 100$ и $f_2 > f_1/5$, то $\alpha = 0.01$.

Для остальных случаев формула Саттертвайта не пригодна. Так как отношение ρ обычно неизвестно, то Гейлор и Хоппер предлагают критерий

$$MS_1 / MS_2 \geq F^{\alpha}(f_2,f_1) \times F^{0.5}(f_1,f_2)$$

для проверки пригодности формулы Саттертвайта в случае разницы двух квадратичных форм.

В данной системе процентные точки F-распределения для приближенного F-критерия и критерия Гейлора и Хоппера вычисляются кумулянтными формулами порядка 7 Корниша и Фишера

(см. [2] или [5]). Как показали Сахай и Томпсон (см. [5]), эти формулы дают наилучшее приближение для F-распределения в сравнении с другими известными приближениями F распределения. Как показали практические расчеты для большинства случаев относительная ошибка при применении этих формул меньше чем 0.5% .

В данной системе, кроме точечных оценок для дисперсионных компонент и результатов применения F-критерия, выдается и обыкновенная таблица дисперсионного анализа.

В случае обнаружения влияния какого-то из факторов есть возможность дальнейшего исследования влияния факторов при помощи S-метода Шеффе (см. [1], стр. 101-111).

Литература

1. Шеффе Г., Дисперсионный анализ. М., 1963.
2. Fisher, R.A., Cornish, E.A., The Percentile points of Distributions Having known Cumulants. Technometrics, 1960, 2, 209-226.
3. Gaylor, D.W., Hopper, F.W., Estimating the Degrees of Freedom for Linear Combinations of Mean Squares by Satterthwaite's Formula. Technometrics, 1969, 11, № 4, 691-706.
4. Henderson, C.R., Estimation of Variance and Covariance Components. Biometrics, 1953, 9, 226-252.
5. Sahai, H., Thompson, W.O., Comparisons of Approximations to the Percentiles of the t-, χ^2- and F-distributions. J. Statist. Comput. Simul., 1974, 3, 81-93.
6. Searle, S.R., Topics in Variance Components Estimation. Biometrics, 1971, 27, № 1, 1-76.
7. Searle, S.R., Henderson, C.R., Another look of Henderson's Methods of estimating Variance Components. Biometrics, 1968, 24, № 4, 749-772.

МЕТОДИКА ЛИНЕЙНОГО АНАЛИЗА ВЫСОКОМЕРНЫХ ПРИЗНАК-ВЕКТОРОВ С ЭЛИМИНИРОВАНИЕМ "МЕШАЮЩИХ" ПРИЗНАКОВ

Э.А. Тийт, Х.Р. Ридала, П.Э. Ридала

При обработке данных, описываемых с помощью очень большого числа признаков, возникают следующие задачи:

1° сокращение числа изучаемых признаков при условии, что сохраняется требуемая часть информации, передаваемой этими признаками;

2° элиминирование влияния таких признаков, включение которых в модель нежелательно;

3° учитывание априорной группировки признаков (по их содержанию);

4° нахождение показателей, характеризующих статистическую связь между отдельными группами признаков.

Для решения отдельных задач из этого списка имеются некоторые общеизвестные методы. Например, сократить числа признаков можно методом компонентного анализа (см. напр. [1], стр. 530); корреляции между двумя группами признаков измеряются при помощи канонического анализа (см. напр. [1], стр. 522) и т.д.

Однако, пока нет методики, которая решила бы комплексно

все вышеописанные задачи.

Авторы настоящей заметки сделали попытку выработать методику статистического анализа, реализующую комплексно все упомянутые задачи на базе линейных моделей.

Хотя линейные модели являются оптимальными только в случае совместного нормального распределения всех рассматриваемых признаков, они дают достаточно хорошее приближение к оптимальной модели при обработке сравнительно неточно измеренных данных. Такими являются, например, результаты разных социологических, психологических, педагогических, медицинских исследований. Существенное преимущество основанных на линейных моделях методик обработки данных, выявляется при их реализации. Именно, для них исходным материалом является корреляционная матрица исследуемых признаков, и таким образом, объем исходной информации не зависит от количества исследуемых объектов (объема выборки), которое при таких исследованиях может быть очень большим (до порядка десяти тысяч).

Для изложения вышеописанной методики вводим следующие понятия и обозначения.

Пусть X – исследуемый признак-вектор, $X' = (X_1,\ldots,X_m)$. Заданы подвекторы (см. [2]) $X(I_j)$ $(j=1,\ldots,k)$ этого вектора, притом выполняются следующие условия:

$$I_i \cap I_j = \emptyset \quad (i \neq j),$$
$$I_i \subset 1_m \quad (i=1,\ldots,k).$$

Предполагается, что каждый подвектор $X(I_j)$ состоит из содержательно связанных признаков, и для него существует т.н. ме-

шающий подвектор $X(I_j^0)$, притом имеет место равенство

$$I_j \cap I_j^0 = \emptyset \quad (j=1,\ldots,k).$$

Заметим, что мешающие подвекторы $X(I_j^0)$ могут быть различными, а также частично или даже полностью совпадать. Некоторые (или даже все) мешающие подвекторы могут быть пустыми:

$$I_{j_i}^0 = \emptyset \quad (i=1,\ldots,l; \quad i \leq k).$$

Состав каждого подвектора (группы признаков) $X(I_j)$ и соответствующего ему мешающего подвектора $X(I_j^0)$ $(j=1,\ldots,k)$ считается известным априори, до начала исследования.

Пусть Y - вектор, полученный из вектора X центрированием и нормированием:

$$Y = D^{-0,5}(X - EX)D^{-0,5}, \qquad (1)$$

где EX - вектор средних значений X, $EX' = (EX_1,\ldots,EX_m)$, а $D = diag(E(X-EX)(X-EX)')$ - диагональная матрица дисперсий вектора X.

Если рассмотреть X как вектор, имеющий эмпирическое распределение, заданное конкретной выборкой, то и вектор Y определяется при помощи выборочных средних и дисперсий: $EX = (\bar{x}_1,\ldots,\bar{x}_m)$,

$$D = \begin{pmatrix} \bar{\sigma}_1 & 0 & \ldots & 0 \\ 0 & \bar{\sigma}_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \bar{\sigma}_m \end{pmatrix}.$$

Тогда и все полученные в дальнейшем характеристики являются выборочными, и их можно рассматривать как точечные

оценки соответствующих генеральных характеристик.

Обозначим корреляционную матрицу исходного признак-вектора X символом R, $R = EYY'$, и введем следующие символы для взаимных корреляционных матриц отдельных подвекторов:

$$R_{ij} = EY(I_i)Y'(I_j), \quad R_{ij^o} = EY(I_i)Y'(I_j^o),$$

$$R_{i^oj} = EY(I_i^o)Y'(I_j), \quad R_{i^oj^o} = EY(I_i^o)Y'(I_j^o) \quad (i,j=1,\ldots,k).$$

В частных случаях $i = j$ получаются автокорреляционные матрицы R_{ii} и $R_{i^oi^o}$ $(i=1,\ldots,k)$.

Определим вектор $\tilde{X}$, $\tilde{X}' = (X'(I_1) : X'(I_1^o) : X'(I_2) : X'(I_2^o) : \ldots : X'(I_k) : X'(I_k^o))$ и соответствующий центрированный и нормированный вектор $\tilde{Y}$. Корреляционная матрица $\tilde{R} = E\tilde{Y}\tilde{Y}'$ вектора $\tilde{X}$ имеет следующую блочную структуру:

$$\tilde{R} = \left(\begin{array}{c|c|c|c|c|c|c}
R_{11} & R_{11^o} & R_{12} & R_{12^o} & \cdots & R_{1k} & R_{1k^o} \\
\hline
R'_{11^o} & R_{1^o1^o} & R_{1^o2} & R_{1^o2^o} & \cdots & R_{1^ok} & R_{1^ok^o} \\
\hline
R'_{12} & R'_{1^o2} & R_{22} & R_{22^o} & \cdots & R_{2k} & R_{2k^o} \\
\hline
R'_{12^o} & R'_{1^o2^o} & R'_{22^o} & R_{2^o2^o} & \cdots & R_{2^ok} & R_{2^ok^o} \\
\hline
\multicolumn{7}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \\
\hline
R'_{1k} & R'_{1^ok} & R'_{2k} & R'_{2^ok} & \cdots & R_{kk} & R_{kk^o} \\
\hline
R'_{1k^o} & R'_{1^ok^o} & R'_{2k^o} & R'_{2^ok^o} & \cdots & R'_{kk^o} & R_{k^ok^o}
\end{array}\right).$$

Заметим, что так как мешающие подвекторы не обязательно взаимно исключающие, то порядки m и $\tilde{m}$ матриц R и $\tilde{R}$ связаны неравенством

$$\tilde{m} \geq m.$$

Дальнейший анализ проводим сперва для признак-вектора Y.

Первым шагом является элиминирование линейного влияния мешающих признак-векторов. Это осуществляется путем вычисления остатков прогноза $\hat{Y}(I_j)$ векторов $Y(I_j)$ при их линейном прогнозировании через мешающие подвекторы $Y(I_j^o)$:

$$\hat{Y}(I_j) = Y(I_j) - B_j Y(I_j^o), \qquad (2)$$

где

$$B_j = R_{jj^o} R^-_{j^o j^o} \quad (j=1,\dots,k). \qquad (3)$$

Здесь $R^-_{j^o j^o}$ – обобщенная обратная матрица матрицы $R_{j^o j^o}$. Если $R_{j^o j^o}$ не сингулярна (в практических задачах в большинстве случаев так и бывает), то $R^-_{j^o j^o}$ является обычной обратной матрицей $R^{-1}_{j^o j^o}$ матрицы $R_{j^o j^o}$.

Определим теперь признак-вектор $\hat{Y}$ остатков прогноза, $\hat{Y}' = (\hat{Y}'(I_1) \vdots \hat{Y}'(I_2) \vdots \dots \vdots \hat{Y}'(I_k))$ и вычислим его корреляционную матрицу $\hat{R}$:

$$\hat{R} = \begin{pmatrix} \hat{R}_{11} & \hat{R}_{12} & \cdots & \hat{R}_{1k} \\ \hat{R}'_{12} & \hat{R}_{22} & \cdots & \hat{R}_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{R}'_{1k} & \hat{R}'_{2k} & \cdots & \hat{R}_{kk} \end{pmatrix}.$$

Порядок $\hat{m}$ матрицы $\hat{R}$ удовлетворяет неравенству $\hat{m} \leq m$, а блоки выражаются через блоки исходной матрицы, как легко проверить посредством формул (2) и (3), следующим образом:

$$\begin{cases} \hat{R}_{ij} = H_i^{-0,5}(R_{ij} - R_{ii} \circ R_{i\circ i}^{-} \circ R_{i\circ j} - R_{ij} \circ R_{j\circ j}^{-} \circ R_{j\circ j} + \\ \quad + R_{ii} \circ R_{i\circ i}^{-} \circ R_{i\circ j} \circ R_{j\circ j}^{-} \circ R_{j\circ j}) H_j^{-0,5}, \\ \\ H_q = (R_{qq} - R_{qq} \circ R_{q\circ q}^{-} \circ R_{q\circ q}) \quad (q=i,j;\ i,j=1,\ldots,k). \end{cases} \tag{4}$$

В частном случае, когда $I_j^o = \emptyset$, считаем[1] $R_{j\circ j\circ} = \mathfrak{I}$ и $R_{ij\circ} = 0$ $(i=1,\ldots,k)$.

Автокорреляционная блок-матрица $\hat{R}_{ii}$ $(i=1,\ldots,k)$ является матрицей частных коэффициентов корреляции подвектора $Y(I_i)$ относительно подвектора $Y(I_i^o)$, и, следовательно, также матрицей частных коэффициентов корреляции исходного подвектора $X(I_i)$ относительно мешающего подвектора $X(I_i^o)$.

На основании автокорреляционной матрицы $\hat{R}_{ii}$ $(i=1,\ldots,k)$ возможно вычислить главные компоненты. Для этого найдем матрицу Λ_i ее собственных значений

$$\Lambda_i = \begin{pmatrix} \lambda_1^i & 0 & \ldots & 0 \\ 0 & \lambda_2^i & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_{m_i}^i \end{pmatrix},$$

где $\lambda_s^i \geq \lambda_{s+1}^i$ $(s=1,\ldots,m_i-1)$, а m_i – порядок матрицы $\hat{R}_{ii}$, и матрицу U_i соответствующих нормированных (и ортогональных) собственных векторов,

$$U_i = (U_1^i : U_2^i : \ldots : U_{m_i}^i),$$

где $U_s^{i\prime} = (U_{s1}^i, U_{s2}^i, \ldots, U_{sm_i}^i)$ – собственный вектор матрицы $\hat{R}_{ii}$, соответствующий собственному числу λ_s^i, и

---

[1] Здесь $\mathfrak{I}$ – единичная матрица.

$$U_i' U_i = U_i U_i' = 1.$$

Пусть задано число δ_i, $0 < \delta_i < 1$. Требуется при сокращении числа признаков в подвекторе $X(I_i)$ сохранить долю δ_i информации, передаваемой им. При линейной модели это условие понимается как образование при помощи линейного преобразования $X(I_i)$ нового признак-вектора так, что при линейном прогнозировании исходного вектора по новому вектору, доля прогноза (в смысле отношений суммарных дисперсий векторов) будет не меньше δ_i. Хорошо известно, что таким оптимальным прогнозирующим вектором является комплект из p_i первых главных компонентов исходного признак-вектора, где p_i - минимальное число, удовлетворяющее неравенству

$$\sum_{s=1}^{p_i} \lambda_s^i \geq \delta_i m_i.$$

Определим теперь матрицы $\bar{\Lambda}_i$ и $\bar{U}_i$ соответственно соотношениями:

$$\bar{\Lambda}_i = \begin{pmatrix} \lambda_1^i & 0 & \dots & 0 \\ 0 & \lambda_2^i & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_{p_i}^i \end{pmatrix},$$

$$\bar{U}_i = (U_1^i : U_2^i : \dots : U_{p_i}^i),$$

и по ним искомый новый признак-вектор Z_i:

$$Z_i = F_i \hat{Y}(I_i), \qquad (5)$$

где матрица преобразования F_i выражается следующей формулой:

$$F_i = \bar{\Lambda}_i^{-0,5} \bar{U}_i'. \qquad (6)$$

Обратное соотношение, выражающее исходный признак-вектор $\hat{Y}(I_i)$ по новому, имеет место лишь приблизительно (прогнозируется лишь доля δ_i из дисперсии $\hat{Y}(I_i)$):

$$\hat{Y}(I_i) \approx C_i Z_i, \qquad (7)$$

где

$$C_i = \bar{U}_i \bar{\Lambda}_i^{-0,5} .$$

В результате получается новый признак-вектор Z, $Z' = (Z_1' \vdots Z_2' \vdots \ldots \vdots Z_k')$, имеющий размерность p, $p = \sum_{s=1}^{k} p_s$, которая при целесообразном выборе чисел δ_i (например, при $0.5 \leq \delta_i \leq 0.7$) практически является несколько раз меньше чем m.

Для дальнейшего анализа необходимо вычислить и корреляционную матрицу нового признак-вектора Z. Обозначаем эту матрицу символом P:

$$P = \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{12}' & P_{22} & \cdots & P_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ P_{1k}' & P_{2k}' & \cdots & P_{kk} \end{pmatrix}$$

и вычислим значения отдельных блоков. Для этого заметим, что $EZ_i = 0$ $(i=1,\ldots,k)$, а из формул (4) и (5) получается:

$$EZ_i Z_i' = F_i E\hat{Y}(I_i)\hat{Y}'(I_i)F_i' = F_i \hat{R}_{ii} F_i' = \bar{\Lambda}_i^{-0,5} \bar{U}_i \hat{R}_{ii} \bar{U}_i' \bar{\Lambda}_i^{-0,5} = J.$$

Тогда

$$P_{ij} = EZ_i Z_j' = F_i \hat{R}_{ij} F_j',$$
$$P_{ii} = J .$$

Заметим, что полученная матрица P похожа по структуре на результаты канонического анализа. Имеется все-таки принципиальное различие: в каноническом анализе новые признаки выбираются так, чтобы они максимизировали взаимные корреляции, в данном случае же так, чтобы они максимально описывали соответствующий подвектор. Последний принцип оптимизации не ограничивает число исследуемых групп признаков, которое при каноническом анализе должно всегда равняться двум; кроме того оно имеет и некоторые преимущества при интерпретации.

В целях упрощения интерпретации можно полученному новому признак-вектору применить некоторый метод поворота осей (см. напр. [3], стр. 269-292); в результате получается некоторый новый признак-вектор Z^*, $Z^{*'} = (Z_1^{*'}:Z_2^{*'}:\ldots:Z_k^{*'})$, где

$$Z_i^* = T_i Z_i \quad (i=1,\ldots,k),$$

T_i – ортогональная матрица (матрица поворота), $T_i T_i' = J$.

Признак-вектор Z_i^* прогнозирует исходный признак-вектор столь же хорошо, как и признак-вектор Z_i, но каждый его компонент $Z_{i_s}^*$ $(s=1,\ldots,p_i)$ имеет высокие корреляции с некоторыми компонентами исходного признак-вектора, и практически некоррелирован с остальными. Такая ситуация облегчает содержательную интерпретацию признак-вектора Z_i^*.

В результате поворота T_i получим новое выражение матрице преобразования

$$Z_i^* = T_i' F_i \hat{Y}(I_i) = F_i^* \hat{Y}(I_i), \qquad (8)$$

где

$$F_i^* = T_i' F_i.$$

Вычислим и корреляционную матрицу P^* признак-вектора Z^*:

$$P^* = \begin{pmatrix} P^*_{11} & P^*_{12} & \cdots & P^*_{1k} \\ P^{*\prime}_{21} & P^*_{22} & \cdots & P^*_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ P^{*\prime}_{k1} & P^{*\prime}_{k2} & \cdots & P^*_{kk} \end{pmatrix}.$$

Так как

$$EZ^*_i Z^{*\prime}_i = T_i E Z_i Z_i' T_i' = T_i \Im T_i' = T_i T_i' = \Im,$$

то

$$P^*_{ij} = EZ^*_i Z^{*\prime}_j = P^*_i \hat{R}_{ij} P^{*\prime}_j = T_i' P_{ij} T_j.$$

Таким образом, для прогноза исходных признаков через новые получим приближенную формулу

$$\hat{Y}(I_i) \approx C^*_i Z^*_i,$$

где

$$C^*_i = C_i T_i = \bar{U}_i \bar{\Lambda}_i^{0,5} T_i.$$

Матрица C^*_i является факторной матрицей для признак-вектора $Y(I_i)$, вычисленной по нередуцированной корреляционной матрице.

Заметим, что все вышесказанное сохраняется и для первоначального признак-вектора X, так как R является и его корреляционной матрицей. Однако при выражении новых признаков через первоначальные надо учесть нормирование и центрирование. Таким образом, применяя формулы (1), (2), (3) и (8),

получим:

$$Z_i^* = F_i^*\hat{Y}(I_I) = F_i^*\left[D_i^{-0,5}X(I_i)D_i^{-0,5} - B_iD_{i^o}^{-0,5}X(I_i^o)D_{i^o}^{-0,5} + G_i\right],$$

где

$$G_I = D_i^{-0,5}BX(I_i)D_i^{-0,5} - D_iD_{i^o}^{-0,5}BX(I_i^o)D_{i^o}^{-0,5} .$$

является свободным членом, а

$$B_i = \mathrm{diag}\left\{E[(X(I_i) - BX(I_i))(X(I_i) - BX(I_i))']\right\} .$$

Вышеописанная методика, реализованная в ВЦ ТГУ в виде комплекта FORTRAN-программ для ЭВМ "Минск-32", оправдала себя в практических задачах обработки данных.

Литература

1. Рао С.Р., Линейные статистические методы и их применения. М., 1968.
2. Тийт Э.А., Обобщенное понятие признак-вектора. Труды ВЦ ТГУ, 1975, 32, 3-29.
3. Харлан Г., Современный факторный анализ. М., 1972.

КЛАСТЕР-АНАЛИЗ ПРИ ЗАДАЧЕ ТАКСОНОМИИ

Р.В. Зэремаа

Рассмотрим проблемы, связанные с разбиением эмпирического материала научных исследований. Отметим, что обычно разбиение различных явлений на классы "похожих" называется задачей классификации.

Используем такую терминологию: назовем объектами те реальные элементы, которые предъявлены для классификации, т.е. для разбиения на систему классов. Предположим, что об этих классах мы не имеем никаких априорных сведений. Такая задача часто называется задачей таксономии, а соответствующие классы таксонами.

Пусть исходный материал представлен так, что данные составляют кластер-структуру, т.е. разумно искать в данных довольно однородные группы элементов, которые сравнительно изолированы от членов других групп. Предположим, что такие группы в свою очередь разбиваются на сравнительно однородные, довольно изолированные друг от друга, группы - кластеры. Методы конструирования системы кластеров можно рассматривать как формализацию такого представления, что образование кластера зависит от узнавания всеобщего межобъектного различия

(или подобия). В этой статье кластер-анализ рассматривается как двухэтапный процесс:

к первому этапу относится конструирование матрицы различия между классифицируемыми объектами;

на втором этапе происходит конструирование кластер-системы, которая объединяет объекты при различных уровнях различия.

В настоящее время строго математическая обработка как первого, так и второго этапа находится еще в начальной стадии: почти полностью отсутствуют исследования описания объектов и выбора меры близости между объектами с учетом цели классифицирования; при проведении второго этапа используются методы, которые в конечном результате дают как непересекающиеся, так и пересекающиеся кластеры, - но и этот этап еще требует исследования.

В литературе имеется (особенно начиная с шестидесятых годов) много методов автоматического классифицирования. Но, сравнивая их, мы видим, что использование различных методов при одних и тех же исходных данных дает довольно расходящиеся результаты. Для того, чтобы предпочесть один результат другому, нам необходимы критерии, на основе которых можно определить качество классификации. Несмотря на многочисленность публикаций на тему автоматического классифицирования, мало встречается попыток конструирования математических рамок, в которых можно исследовать качества таких методов.

На качество метода можно вывести две диаметрально противоположные точки зрения.

I. Классифицирование должно быть интенсивнее по сравне-

нию с тем, что было бы естественным на основе первоначального коэффициента различия (отметим, что коэффициент различия определяется как функция, ставящая каждой паре объектов в соответствие вещественное число; общие требования к коэффициенту различия рассмотрим в дальнейшем). Искусственно заостренный анализ должен гарантировать продуцирование небольшого количества ясно выделяющихся классов.

Из этих принципов исходит так называемая Австралийская группа (Ланс, Дейл, Клифорд, Лэмберт, Мэкнотн-Смит; см. работы [7-10, 14-16]), которая вырабатывает своим заказчикам алгоритмы для классифицирования материалов научных исследований. Ею выработана так называемая общая комбинированная стратегия, сущность которой можно представить следующим образом.

Пусть $S_l, S_m, S_q, \ldots$ - группы в какой-то стадии иерархического агломеративного алгоритма. Учитывая, что принцип работы такого алгоритма состоит в последовательном объединении групп элементов, сначала самых близких, а затем друг от друга все более отдаленных, - здесь существенно определить модифицированное расстояние (расстояние не в строгом смысле!), когда его надо вычислить между группой S_l и группой $S_{mq}=S_m \cup S_q$ на основе расстояний $d_{lm}=d(S_l,S_m)$, $d_{lq}=d(S_l,S_q)$, $d_{mq}=d(S_m,S_q)$ и соответственно с числами элементов n_m, n_q, n_l, n_{mq} в группах S_m, S_q, S_l, S_{mq}. Австралийская группа выдвигает общую формулу

$$d_{l(mq)}=d(S_l,S_{mq})=\alpha_1 d_{lm}+\alpha_2 d_{lq}+\beta d_{mq}+\gamma|d_{lm}-d_{lq}|. \qquad (1)$$

Параметры α_1, α_2, β и γ определяют сущность стратегии: выбо-

ром подходящих параметров определяются стратегии, соответствующие нескольким известным методам с названиями:

"Ближайший сосед" $\alpha_1=\alpha_2=1/2,\ \beta=0,\ \gamma=-1/2;$

"Дальний сосед" $\alpha_1=\alpha_2=1/2,\ \beta=0,\ \gamma=1/2;$

"Центроид" $\alpha_1=n_m/n_{mq},\ \alpha_2=n_q/n_{mq},\ \beta=-n_m n_q/n_{mq}^2,\ \gamma=0;$

"Медиана" $\alpha_1=\alpha_2=1/2,\ \beta=-1/4,\ \gamma=0;$

"Групповое среднее" $\alpha_1=n_m/n_{mq},\ \alpha_2=n_q/n_{mq},\ \beta=\gamma=0.$

Можно сказать, что в алгоритме "ближайшего соседа" расстояние между любыми таксонами равно расстоянию между двумя самыми близкими элементами, представляющими свои таксоны. В этом алгоритме два объекта попадают в один таксон, если существует соединяющая их цепочка близких между собой элементов, - часто называют этот алгоритм алгоритмом "одной связи". Также можно сказать, что в алгоритме "дальнего соседа" (или "полной связи") расстояние между двумя таксонами определяется как расстояние между двумя самыми отдаленными друг от друга представителями этих таксонов. А, например, в алгоритме с названием "Групповое среднее" под расстоянием между таксонами понимается среднее расстояний между всевозможными парами представителей этих таксонов.

Австралийская группа выработала стратегию, которая разрешает изменить интенсивность классификации варьирования параметра β. Эта стратегия получается, если наложим параметрам из формулы (1) ограничения

$$\alpha_1 = \alpha_2 = \alpha;\quad 2\alpha+\beta = 1;\quad \beta < 1;\quad \gamma = 0.$$

Эта полученная стратегия называется "гибкой стратегией", причем гибкость состоит в том, что интенсивность классифицирования можно изменить с помощью значений параметра β: при

отрицательном β имеем дело с так называемой расширяющей пространство стратегией, при положительном β - с так называемой суживающей пространство стратегией, при значении ноль - со сохраняющей пространство стратегией. Австралийская группа рекомендует выбрать β из промежутка $(-1,0)$, так как полученные при этом таксоны являются высоко однородными.

Итак, некоторые авторы алгоритмов классифицирования стараются образовать искусственно "явные" кластеры, т.е. если исследователь обнаруживает, что результаты классифицирования материала недостаточно подтверждают заранее (интуитивно) поставленную гипотезу, то, повышая или понижая интенсивность классифицирования, он все-таки может получить такую классификацию, которая подтверждает поставление или закрепление такой гипотезы.

II. В другом подходе к оценке качества классификации целью классифицирования является описание разницы между объектами способом, позволяющим предложить или подтвердить гипотезы о факторах, которые обуславливают дифференциацию объектов, на основе структуры начального коэффициента различия. Классифицирование с такой целью можно называть упрощением данных. Из таких принципов исходит так называемая Кембриджская группа (Сибсон, Жардин, Ван Рийсберген; см. работы [1-6, 11-13]).

Логический подход к проблеме упрощения содержит следующие этапы:

1) точное математическое определение данных и результата упрощения, позволяющее рассматривать методы упрощения как преобразования из одной структуры в другую;

2) определение критериев на такие преобразования (такие критерии могут содержать детализацию операций над данными, при которых представление является инвариантным или ковариантным, спецификацию структуры данных, подлежащих сохранению, и определение условий оптимальности);

3) определение существования, единственности и свойств методов, удовлетворяющих наложенным требованиям;

4) нахождение эффективных алгоритмов.

Предположим, что данные для кластер-анализа представлены в виде коэффициента различия над всеми парами n-элементной совокупности P. Определяем коэффициент различия (КР) как действительную функцию d на множестве $P \times P$ с ограничениями:

КР1: $d(a,b) \geq 0, \quad \forall a,b \in P;$

КР2: $d(a,a) = 0, \quad \forall a \in P;$

КР3: $d(a,b) = d(b,a), \ \forall a,b \in P.$

Это - общие требования к коэффициенту различия. Можно наложить и дополнительные ограничения; назовем из них такие, как требование дефинитивности:

КРД: $d(a,b) = 0 \Longrightarrow a = b, \quad \forall a,b \in P;$

требование метричности:

КРМ: $d(a,b) + d(b,c) \geq d(a,c), \quad \forall a,b,c \in P$

или более сильное требование: требование ультраметричности

КРУ: $\max\{d(a,b),d(b,c)\} \geq d(a,c), \ \forall a,b,c \in P.$

Отметим, что дефинитивные метрические коэффициенты различия являются метриками в множестве P в обычном смысле. Дефинитивные ультраметрические коэффициенты различия являются ультраметриками в множестве P в обычном смысле.

* λ-мерным пространством является множество P с введенной λ-метрикой, которой можно задать при помощи некоторой функции d, ставящей каждой паре элементов из P в соответствие вещественное число; общие требования к метрике таковы:

M1: $d(a,b) \geq 0$, $\forall a,b \in P$;

M2: $d(a,b) = 0 \Leftrightarrow a = b$;

M3: $d(a,b) = d(b,a)$, $\forall a,b \in P$;

M4а: $d(a,c)^{1/\lambda} \leq d(a,b)^{1/\lambda} + d(b,c)^{1/\lambda}$, $\forall a,b,c \in P$, если $0 < \lambda \leq 1$;

M4б: $d(a,c) \leq \max\{d(a,b),d(b,c)\}$, $\forall a,b,c \in P$, если $\lambda = 0$.

1-мерное пространство обычно называется метрическим пространством, 0-мерное пространство - ультраметрическим пространством.

Напомним некоторые определения и свойства.

1. Если $d^{(1)}$ и $d^{(2)}$ две метрики в множестве P, мы говорим, что $d^{(1)}$ доминирует над $d^{(2)}$ (символически $d^{(1)} \geq d^{(2)}$) тогда и только тогда, если $d^{(1)}(a,b) \geq d^{(2)}(a,b)$ для всех $a,b \in P$.

2. Если d является метрикой в P, то для любого λ существует однозначно определенная наибольшая λ-метрика, над которой d доминирует. Эту λ-метрику называем λ-поддоминантом d_λ метрики d.

3. Последовательность точек $s_0,s_1,\ldots,s_n \in P$ называется цепью от точки a до точки b, если $s_0 = a$, $s_n = b$ и все точки s_i $(i=0,1,\ldots,n)$ разные. Если C является цепью от a до b, определяем

$$f_\lambda(C) = \left(\sum_{i=0}^{i=n-1} d(s_i,s_{i+1})^{1/\lambda}\right)^\lambda, \quad 0 < \lambda \leq 1;$$

$$f_\lambda(C) = \max_{i=0,\ldots,n-1} d(s_i,s_{i+1}), \quad \lambda = 0.$$

Тогда $d_\lambda(a,b) = \min\{f_\lambda(C) : C$ является цепью от a до b$\}$.

4. Из вышесказанного следует, что d_λ – непрерывная функция от переменных $d(a,b)$ $(a,b\in P)$ и λ.

5. Пусть $\mathfrak{M}$ множество всех метрик в конечном множестве P. Определяем некоторое множество этих метрик в $\mathfrak{M}$ следующим образом:

$$\Delta_\mu(d^{(1)},d^{(2)}) = \left(\frac{1}{2}\sum_{a,b\in P} |d^{(1)}(a,b)-d^{(2)}(a,b)|^{1/\mu}\right)^\mu, \quad 0<\mu\leq 1;$$

$$\Delta_\mu(d^{(1)},d^{(2)}) = \max_{a,b\in P} |d^{(1)}(a,b)-d^{(2)}(a,b)|, \quad \mu = 0.$$

Каждая Δ_μ является метрикой в $\mathfrak{M}$, которая удовлетворяет дополнительному требованию: инвариантность от перестановок в множестве P. Все Δ_μ $(0 \leq \mu \leq 1)$ топологически эквивалентные. Если множество P имеет n элементов, то множество $\mathfrak{M}$ можно рассматривать как подмножество $1/2n(n-1)$-размерного эвклидова пространства; каждая Δ_μ дает топологию подпространства, так что мы можем говорить о непрерывности с точки зрения метрики Δ_μ.

6. d_λ дает глобальный минимум каждой метрике $\Delta_\mu(d,d')$ как функцию от d' при условии, если $d' \leq d$ и d' является λ-метрикой.

Исследуем теперь результат классифицирования. В начале рассмотрим случай непересекающихся таксонов. Удобным матема-

тическим инструментом для описания подобных разбиений является отношение эквивалентности. Отметим важное свойство отношений эквивалентности: всякое отношение эквивалентности, заданное в множестве P, определяет разбиение этого множества на (непересекающиеся) классы, и, наоборот, всякое (иерархическое) разбиение множества P определяет в нем отношение эквивалентности.

Итак, в иерархических моделях таксоны являются помеченными классами эквивалентности при заданном отношении эквивалентности. Обобщенная модель, в которой разрешается частичное покрытие между таксонами, конструируется, используя подходящим образом ограниченные симметричные и рефлексивные отношения. Прогрессирующее ослабление условия транзитивности увеличивает возможность частичного покрытия в таксонах.

Определим теперь отношение, которое будем использовать при построении более общей модели. Пусть P - конечное множество, $|P| = n$, и пусть R - подмножество множества $P \times P$.

k-отношением в множестве P является бинарное отношение, которое удовлетворяет следующим условиям:

рефлексивность: $\forall a \in P \Rightarrow (a,a) \in R$;

симметричность: $\forall a,b \in P,\ (a,b) \in R \Rightarrow (b,a) \in R$;

k-транзитивность: $[S \subseteq P$ и $|S| = k;\ \{a,b\} \subseteq P - S$ и $S \times (S \cup \{a,b\}) \subseteq R] \Rightarrow (a,b) \in R$.

Нас интересуют значения $0 \leq k \leq n$. Отметим, что 1-отношение и отношение эквивалентности одно и тоже.

Чаще используемым коэффициентом различия является дефинитивный метрический коэффициент различия. Исследуем теперь методы, переводящие дефинитивный метрический коэффициент раз-

личия в систему иерархически расслоенных кластеров (т.е. в последовательность расслоенных разбиений совокупности объектов, при этом с каждым разбиением связан числовой уровень). Такая иерархия с числовыми уровнями называется дендрограммой. Выведение иерархической классификации из дендрограммы может происходить путем идентификации ординарных уровней (представляемых натуральными числами) иерархической классификации с числовыми уровнями дендрограммы.

Определим таксономическую иерархию как упорядоченную тройку (P, J, M), где

P - конечное множество;

J - ненулевое натуральное число;

M - соответствие между натуральными числами j, удовлетворяющими условию $0 \leq j \leq J$, и отношениями эквивалентности в множестве P следующим образом:

$$M(0) = \{(a,a) : a \in P\};$$

$$M(J) = P \times P;$$

$$0 \leq j \leq j' \leq J \Rightarrow M(j) \subseteq M(j').$$

Обозначим множество классов эквивалентности при отношении M(j) в множестве P символом P/M(j). Обычно называют

j-категорией упорядоченную пару $\{(A,j):A \in P/M(j)\}$, $(0 \leq j \leq J)$;

j-таксоном - множество из j-категории, $(0 \leq j \leq J)$;

таксоном - любой j-таксон.

Таксон T выпишем упорядоченной парой, первый член которой представляет распространение таксона и второй член - ординарный уровень; для первого члена используем обозначение ext T, а для второго ord T.

Пусть $a,b \in P$ и найдем такой таксон T при самом низком

уровне, где $a,b \in \mathrm{ext}\ T$; определяем теперь

$$\theta(a,b) = \mathrm{ord}\ T.$$

Легко можно доказать разные свойства, например:

1) если $\mathrm{ord}\ T = J$, то $\mathrm{ext}\ T = P$;

2) если $\mathrm{ord}\ T = 0$, то $\mathrm{ext}\ T$ состоит из одного элемента P;

3) если $a,b,c \in P$, то $\theta(a,c) \leq \max\{\theta(a,b),\theta(b,c)\}$;

и т.д.

Дендрограмму (иерархическую систему кластеров) можно математически характеризовать как ультраметрический коэффициент различия, зафиксировав для каждой пары объектов самый низкий уровень, при котором они находятся в том же таксоне.

Пусть у нас имеется таксономическая иерархия (P, J, M) и функция τ определена над натуральными числами $0 \leq j \leq J$, значениями которой являются неотрицательные действительные числа таким образом, что

$$\tau(0) = 0;$$

$$0 \leq j < j' \leq J \Rightarrow \tau(j) < \tau(j').$$

Возьмем $D_\tau(a,b) = \tau[\theta(a,b)]$. Пользуясь отмеченным свойством 3), можно сказать, что D_τ ультраметрика в P. Так, каждая иерархия определяет семейство ультраметрик, - одну для каждой функции τ.

Наоборот, пусть D некоторая ультраметрика в P и σ функция над D(P,P), имеющая значения в множестве $0 \leq j \leq J$, которая удовлетворяет условиям

$$\sigma(0) = 0;$$

$$r \in D(P,P),\ r > 0 \Rightarrow \sigma(r) > 0;$$

$$r,r' \in D(P,P),\ r < r' \Rightarrow \sigma(r) \leq \sigma(r').$$

Определяем отношение эквивалентности $M(j)$ так:

$(a,b) \in M(j)$ тогда и только тогда, если $\sigma[D(a,b)] \leq j$,

– получаем иерархию.

Итак, методы, генерирующие из коэффициента различия таксономическую иерархию, можно рассматривать как переход из метрики в таксономическую иерархию через дендрограммы. Дендрограмма на конечном множестве имеет взаимно однозначное соответствие с ультраметрикой. Переход с дендрограммы к иерархии зависит от выбора функции σ, которая определяет степень таксономической упорядоченности; в таксономической иерархии мы можем собрать и такие уровни, которые в дендрограмме различаются.

Отметим, что обобщая такое взаимно однозначное соответствие между дендрограммой и ультраметрикой, можно утверждать и взаимно однозначное соответствие между множеством коэффициентов различия $C(P)$ и системами расслоенных кластеров.

Пусть $\sum(P)$ является множеством симметричных отношений в множестве P. Конструирование системы расслоенных кластеров рассматриваем как действие функции

$$c : [0,\infty) \to \sum(P),$$

удовлетворяющей условиям

КЛ1: $0 \leq h \leq h' \Rightarrow c(h) \subseteq c(h')$;

КЛ2: если h достаточно большое, то $c(h) = P \times P$;

КЛ3: для любого h $\exists \delta > 0$, так что $c(h) = c(h + \delta)$.

Если дополнительно функция c удовлетворяет условию

КЛД: $c(0) = \{(a,a) : a \in P\}$,

то получаем дефинитивные кластер-системы, а если

КЛУ: $c(h)$ является отношением эквивалентности для любого h,

то получаем иерархические кластер-системы (дендрограммы).

Взаимно однозначное соответствие между множеством C(P) и системами расслоенных кластеров множества P можно представить следующим образом:

$$[L(d)](h) = \{(a,b) : d(a,b) \leq h\},$$

$$[L^{-1}(c)](a,b) = \inf \{h : (a,b) \in c(h)\},$$

преобразование L индуцирует взаимно однозначное соответствие между множеством C'(P) дефинитивных коэффициентов различия и дефинитивными кластер-системами; между множеством U(P) ультраметрических коэффициентов различия и дендрограммами. L и L^{-1} взаимно обратимые из-за условия КЛЗ. Можно сказать, что рассматривая функцию c, какое-то максимальное подмножество K множества P является кластером, если c(h) - общее отношение в множестве K.

Итак, кластер-системы и коэффициенты различия можно идентифицировать. Но, в общем, не все рассматриваемые коэффициенты различия находятся в этом подмножестве множества C(P), которое соответствует особой требуемой системе кластеров (например, иерархической системе). Те, которые не принадлежат этому подмножеству, надо модифицировать, чтобы получить желаемую систему. Пусть X обозначает подмножество множества C(P), которое соответствует возможным данным задачи исследования. Пусть Y обозначает подмножество C(P), соответствующее искомому типу кластер-системы. Мы можем представить кластер-метод как функцию

$$F : X \to Y.$$

Любое интуитивное требование на кластер-метод является ограничением функции F.

Например, пусть X = C(P) и У = U(P), и поставим такие естественные требования:

Т1: метод должен быть четко определен (при фиксированной совокупности данных он должен дать однозначный результат).

Т2: малым изменениям в данных должны соответствовать малые изменения в результатах (называем это требованием стабильности метода или требованием непрерывности преобразования).

Т3: получаемый результат должен быть в некотором смысле самым лучшим при налаженных требованиях, то есть давать коэффициенту различия минимальное искажение.

* Отметим, что на странице 66 под номером 5 приведенные функции Δ_μ являются мерами искажения нужного типа.

Тут мы можем и уточнить ранее упомянутые понятия сужающей и расширяющей пространство стратегий.

Некоторое преобразование удовлетворяет условию сужения тогда и только тогда, если $F(d) \leq d$. Если мерой искажения взята некоторая метрика Δ_μ и дополнительные условия не налагаются, то, учитывая пункты 2, 5, 6, можно показать, что $F(d) = d_0$ и требования Т1 и Т2 удовлетворены. Также можно показать (см. пункт 4), что преобразование $d \to d_0$ можно рассматривать как непрерывную деформацию d_λ при изменении λ от 1 до 0. Удается показать, что d_0 можно вычислить методом "одной связи".

Некоторое преобразование удовлетворяет условию расширения тогда и только тогда, если $F(d) \geq d$. При более подробном исследовании можно увидеть,

что если не наложены дополнительные условия и оптимизацию проводят на основе некоторого Δ_μ, то требование T1 не удовлетворено. Отметим, что такие преобразования получаются методом "полной связи".

Большинство методов типа "средней связи" не удовлетворяют ни требованию суживания, ни требованию расширения; соответствующие методы не удовлетворяют и требованию T2.

T4: если коэффициент различия уже является ультраметрической, то при преобразовании он должен остаться неизменным, – так как если коэффициент различия уже представляет классификацию желательного вида, то неразумно его менять.

T5: операция преобразования должна коммутировать с умножением коэффициента различия со строго положительным параметром, – такое требование гарантирует независимость преобразования от шкалы.

T6: операция преобразования должна коммутировать с любой перестановкой множества P, – это требование обеспечивает независимость преобразования от любого обозначения объектов.

T7: если в данных некоторые кластеры четко вырисованы, то они должны сохраняться и после применения преобразования на начальный коэффициент различия.

Каждое такое интуитивное требование является ограничением на функцию F. Чтобы придать смысл этим ограничениям, нужно также наложить ограничения на множества X и Y. Приведем здесь как пример математические формы требования, которые находятся более просто по сравнению с остальными, требующими более подробного исследования.

T1а: $\emptyset \neq У \subseteq X$;

T1б: $F : У \to У$.

T5а: $d \in X \Rightarrow \alpha d \in X$; $d \in У \Rightarrow \alpha d \in У$, $\forall \alpha > 0$;

T5б: $F(\alpha d) = \alpha F(d)$, $\forall \alpha > 0$.

T6а: $\alpha \in X \Rightarrow d(\varrho \times \varrho) \in X$; $d \in У \Rightarrow d(\varrho \times \varrho) \in У$,
ϱ является перестановкой множества P;

T6б: $[F(d)](\varrho \times \varrho) = F[d(\varrho \times \varrho)]$,
ϱ является перестановкой множества P.

T7а: $d \in X$, то существует $d' \in У$ так, что $d' \trianglelefteq d$;

T7б: $F(d) \trianglelefteq d$.

Строгое исследование всех перечисленных требований приведено в работах [1, 4, 12]. Там же показывается, что такие естественные требования к кластер-методу определяют метод "одной связи" и это единственный подходящий иерархический метод в модели, где $X = C(P)$ и $У = U(P)$.

На рис. 1 иллюстрируем метод "одной связи", пользуясь графом: при коэффициенте различия h вершины графа представляют объекты; ребра соединяют объекты, различие которых не превышает h. При некотором значении коэффициента как при данном уровне, метод "одной связи" дает кластеры как полные подграфы.

Цепочная тенденция метода "одной связи" широко известна (т.е. при относительно низком уровне соединяют группы, связанные цепями). Разные методы "средней связи" стараются избегать цепочного эффекта, выбирая кластеры в некотором смысле более однородные, чем полученные методом "одной связи". Однако, методы неудовлетворяющие требованию непрерывности, скромно говоря, имеют подозрительное значение.

Рис. 1. Образование дендрограммы из коэффициента различия.

А. Коэффициент различия при пяти объектах.

В. Граф-представление коэффициента различия и образование кластеров.

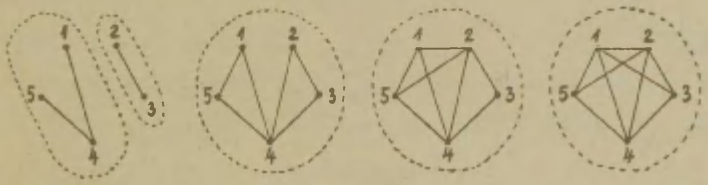
А

	1	2	3	4	5	
		0,3	0,4	0,1	0,2	1
			0,1	0,2	0,3	2
				0,2	0,4	3
					0,1	4
						5

метод "одной связи" →

1	2	3	4	5	
	0,2	0,2	0,1	0,1	1
		0,1	0,2	0,2	2
			0,2	0,2	3
				0,1	4
					5

В



h = 0,1 h = 0,2 h = 0,3 h = 0,4

Так как недостатки метода "одной связи" надо рассматривать как недостатки самого иерархического классифицирования, то лучший путь вновь получить информацию, скрытую от "уцепления" (например, информацию об относительной однородности) - это рассмотреть кластер-методы, выводящие к частично покрывающим (неиерархическим) системам классификации. Интуитивно можно полагать, что увеличивая покрываемость, увеличивается и точность представления данных (однако это за счет повышения сложности представления результата).

Можно обобщить понятия разбиения, определяя k-разбиение, где допускается частичное покрытие максимально $k-1$ объектов между входящими в это разбиение классами. Классифицирующую систему можно рассматривать как последовательность расслоенных k-разбиений; система является иерархической, если $k=1$ и частично покрывающей, если $k>1$. Аналогичным обобщением можем определить понятие k-дендрограммы. В k-дендрограмме при данном уровне кластеры могут частично покрываться до $k-1$ объекта. Метод "одной связи" можно обобщить и указать, что каждый член так определенной последовательности (B_k) кластер-методов удовлетворяет приведенным выше критериям (конечно, обобщая и некоторые требования); первый член последовательности - метод "одной связи", выводящий к иерархической дендрограмме (к 1-дендрограмме); второй член, который можно назвать методом "двойной связи", дает 2-дендрограмму, в которой кластеры могут иметь один общий элемент, и т.д. Если в множестве P n элементов, то B_{n-1} дает точное представление начального коэффициента различия. Можно показать, что семейство мер искажений $\hat{\Delta}_\mu(d,B_k(d))$ монотонно убывает при увеличении k и достигает нуля при $k=n-1$. Подробно исследуют получение таких методов и алгоритм для нахождения $B_k(d)$ в работах [6] и [12].

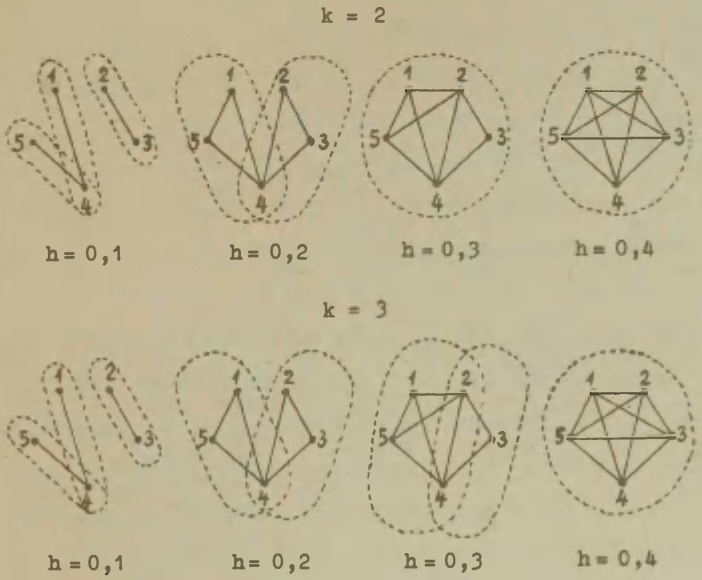
Графическое изображение на рис. 2 дает достаточно хорошее представление о методах построения частично покрывающих систем классификации при выше отмеченных требованиях к методу классифицирования. Применяя метод B_k, получаем кластеры при уровне h следующим образом: вырисовывается граф, ребра которого соединяют объекты с различием не выше h. Найдем

Рис. 2. Образование k-дендрограммы из коэффициента различия.

A. Граф-представление k-дендрограммы.

B. Числовое представление коэффициента различия в результате применения метода B_k.

A



B

k = 2

1	2	3	4	5	
	0,3	0,3	0,1	0,2	1
		0,1	0,2	0,3	2
			0,2	0,3	3
				0,1	4
					5

k = 3

1	2	3	4	5	
	0,3	0,4	0,1	0,2	1
		0,1	0,2	0,3	2
			0,2	0,4	3
				0,1	4
					5

максимальные подграфы (т.е. максимальные подмножества вершин, имеющих все возможные ребра), учитывая, что при применении метода B_k разрешается k-1 общих вершин подграфов.

Отметим, что рассматривая искажение, которое образовано членами последовательности ($B_k(d)$), можем решить, как далеко можно отойти от иерархической классификации. Для выбора уровня в k-дендрограмме, при котором узнаваемые кластеры можно взять за основу классификации, можно использовать меры относительной изоляции и однородности кластеров.

На базе работ Кембриджской группы мы постарались показать строгий подход к выбору метода классифицирования, исходя из некоторых основных предположений, прежде всего из структуры коэффициента различия.

В настоящей статье мы описывали самые распространенные в зарубежной литературе сопоставляемые взгляды двух школ на методы кластер-анализа. Спор о критериях на методы автоматического классифицирования между этими школами можно было бы решить, рассматривая соответствующие цели классифицирования: в одном случае получают высоко однородные кластеры, в другом - упрощенное представление. Если желательно получить и то и другое, то надо использовать неиерархические кластер-методы. При необходимости получить иерархическую кластер-систему следует делать выбор между адекватностью представления и однородностью, бережно анализируя поставленные цели. Если выбирают однородность представления, надо иметь обзор о свойствах используемого метода и о критериях, лежащих в основе конструирования кластеров.

Литература

1. Jardine, C.J., Jardine, N., Sibson, R., The structure and construction of taxonomic hierarchies. Math. Biosciences, 1967, 1, 173-179.
2. Jardine, N., Algorithms, methods and models in the simplication of complex data. The Computer Journal, 1970, 13, 116-117.
3. Jardine, N., A new approach to pattern recognition. Nature, 1971, 234, 526-528.
4. Jardine, N., Sibson, R., A model for taxonomy. Math. Biosciences, 1968, 2, 465-482.
5. Jardine, N., Sibson, R., Choice of methods for automatic classification. The Computer Journal, 1971, 14, 404-406.
6. Jardine, N., Sibson, R., The construction of hierarchic and non-hierarchic classifications. The Computer Journal, 1968, 11, 177-184.
7. Lance, G.N., Williams, W.T., A generalised sorting strategy for computer classifications. Nature, 1966, 212, 218.
8. Lance, G.N., Williams, W.T., A general theory of classificatory sorting strategies. I Hierarchical systems. The Computer Journal, 1967, 9, 373-380.

9. Lance, G.N., Williams, W.T., A general theory of classificatory sorting strategies. II Clustering systems. The Computer Journal, 1967, 10, 271-277.

10. MacNaughton-Smith, P., Some statistical and other numerical techniques for classifying individuals. Home Office Research Unit Report, 6, H.M.S.O., London, 1965.

11. Sibson, R., Some observations on a paper by Lance and Williams. The Computer Journal, 1971, 14, 156-157.

12. Sibson, R., A model for taxonomy II. Math. Biosciences, 1970, 6, 405-430.

13. Van Rijsbergen, C.J., An algorithm for information structuring and retrieval. The Computer Journal, 1971, 14, 407-412.

14. Williams, W.T., Lance, G.N., Dale, M.B., Clifford, H.T., Controversy concerning the criteria for taxonometric strategies. The Computer Journal, 1971, 14, 162-165.

15. Williams, W.T., Dale, M.B., Fundamental problems in numerical taxonomy. Advances in Botanical Research, 1965, 2, 35-68.

16. Williams, W.T., Clifford, H.T., Lance, G.N., Group-size dependence: a rationale for choice between numerical classifications. The Computer Journal, 1971, 14, 157-162.

АЛГОРИТМ ГРУППИРОВКИ ДЛЯ ОБЪЕКТОВ ЗАДАВАЕМЫХ НОМИНАЛЬНЫМИ ПРИЗНАКАМИ

К.А. Пярна

1. Введение

В разных эмпирических исследованиях часто возникает проблема группирования объектов, которая состоит в разбиении данного множества объектов на гомогенные в определенном смысле группы.

Существующие алгоритмы таксономии ориентированы преимущественно на использование количественных признаков. Случай описания объектов значениями качественных (т.е. номинальных) признаков относительно мало рассмотрен в соответствующей литературе. Традиционной техникой в таком случае является вычисление т.н. расстояний Гэмминга между группируемыми объектами (см. напр. [7]). Это дает возможность и при номинальных признаках использовать алгоритмы группировки, которые исходят из матрицы расстояний между объектами. Однако, такой подход имеет по меньшей мере два недостатка: во-первых, если число объектов довольно велико, то при вычислении матрицы расстояний возникают технические трудности и, во-вторых, замена значений признаков на объектах в качестве исходных дан-

ных с матрицей расстояний между этими объектами в общем-то сопряжена с потерей информации.

Но свободные от этих недостатков алгоритмы, использующие в качестве исходных данных значения самих признаков на объектах (вместо матрицы расстояний), в настоящее время являются неудовлетворительными, или в смысле практической применимости их, или из-за недостаточной теоретической обоснованности. Так, на основе "расстояния" между двумя разбиениями объектов [3, 9], в статье [5] приведено определение "оптимального" разбиения, а проблема о практическом нахождении такого разбиения из множества всевозможных разбиений не рассматривается. Аналогичным примером является статья [6], где при определении "оптимального" разбиения используется информационно-теоретический подход. Примером иного типа является метод "последовательной группировки" [2, 4], который, будучи легкоприменимым в практике, в то же время не может считаться достаточно обоснованным в теоретическом плане.

В настоящей статье делается попытка усовершенствовать информационно-теоретический подход, рассматриваемый ранее в работе [6], к задаче группировки в случае номинальных признаков и приводится метод для практической реализации этого подхода. Сначала мы определяем "расстояние" между двумя разбиениями, а через него - некоторый функционал на множестве разбиений. "Оптимальным" считаем такое разбиение, на котором этот функционал достигает минимального значения. Затем предлагаем алгоритм для группировки объектов, который позволяет найти в определенном смысле лучшее приближение к оптимально-

му разбиению. Приведены некоторые свойства алгоритма и небольшой искусственный пример его применения.

2. Определение расстояния между признаками

Сначала описываем исходные данные задачи. Пусть $\mathcal{O} = \{o_1, o_2, \ldots, o_N\}$ является множеством объектов, которое надо сгруппировать по номинальным признакам $X_1, X_2, \ldots, X_M$. Исходные данные можно представить в виде $M \times N$ матрицы $\| X_i(o_j) \|$, где $X_i(o_j)$ обозначит значение признака X_i на объекте o_j. Таким образом, мы не исходим из матрицы расстояний между объектами, как это обычно делают при задаче группировки.

Особенностью номинальных признаков является то обстоятельство, что не существует никакого другого отношения между отдельными значениями признака кроме отношения эквивалентности. Это существенно ограничивает возможности математической обработки номинальных данных. Одним из более подходящих способов анализа таких данных является информационно-теоретический подход, который мы и используем в дальнейшем.

Пусть $H(X_i)$, $H(X_j)$ обозначат энтропии признаков X_i и X_j, а $H(X_i, X_j)$ их общую энтропию (по формуле Шэннона). Выражение

$$T(X_i, X_j) = H(X_i) + H(X_j) - H(X_i, X_j) \qquad (1)$$

называется, как известно, передаваемой информацией признаков X_i и X_j, так как она показывает количество информации, которую мы дополнительно получаем об одном признаке, если известно значение другого признака (более подробно об этих понятиях см. напр. [1]). Благодаря названному свойству, пере-

даваемая информация (1) может быть рассмотрена как некоторая мера тесноты связи между признаками X_i и X_j. Более подробный анализ показывает, что в качестве такой связи выступает взаимно-однозначное соответствие между значениями признаков X_i и X_j (которое, по соглашению, называем взаимно-однозначным соответствием признаков X_i и X_j).

Расстояние между признаками X_i и X_j определяем следующим образом:

$$d(X_i,X_j) = 1 - \frac{T(X_i,X_j)}{H(X_i,X_j)} . \qquad (2)$$

Как видно из формулы, величина $1 - d(X_i,X_j)$ является мерой тесноты взаимно-однозначного соответствия рассматриваемых признаков, нормированная с помощью $H(X_i,X_j)$. Легко показать, что $0 \leq d(X_i,X_j) \leq 1$, при этом $d(X_i,X_j) = 0$ тогда и только тогда, когда признаки находятся во взаимно-однозначном соответствии, и $d(X_i,X_j) = 1$, если признаки являются статистически независимыми. Кроме того, доказано, что расстояние (2) удовлетворяет все аксиомы метрики [8, 10].

3. Экстремизируемый функционал

В целях более удобного рассмотрения задачи введем понятие группирующего признака. Пусть дана некоторая группировка объектов множества $\mathcal{O}$, т.е. разбиение множества $\mathcal{O}$ на непересекающие подмножества (группы) $\mathcal{O}_l$:

$$\mathcal{O} = \cup \mathcal{O}_l \quad \text{и} \quad \mathcal{O}_l \cap \mathcal{O}_m = \emptyset, \quad \text{если} \quad l \neq m .$$

Группирующим признаком, соответствующим этому разбиению, называем признак s, который определяется равенством $s(o_j) = 1$ если $o_j \in \mathcal{O}_1$. Группирующий признак s будет трактован нами как номинальный признак.

Обозначим через s' некоторый другой группирующий признак на множестве $\mathcal{O}$. Группирующие признаки s и s' называются эквивалентными, если они дают одно и то же разбиение множества $\mathcal{O}$, т.е. существует связь $s(o_{j_1}) = s(o_{j_2}) \Leftrightarrow s'(o_{j_1}) = s'(o_{j_2})$. В противном случае группирующие признаки называются различными. Множество всех различных группирующих признаков на множестве $\mathcal{O}$ обозначается через S. Для каждого признака s из множества S можно вычислить его расстояние от любого исходного признака X_i (i=1,2,...,M). По формуле (2) это расстояние $d(s,X_i)$ равно

$$d(s,X_i) = 1 - \frac{T(s,X_i)}{H(s,X_i)}, \quad i=1,2,\ldots,M \qquad (3)$$

при этом, как указано уже выше, величина $1 - d(s,X_i)$ является мерой тесноты взаимно-однозначного соответствия признаков s и X_i.

Определяем функционал

$$d(s) = \sum_{i=1}^{M} d(s,X_i), \qquad (4)$$

значение которого тем меньше, чем большей является сумма мер взаимно-однозначных соответствий группирующего признака s и исходных признаков X_i (i=1,...,M). Оптимальным группирующим признаком на множестве $\mathcal{O}$ называется такой признак $s^* \in S$, на

котором функционал d(s) достигает своего минимального значения по всему множеству S. По этому определению оптимальный признак оказывается наиболее сильно связанным в смысле взаимно-однозначного соответствия с исходными признаками. Задача группирования объектов теперь переходит в задачу нахождения оптимального группирующего признака из множества S.

Данная задача может быть поставлена и в более общем плане, требуя минимизации функционала

$$d'(s) = \sum_{i=1}^{M} a_i d(s, X_i), \quad a_i > 0 . \qquad (4)$$

В таком случае при группировке можно учитывать и важность (удельный вес) разных признаков. Поскольку дальнейшее рассмотрение задачи не зависит от вида минимизирующего функционала, то пользуемся единым обозначением d(s), под которым можно при желании подразумевать и функционал d'(s).

Учитывая метрику (2) в пространстве признаков, оптимальный группирующий признак s^* имеет наглядную геометрическую трактовку. Рассматривая признаки как точки в пространстве всевозможных разбиений, оптимальный группирующий признак является ближайшей к исходным признакам (точкам) $X_1, X_2, \ldots, X_M$ точкой, который, следовательно, находится в "середине" их. Поэтому уместно называть оптимальный группирующий признак средним признаком исходных признаков $X_1, X_2, \ldots, X_M$.

4. Алгоритм группировки

Следующей задачей после определения оптимального разбиения s^* выступает его практическое нахождение из множества S.

Теоретически эта задача разрешима посредством полного перебора всех элементов множества S и выбором элемента s^*, на котором функционал $d(s)$ достигает минимального значения. Однако, на практике такой подход не применим, так как уже при относительно малом числе объектов (N), множество S различных разбиений этих объектов оказывается столь мощным, что при реализации метода возникают технические трудности. Например, для N = 3 число различных разбиений $\nu = 5$, для N = 6 $\nu = 203$, для N = 9 $\nu = 21181$, а для N = 12 число разбиений равняется уже $\nu = 4217427$, не говоря о значении ν при N≈1000, число которое в практике встречается довольно часто.

Задача не может быть решена и посредством аналитических методов решения оптимизационных задач, поскольку число оптимизируемых переменных (значения группирующего признака на каждом объекте) является слишком большим (равняется числу объектов N).

Учитывая вышесказанное, следует считать целесообразной выработку таких алгоритмов группировки, которые не обязательно должны дать точное оптимальное решение, но в то же время являются практически применимыми. Ниже мы приведем один из таких методов группировки объектов. Получаемое при этом решение может быть отличным от оптимального, но в определенном смысле является наилучшим приближением к нему.

В целях ясности изложения введем следующее обозначение. Символом s_n обозначим некоторый группирующий признак на n-элементном подмножестве $\{o_{j_1}, o_{j_2}, \ldots, o_{j_n}\}$ множества $\mathcal{O}$. Пусть $d(s_n; j_1, j_2, \ldots, j_n)$ обозначает значение функционала $d(s)$ на разбиении s_n данного подмножества. Предлагаемый алгоритм

группировки состоит в вычислении последовательности разбиений $s_2^*, s_3^*, \ldots, s_N^*$, где s_n^* $(n=2,3,\ldots,N)$ является "лучшим" (в смысле значения функционала $d(s)$) разбиением объектов на некотором n-элементном подмножестве множества $\mathcal{O}$.

Сам алгоритм группировки следующий:

1. Из C_2^N возможных пар объектов множества $\mathcal{O}$ выбираем пару $\{o_{\alpha_1}, o_{\alpha_2}\}$ и из всевозможных разбиений s_2 на двухэлементном подмножестве выбираем разбиение s_2^*, при котором

$$d(s_2^*;\alpha_1,\alpha_2) = \min_{\substack{o_k,o_l\in\mathcal{O},\ k\neq l \\ s_2(o_k)=1 \\ s_2(o_l)=1 \text{ или } 2}} d(s_2;k,l) . \qquad (5)$$

Объекты $o_{\alpha_1}, o_{\alpha_2}$ называем сгруппированными (группировка определяется признаком s_2^*), а все другие объекты - негруппированными. Пусть r_2 означает число групп в разбиении s_2^* т.е. $r_2 = \max\limits_{i=1,2} s_2^*(o_{\alpha_i})$.

2. Из множества всех негруппированных объектов выбираем объект o_{α_3}, и в то же время разбиение s_3^*, которые удовлетворяют условию

$$d(s_3^*;\alpha_1,\alpha_2,\alpha_3) = \min_{\substack{o_k\in\mathcal{O}\setminus\{o_{\alpha_1},o_{\alpha_2}\} \\ s_3(o_{\alpha_j})=s_2^*(o_{\alpha_j}),\ j=1,2 \\ 1\le s_3(o_k)\le r_2+1}} d(s_3;\alpha_1,\alpha_2,k) . \qquad (6)$$

Объект o_{α_3} называется сгруппированным, а r_3 пусть обозначит число групп в разбиении множества $\{o_{\alpha_1}, o_{\alpha_2}, o_{\alpha_3}\}$ т.е. $r_3 = \max\limits_{j=1,2,3} s_3^*(o_{\alpha_j})$. Очевидно, что $r_2 \le r_3 \le r_2+1$.

3. В общем, при (n-1)-ом шаге из множества всех негруппированных объектов (число их равняется N - n + 1) выбираем объект o_{α_n}, а из всевозможных разбиений s_n выбираем s_n^*, которые удовлетворяют требованию

$$d(s_n^*;\alpha_1,\alpha_2,\ldots,\alpha_n) = \min_{\substack{o_k\in O\setminus\{o_{\alpha_1},\ldots,o_{\alpha_{n-1}}\} \\ s_n(o_{\alpha_j})=s_{n-1}^*(o_{\alpha_j}),\ j=1,\ldots,n-1 \\ 1\le s_n(o_k)\le r_{n-1}+1}} d(s_n;\alpha_1,\alpha_2,\ldots,\alpha_{n-1},k). \quad (7)$$

Объект o_n называется сгруппированным, число групп в разбиении s_n^* обозначается через $r_n = \max\limits_{i=1,2,\ldots,n} s_n^*(o_{\alpha_i})$.

4. Шаг 3 повторяется до тех пор, пока еще есть негруппированные объекты. Общее число шагов будет N - 1. В случае нескольких объектов (или разбиений), удовлетворяющих условию минимума (7), новым группированным объектом рассматривается любой из них, а другие объекты по-прежнему считаются негруппированными. После последнего шага получается разбиение s_N^*, которое и считается решением задачи группировки.

5. Некоторые свойства алгоритма и оценивание результатов

1) Одним из положительных свойств алгоритма является факт, что решение, как правило, не зависит от порядка взятия объектов. Это гарантируется критерием выбора объектов, по которому на каждом шагу из множества всех негруппированных объектов сгруппированным считается тот объект, который минимизирует значение функционала d(s) по всему множеству не-

группированных объектов.

2) Общее число "проб" при использовании этого метода равняется величине

$$2 \cdot C_2^N + \sum_{n=2}^{N-1} (r_{n+1})(N - n) < \frac{N^2}{2} (r_N + 4),$$

где r_n - число групп после группирования n объектов. Видим, что число проб зависит как от числа объектов N, так и от структуры исходных данных (через r_n). Легко показать, что максимальное возможное число проб при N = 12 равняется 396-и, то есть более чем 10^4 раз меньше числа 4217427 (число проб при методе полного перебора всевозможных разбиений). При этом вычислительные процедуры могут быть еще сокращены при использовании на каждом шагу результатов предыдущего шага (например, можно использовать двухмерные распределения группирующего и исходных признаков, найденных на предыдущем шагу).

3) Как вообще при экстремальном подходе, и при этом методе возможно оценить качество решения. Для этого мы используем величину $d(s_N^*)$, т.е. значение функционала на последнем шагу алгоритма.

4) Можно сравнить (по качеству) между собой и решения нескольких задач. В таком случае надо учитывать число M в каждой отдельной задаче. Таким образом, для сравнения результатов различных задач надо вычислить величины $d(s_N^*) / M$.

5) В некоторых случаях исследователя интересует и то, насколько точно на основании группировки можно, "восстановить" на объектах значения исходных признаков $X_1, \ldots, X_M$. Эта точность может быть оценена с помощью арифметического сред-

него (K) отношений передаваемой информации группирующего и исходного признаков энтропии исходного признака.

$$K = \frac{1}{M} \sum_{i=1}^{M} \frac{T(s_N^*, X_i)}{H(X_i)} .$$

При $K = 1$ все исходные признаки являются полно восстанавливаемыми (прогнозируемыми) на основе группировки. Как правило, $0 < K < 1$.

6) Так как при реализации алгоритма максимизируется теснота взаимно-однозначного соответствия между группирующим и исходными признаками, то получаемое в результате группирования число групп r_N должно быть порядка $r_N \sim v_i$ (v_i – число значений признака X_i). Такое значение r_N можно считать "естественным".

7) Добавляем, что описанный выше метод группировки придуман прежде всего для номинальных признаков. Хотя метод применим и для количественных (или порядковых) признаков, однако, в таких случаях он не эффективен, поскольку он не использует всю содержащуюся в таких признаках информацию.

6. Пример

Рассмотрим случай $N = 4$; $M = 3$ и $a_i = 1$ $(i=1,2,3)$. Пусть

$$o_1 = (A,a,\alpha)$$

$$o_2 = (A,a,\beta)$$

$$o_3 = (B,b,\gamma)$$

$$o_4 = (A,b,\gamma) ,$$

где в скобках отмечены значения признаков X_1, X_2 и X_3 (соответственно). Представим ход применения вышеописанного метода группировки.

1. Из множества $C_2^4 = 6$ возможных пар объектов выбираем пару $\{o_1, o_3\}$, которая при $s_2^*(o_1) = 1$ и $s_2^*(o_3) = 2$ минимизирует значение функционала $d(s_2)$ по всем парам и по всевозможным разбиениям двухэлементного множества на группы. (Отметим, что минимальное значение $d(s_2)$ реализируется и на паре $\{o_2, o_3\}$, так что выбирать можно и эту пару.) Выбранные объекты o_1 и o_3 называются сгруппированными. Таким образом, после первого шага получается группировка $\{o_1\}$, $\{o_3\}$.

2. Третьим сгруппированным объектом оказался объект o_2, так как именно o_2 дает (при $s_3^*(o_2) = 1$) минимальное значение функционала $d(s_3)$. После второго шага получается, таким образом, группировка $\{o_1, o_2\}$, $\{o_3\}$.

3. В отношении объекта o_4 остается лишь проверить, входит ли он в первую или во вторую группу, или же составляет отдельную группу. Оказывается, что $d(s_4)$ минимален при $s_4^*(o_4) = 2$.

Следовательно, решением задачи является группировка $\{o_1, o_2\}$, $\{o_3, o_4\}$.

Качество полученной группировки характеризуется значением функционала $d(s_4^*) = 1{,}11$; а в нормированном виде $d(s_4^*)/3 = 0{,}37$, величина которая является средним расстоянием между группирующим и исходными признаками X_1, X_2, X_3.

Значение коэффициента $K = 0{,}69$ показывает, что энтропия исходных признаков описывается группирующим признаком в среднем в объеме 69% (т.е. точность прогнозирования исходных признаков).

Литература

1. Голдман С., Теория информации. М., 1957.
2. Журавель Н.М., Журавель Ф.А., Последовательная группировка на основе качественных признаков. В сб.: Распознавание образов и регрессионный анализ в экономических исследованиях, Новосибирск, 1972.
3. Миркин Б.Г., Черный Л.Б., Измерение близости между различными разбиениями конечного множества объектов. Автоматика и телемеханика, 1970, № 5.
4. Розин Б.Б., Учет влияния качественных признаков при моделировании экономических показателей. В сб.: Вопросы экономико-статистического моделирования и прогнозирования в промышленности, Новосибирск, 1970.
5. Трус Л.С., Черный Л.Б., Распознавание образов в пространстве разбиений. В сб.: Социальная мобильность и проблемы формирования и использования трудовых ресурсов, Новосибирск-Иркутск, 1970.
6. Estabrook, G.F., Some information theoretic optimality criteria for general classification. J.Int.Assoc.Math. Geol., 1971, 3, 203-207.
7. Gower, J.C., A general coefficient of similarity and some its properties. Biometrics, 1971, 27, 857-874.

8. Rajski, C., A metric space of discrete probability distributions. Information and Control, 1961, 4, 371-377.

9. Hand, W.M., Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc., 1971, 66, 846-850.

10. Yasuichi, H., A note on entropic metrics. Information and Control, 1973, 22, 403-404.

Содержание

ТРУДЫ ВЫЧИСЛИТЕЛЬНОГО ЦЕНТРА. Выпуск 36. На русском языке. Тартуский государственный университет, ЭССР, г. Тарту, ул. Юликооли, I8. Сдано в печать 9/07 I976 Бумага офсетная 30х42 1/4. Печ. листов 6,0 (условных 5,58). Учетно-изд. листов 3,82. Тираж 300. МВ 05332. Типография ТГУ, ЭССР, г. Тарту, ул. Пялсони, I4.Зак. № 874. Цена 38 коп.

38 коп.