EESTI NSV TEADUSTE AKADEEMIA TOIMETISED. ÜHISKONNATEADUSED
ИЗВЕСТИЯ АКАДЕМИИ НАУК ЭСТОНСКОЙ ССР. ОБЩЕСТВЕННЫЕ НАУКИ
PROCEEDINGS OF THE ACADEMY OF SCIENCES OF THE ESTONIAN SSR. SOCIAL SCIENCES

1986, 35, 3

https://doi.org/10.3176/hum.soc.sci.1986.3.06

Аза МААМЯГИ

ЭКСПЕРТНАЯ КЛАССИФИКАЦИЯ ОБЪЕКТОВ. РАСПРЕДЕЛЕНИЕ РАССТОЯНИЯ МЕЖДУ УПОРЯДОЧЕННЫМИ РАЗБИЕНИЯМИ

1. Когда «сходство» исследуемого множества объектов трудно оценить с помощью какого-то одного количественного показателя, классификация проводится экспертами обычно без четкой формулировки понятия «сходства», интуитивно.[1] Как правило, из-за различного понимания последнего предложенные различными экспертами разбиения между собой не совпадают. Возникают две проблемы.

Во-первых, не ясно, как по нескольким, весьма противоречивым, быть может, разбиениям найти «истинное».

Во-вторых, при очень большом расхождении мнений экспертов (при полной их рассогласованности, при отсутствии всякой связи между ними) необходимо проверить, не является ли исходная совокупность однородной. Другими словами, не исключено, что «истинного» разбиения не существует и группы, выделенные экспертами, даны наугад, так как если исходное множество не распадается естественным образом на классы, т. е. если оно представляет собой как бы один неделимый класс (или наоборот: все объекты друг на друга одинаково мало похожи, т. е. отсутствует стратификационная природа в множестве), то эксперты, не предпочитая одно разбиение другому, очевидно, предложат любое из возможных разбиений с одинаковой вероятностью. Ясно, что данная схема охватывает более широкий круг причин, чем только что описанные. Равновероятность выбора разбиений (предполагается и независимый друг от друга выбор) может быть обусловлена не только однородностью рассматриваемой совокупности, ее причиной может быть, например, полная некомпетентность привлекаемых к классификации экспертов, непонимание ими поставленной задачи, их полное безразличие к результатам исследования, неинформативность признаков, с помощью которых пытаются оценить истинное разбиение. Перечисленные обстоятельства — не все возможные причины выпадения независимых, равновероятных наблюдений.

Итак, предложив m экспертам разбить n объектов точно на k (не более чем на k) классов, получим m разбиений $\{S_i\}_{i=1}^{m}$, которые могут друг от друга сильно отличаться.[2] Как найти оценку истинного разбиения $\widehat{S} = f(S_1, S_2, \ldots, S_m)$? Как проверить гипотезу, что какое-то

[1] Необходимость выделения из основной рассматриваемой совокупности объектов некоторого подмножества «схожих» между собой объектов возникает при решении многих экономических проблем. Так, при оценке работы предприятий одного профиля целесообразно разделить их на крупные и на мелкие и рассматривать эти группы отдельно.

[2] Разбиения $\{S_i\}_{i=1}^{m}$ могут быть получены и каким-нибудь другим образом (не обязательно экспертным), например при проектировании многомерных точек на m пар компонент (предварительная визуализация многомерных данных), и т. д.

заданное разбиение может быть «истинным», или гипотезу, что исследуемая совокупность на самом деле однородна? Последняя гипотеза должна бы проверяться всегда, с самого начала, до проведения более глубокого статистического исследования совокупности «выборочных значений» исследуемых классификационных переменных.

Ясно, что если рассматривать некоторое расстояние между разбиениями $\{S_i\}_{i=1}^m$ как меру «различия» между ними, то, зная закон распределения его при справедливости проверяемых гипотез о распределении рассматриваемой классификационной переменной, можно проверить эти гипотезы по имеющимся выборочным значениям $\{S_i\}_{i=1}^m$ этой переменной. Поэтому и поставленная задача может быть сведена к нахождению закона распределения подходящего расстояния (случайной величины, заданной на пространстве разбиений) при справедливости изучаемых гипотез о распределении изучаемой классификационной переменной.

Отметим здесь во избежание недоразумений, что в дальнейшем выражения «разбиение», «классификация», «экспериментальная реализация классификационной переменной», «наблюдение» надо понимать как «выборочное значение классификационной переменной».

Гипотезу о равновероятном распределении ранжировок исследовал М. Кендел [1], аналогичную гипотезу в пространстве нечетких толерантностей А. И. Орлов [2, 3], близкую гипотезу в пространстве неупорядоченных разбиений Л. А. Панкова [4], ею были также вычислены, хотя в несколько иной постановке, чем в [5], первые моменты расстояния Хемминга в случае справедливости гипотезы о равновероятном распределении исследуемых классификационных переменных.

Ясно, что гипотезу о равновероятном распределении можно рассмотреть как частный случай некоторой более широкой гипотезы:

$$P(R=S^j)=p_j; \quad j=1, 2, \ldots, N,$$

где R — рассматриваемая классификационная переменная, N — количество разбиений в рассматриваемом пространстве разбиений $\mathfrak{R}$.

Модели вероятностного механизма ошибок, допускаемых экспертами в практических задачах, чаще исследуются на материалах парных сравнений [6]. В пространствах упорядоченных разбиений эти проблемы рассматривались Ю. Н. Тюриным и Я. Пинкавой [7]. В пространстве неупорядоченных разбиений упомянутая проблема мало исследована. Для них не рассматривались модели, аналогичные модели Льюса—Терстоуна для парных сравнений объектов.

Исследование данного вопроса очень трудоемко, поскольку расчленить задачу бывает сложно, а количество возможных разбиений с увеличением числа объектов растет очень быстро. Трудно бывает получить также достаточное количество мнений экспертов, необходимое для обоснованных статистических выводов. По мере увеличения их числа, как правило, падает уровень компетентности экспертов, совокупность мнений становится неоднородной, выявляются различные точки зрения. Обычно предполагается, что вероятностный механизм в задачах классификации таков, что некоторое «наблюдение» имеет тем большую вероятность появления, чем оно ближе к «истинному» [7]. В задачах опроса экспертов, как правило, предполагается существование единственного «истинного» разбиения, в отличие от задач анкетирования в целях изучения общественного мнения, в которых существование нескольких «точек сгущения» довольно естественно [7].

Напрашивается предположение, что если «далекие» от «истинного» разбиения классификации должны иметь меньшую вероятность выпадения, чем «близкие», то вероятность «наблюдения» разбиения должна

быть монотонной функцией принятого в пространстве расстояния. Но какие именно функции встречаются в практических задачах? Если в качестве «меры различия» в рассматриваемом пространстве взять расстояние Хемминга, то, с учетом вышесказанного, вероятность выпадения некоторого разбиения S^j будет описываться некоторой невозрастающей функцией $f(x)$:

$$P(R=S^j)=f(d(S^j, S_0)),$$

где S_0 — некоторое фиксированное («истинное») разбиение, $d(S^j, S_0)$ — расстояние Хемминга между разбиениями S^j и S_0. Какой именно класс функций $f(x)$ здесь надо рассматривать, пока не ясно. В [8] рассматривается экспоненциальная функция, но это не очень сильное ограничение, так как при справедливости исследуемой гипотезы распределение искомых статистик там найдено на основе распределения некоторых других статистик в случае справедливости другой гипотезы (равновероятной). Последний же результат применим не только к экспоненциальной функции, но и к более широкому классу функций, она была взята лишь в качестве примера.

Сформулируем теперь сказанное более четко.

Итак, рассматривается пространство разбиений n различных объектов на k классов $\mathfrak{N}=\{S^j\}_{j=1}^{N}$ и вероятностное пространство $\{\mathfrak{N}, \mathfrak{A}, P\}$.

Нас будет интересовать, как распределены некоторые случайные величины, заданные на $\mathfrak{N}$ в случае, если переменная R распределена согласно гипотезе $H_i (i=0,1)$. В данной статье рассматривается гипотеза H_0 в пространстве упорядоченных разбиений (для произвольного k):

$$P(R=S^1)=P(R=S^2)=\ldots=P(R=S^N)=1/N,$$

а гипотеза H_1:

$$P(R=S^j)=f_\theta(d(S_0, S^j)), \quad j=1, 2, \ldots, N;$$

$$f_\theta(x)=C_{S_0\theta}^{nk}\, e^{-\theta x}$$

для случая двух классов рассматривалась в [8]; здесь S_0 — как и ранее, «истинное» разбиение. Как нетрудно видеть, H_0 есть частный случай H_1 (при $\theta=0$). Знание вышеупомянутых законов распределений позволит проверять эту гипотезу, строить доверительные множества для S_0 (если оно существует), оценивать параметр θ, если он неизвестен. Более подробно исследуемые гипотезы (какие пространства рассматриваются, исследуемое расстояние и т. д.) изложены в [8]. Здесь же мы приведем некоторые утверждения и их доказательства, которые не были опубликованы. Постановки перечисленных выше задач принадлежат С. А. Айвазяну.

II. Итак, пусть пространство $\mathfrak{N}=\{W^j\}_{j=1}^{M_k}$ состоит из разбиений, которые содержат точно k упорядоченных непустых классов ($M_k=k!\,\varphi(n,k)$, $\varphi(n,k)$ — числа Стирлинга II рода). На пространстве $\mathfrak{N}$ задается переменная T, и гипотеза H_0 предполагает равную вероятность появления всех разбиений ($N=M_k$). T и S — две независимые, одинаково распределенные (согласно гипотезе H_0) переменные, а $d(T, S)$ — расстояние Хемминга между ними. Ниже будут вычислены первые точные моменты $d(T, S)$. Поскольку третий момент $d(T, S)$ при справедливости H_0 в данном случае, как и в пространстве неупорядоченных переменных, записывается очень громоздко, то мы ограничимся только первыми двумя моментами, а третий приведем в Приложении.

Утверждение 1.

$$\mathrm{III}^{n1}_{pq} = \sum_{i=1}^{M_p}\sum_{j=1}^{M_q} d(W^j, W^i) = \tag{1}$$

$$= \{\varphi(n,p)\varphi(n,q) - \varphi(n-1,p)\varphi(n-1,q)\}\,\frac{p!q!n(n-1)}{2},$$

$$\mathrm{III}^{n2}_{pq} = \sum_{i=1}^{M_p}\sum_{j=1}^{M_q} d^2(W^i, W^j) = \{\varphi(n,p)\varphi(n,q) -$$

$$-\varphi(n-1,p)\varphi(n-1,q)\}\,\frac{p!q!n(n-1)}{2} + p!q!\{6C^4_n\,[\varphi(n,p)\varphi(n,q) -$$

$$-2\varphi(n-1,p)\varphi(n-1,q) + \varphi(n-2,p)\varphi(n-2,q)] +$$

$$+6C^3_n\,[(\varphi(n,p)\varphi(n,q) - 2\varphi(n-1,p)\varphi(n-1,q) + \varphi(n-2,p)\varphi(n-2,q)) +$$

$$+ (\varphi(n-1,p) - \varphi(n-2,p))\,(\varphi(n-1,q) - \varphi(n-2,q))] +$$

$$+C^2_n\,[\varphi(n,p) - \varphi(n-1,p)]\,[\varphi(n,q) - \varphi(n-1,q)]\}.$$

Доказательство.

$$d(W^i, W^j) = \sum_{s=1}^{n}\sum_{v=1}^{n} |w^i_{sv} - w^j_{sv}|,$$

где $\{w^m_{sv}\}$ — матрица смежности разбиения W^m. Вычисляются III^{n1}_{pq}, III^{n2}_{pq} в основном так же, как соответствующие выражения для неупорядоченных разбиений [5], но, как нетрудно видеть, в данном случае имеется некоторая особенность: матрицы смежности несимметричны, при вычислении III^{nr}_{pq} надо рассматривать отдельно w^i_{sv} и w^i_{vs} и в случае произведения $|w^i_{s_1v_1} - w^j_{s_1v_1}| \times |w^i_{s_2v_2} - w^j_{s_2v_2}|$ подлежат учету также $s_1 = v_2$, $v_1 = s_2$.

Задача, однако, существенно упрощается упорядоченностью классов в разбиениях. Покажем это на примере доказательства равенства (1) (подробно вычислять III^{n2}_{pq}, III^{n3}_{pq} мы не будем, так как эти вычисления принципиально не отличаются от нижеследующих).

$U^n_{pq} = p!q!\{\varphi(n,p)\varphi(n,q) - \varphi(n-1,p)\varphi(n-1,q)\}$ — число возможных пар таких разбиений, в которых некоторые выделенные объекты o_l, o_t находятся в обоих разбиениях не в одном и том же классе (в противном случае, очевидно, $|w^i_{lt} - w^j_{lt}| = 0$, $|w^i_{tl} - w^j_{tl}| = 0$). Возможны два случая: 1) o_l, o_t находятся в одном разбиении в одном классе, в другом — в разных классах (общее количество таких пар разбиений обозначим через U^{n1}_{pq}); 2) o_l, o_t находятся в обоих разбиениях в разных классах (их количество U^{n2}_{pq}; $U^n_{pq} = U^{n1}_{pq} + U^{n2}_{pq}$).

1) В этом случае либо $|w^i_{lt} - w^j_{lt}| = 0$ и $|w^i_{tl} - w^j_{tl}| = 1$, либо $|w^i_{lt} - w^j_{lt}| = 1$ и $|w^i_{tl} - w^j_{tl}| = 0$ (U^{n1}_{pq} входит в сумму с коэффициентом 1).

2) Здесь $|w^i_{lt} - w^j_{lt}|$ и $|w^i_{tl} - w^j_{tl}|$ одновременно либо 0, либо 1, однако объект o_l окажется в классе с более высоким номером ровно столько раз, сколько и объект o_t (U^{n2}_{pq} входит в сумму с коэффициентом $2 \times \frac{1}{2} = 1$).

Количество различных o_l, o_t: $\dfrac{n(n-1)}{2}$.

Утверждение 1 доказано.

Утверждение 2.

$$Ed^r(T,S)=\frac{\mathrm{III}_{kk}^{nr}}{(k!)^2\varphi^2(n,k)}$$

в случае справедливости гипотезы H_0 (*значение* III_{pq}^{n3} *приводится в Приложении*).

Доказательство непосредственно следует из утверждения 1.

Введем

$$Y_{TS}^{nk}=\frac{k^2}{(k-1)\sqrt[4]{n^3}}\,d(T,S)-\frac{\sqrt{n}}{2}\,(k+1).$$

Утверждение 3.

$$\lim E(Y_{TS}^{nk})^r=E\xi^r$$

$$r=1,\ 2,\ 3$$

при справедливости гипотезы H_0, *где* ξ — *случайная величина, имеющая стандартное нормальное распределение.*

Доказательство следует из утверждения 2.

В [8] приведены при $n\to\infty$ без доказательства пределы моментов Y_{TS}^{nk} произвольного порядка. Доказательство громоздко и будет опубликовано в отдельной статье.

Приложение

Значение величины III_{pq}^{n3}

$$Ed^3(T,S)=\frac{\mathrm{III}_{kk}^{n3}}{(k!)^2\varphi^2(n,k)},$$

$$\mathrm{III}_{pq}^{n3}=\sum_{i=1}^{M_p}\sum_{j=1}^{M_q}d^3(W^j,W^i)=$$

$$=p!q!\{[\varphi(n,p)\varphi(n,q)-\varphi(n-1,p)\varphi(n-1,q)]C_n^2+$$

$$+6[\varphi(n,p)\varphi(n,q)-2\varphi(n-1,p)\varphi(n-1,q)+$$

$$+\varphi(n-2,p)\varphi(n-2,q)]C_{\frac{n(n-1)}{2}}^2+6[\varphi(n,p)\varphi(n,q)-$$

$$-3\varphi(n-1,p)\varphi(n-1,q)+2\varphi(n-2,p)\varphi(n-2,q)]C_n^3+$$

$$+6[\varphi(n,p)\varphi(n,q)-3\varphi(n-1,p)\varphi(n-1,q)+3\varphi(n-2,p)\varphi(n-2,q)-$$

$$-\varphi(n-3,p)\varphi(n-3,q)][15C_n^6+30C_n^5+16C_n^4]+$$

$$+3[\varphi(n,p)-\varphi(n-1,p)][\varphi(n,q)-\varphi(n-1,q)][6(C_n^4+C_n^3)+C_n^2]+$$

$$+18[\varphi(n-1,p)-\varphi(n-2,p)][\varphi(n-1,q)-\varphi(n-2,q)]\times$$

$$\times[10C_n^5+11C_n^4+C_n^3]-36[\varphi(n-2,p)-\varphi(n-3,p)]\times$$

$$\times[\varphi(n-2,q)-\varphi(n-3,q)][5C_n^5+C_n^4]\}.$$

ЛИТЕРАТУРА

1. *Кендел М.* Ранговые корреляции. М., 1975.
2. *Орлов А. И.* Случайные множества с независимыми элементами (люсианы) и их применение. — В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. Ученые записки по статистике, **36**. М., 1980, 287—309.
3. *Орлов А. И.* Устойчивость в социально-экономических моделях. М., 1979.
4. *Панкова Л. А.* Разработка формализованных методов обработки экспертной информации в задаче классификации объектов. Автореф. канд. дис. М., 1977.
5. *Маамяги А.* Экспертная классификация объектов. Распределение расстояния между классификациями некомпетентных экспертов. — Изв. АН ЭССР. Обществ. н., 1977, № 2, 122—131.
6. *Шмерлинг Д. С., Дубровский С. А., Аржанова Т. Д., Френкель А. А.* Экспертные оценки. Методы и применение. — В кн.: Статистические методы анализа экспертных оценок. Ученые записки по статистике, **29**. М., 1977, 290—382.
7. Экспертные оценки. Вопросы кибернетики, вып. 58, М., 1979.
8. *Маамяги А. В.* Некоторые задачи статистического анализа классификаций. Таллин, 1982. (Препринт АН ЭССР.)

Представил К. Хабихт

Институт экономики
Академии наук Эстонской ССР

Поступила в редакцию
2/XII 1985

Aasa MAAMÄGI

OBJEKTIDE EKSPERTKLASSIFIKATSIOON. JÄRJESTATUD JAOTISTE VAHELISE KAUGUSE JAOTUSSEADUS

On vaadeldud mõningate ülesannete matemaatilist püstitust, mis on seotud statistiliste hüpoteeside kontrolliga jaotiste ruumis. On leitud Kemeny—Snelli kauguse esimese, teise ja kolmanda momendi väärtus jaotiste võrdtõenäosusliku jaotuse puhul (artiklis on toodud ka tõestus).

Eesti NSV Teaduste Akadeemia
Majanduse Instituut

Toimetusse saabunud
2. XII 1985

Aasa MAAMÄGI

EXPERT CLASSIFICATION OF OBJECTS. THE DISTRIBUTION LAW OF THE DISTANCE BETWEEN ORDERED PARTITIONS

The mathematical formulation of some problems connected with the testing of statistical hypotheses in the space of partitions is treated. The values of the first, second and third moment of Kemeny—Snell's distance in case of an equally probable distribution have been found (the author presents also the proof).

Academy of Sciences of the Estonian SSR,
Institute of Economics

Received
Dec. 2, 1985