АНАЛИЗ РАЗНОТИПНЫХ ДАННЫХ
(Вычислительные системы)
1983 год Выпуск 99

УДК 518.74

СОГЛАСОВАНИЕ РАЗНОТИПНЫХ ШКАЛ

Н.Г.Загоруйко

§ I. Введение

Рассмотрению проблемы оценки "близости", "похожести", "расстояния" между объектами, свойства которых измерены в разных шкалах, посвящены многие работы [I,2]. При этом ищутся такие меры, которые удовлетворяли бы обычным аксиомам (непрерывности, симметричности и т.п.), были бы инвариантны к допустимым преобразованиям в данной шкале и не зависели бы от состава изучаемых объектов. Итоги этих рассмотрений сводятся в основном к тому, что меры, инвариантные к допустимым преобразованиям для многих шкал, можно указать, а мер, которые не зависели бы от состава выборки, не существует. Добавление к конечной выборке A или изъятие из нее какого-нибудь объекта может изменить прежние порядковые номера объектов i и j (для шкал порядка) или нормировку (для более сильных шкал), что приводит к изменению d_{ij} расстояния между i-м и j -м объектами (см. [I]).

Какой же вывод нужно сделать из этих результатов? Не следует ли признать, что адекватных мер близости между объектами любой конечной выборки нет, а следовательно, нет и оснований верить результатам решения всех тех задач, в которых существенно используются меры близости или расстояния между объектами, т.е. задач таксономии, распознавания образов, корреляционного, регрессионного анализа и т.п.?

Не будем спешить соглашаться с таким пессимистическим заключением. Вспомним, что d_{ij} меняется не при всяком изменении состава выборки A . Действительно, при нормировке сильных шкал по разности между самым большим и самым малым значением характеристи-

ки x в таблице, т.е. по $(x_{max}-x_{min})$, мера d_{ij} будет сохраняться всегда, пока изменения состава объектов не коснутся x_{max} или x_{min}. Для любых шкал нормированная мера d_{ij} останется неизменной, если в таблице продублировать все объекты любое число раз.

Очевидно, что справедливо и обратное утверждение: если конечная выборка A составлена из объектов, которые правильно отражают закономерности генеральной совокупности G, то меры d_{ij} будут одинаковыми для объектов i и j, независимо от того, рассматриваем ли мы их на фоне выборки A или на фоне генеральной совкупности G. Выводы (т.е. таксономия, решающие правила, региональные уравнения и.т.д.), сделанные на основании мер d_{ij}, найденных по выборке, будут сохраняться и для генеральной совокупности G.

Отсюда мы приходим к существенному заключению: если выборка A представительна, то можно указать меры d_{ij}, которые будут адекватными и по отношению к соответствующим шкалам и по отношению к генеральной совокупности. А если выборка A непредставительна, то никакие гарантии инвариантности d_{ij} к допустимым преобразованиям шкал не имеют смысла - из-за непредставительности A индуктивные выводы для G все равно будут ложными.

Данная работа не ставит своей целью рассмотрение проблемы представительности выборки. Будем исходить из того, что обоснование гипотезы представительности выборки A каким-то путем сделано и нас теперь интересует только вопрос о выборе меры расстояния d_{ij}.

Нам хотелось бы иметь меры d_{ij}, обладающие следующими очевидными свойствами [1,2]:

а) непрерывности: мера $d(x_i,x_j)$ должна быть непрерывной функцией своих аргументов;

б) симметричности: предполагая пространство значений аргументов изотропным [3], потребуем, чтобы выполнялось соотношение $d(x_i,x_j) = d(x_j,x_i)$;

в) нормированности: $d(x_i,x_j)$ должно изменяться в пределах от 0 до 1, причем $d(x_i,x_j) = 0$, если $x_i = x_j$;

г) инвариантности [4]: для преобразования φ, допустимого для данной шкалы, $d(x_i,x_j) = d[\varphi(x_i),\varphi(x_j)]$;

д) свойством треугольника: для любых трех объектов a, b, c мера $d_{ab} \le (d_{ac} + d_{cb})$.

Не для всех задач анализа данных нужны меры, которые удовлетворяли бы всем этим требованиям. Часто достаточно, чтобы сохранялась информация о расстоянии между объектами лишь с точностью

до порядка, так что требование "г" можно было бы ослабить, а "д" снять совсем. Однако мы попытаемся найти более универсальную меру, удовлетворяющую всем требованиям "а"-"д".

§2. Меры близости между объектами

Не будем подробно останавливаться на с и л ь н ы х шкалах (от шкалы интервалов и выше). Для них, как хорошо известно, свойствам "а"-"д" удовлетворяет мера

$$d^{c}_{ij} = \frac{|x_i - x_j|}{x_{max} - x_{min}}.$$

Оговоримся лишь, что для частного случая, когда $x_{max} = x_{min}$, неопределенное отношение $\frac{0}{0}$ принимается за 0.

Рассмотрим шкалу п о р я д к а. Напомним, что при всех допустимых для этой шкалы преобразованиях $\varphi^{П}$ отношения из набора $(<,>,=)$ между двумя числами x_i и x_j должны сохраняться и для чисел $\varphi^{П}(x_i)$ и $\varphi^{П}(x_j)$. Если мы построим матрицу $M^{П}$ размером $m \times m$ (где m - число объектов в выборке А), в которой для каждой пары объектов укажем их отношение в шкале порядка, то эта матрица не изменится при всех преобразованиях $\varphi^{П}$. Этим свойством матрицы и воспользуемся для конструирования меры d_{ij}. Естественно считать, что те два объекта одинаковы, которые имеют одинаковые порядковые отношения со всеми другими объектами из выборки А. И, наоборот, наиболее непохожие объекты будут иметь наибольшие отличия в своих отношениях с другими объектами. Различия (d) в отношениях i-го и j-го объектов к некоторому k-му объекту будем считать по такому правилу:

$$d^{k}_{ij} = 0, \text{ если } \begin{cases} x_i > x_k \text{ и } x_j > x_k; \\ x_i < x_k \text{ и } x_j < x_k; \\ x_i = x_k \text{ и } x_j = x_k; \end{cases}$$

$$d^{k}_{ij} = 1, \text{ если } \begin{cases} x_i > x_k, \text{ а } x_j < x_k; \\ x_i < x_k, \text{ а } x_j > x_k; \end{cases}$$

$$d^{k}_{ij} = 0{,}5, \text{ если } \begin{cases} x_i = x_k, \text{ а } x_j \neq x_k; \\ x_i \neq x_k, \text{ а } x_j = x_k. \end{cases}$$

Суммарное различие между i-м и j-м объектами

$$d^{П}_{ij} = \sum_{k=1}^{m} d^{k}_{ij}.$$

Легко видеть, что если $x_i = x_j$, то $d^{П}_{ij} = 0$, и что для объектов i и j, имеющих максимально разные порядковые позиции, $d^{П}_{ij} = 1$. Очевидно выполнение и других требований к $d^{П}_{ij}$.

Заметим, что существует один канонический способ приписывания чисел упорядоченному множеству объектов, при котором можно получать то же значение $d^{П}_{ij}$, не строя матрицу отношений $M^{П}$. Эти числа (назовем их "нормированными рангами") строятся по такому правилу: первому по порядку объекту приписывается число 1, второму - 2 и так до конца. Если встречается l объектов с одинаковым порядковым номером (так называемые "серии"), то всем этим объектам приписывается номер $x' = \frac{1}{l}\sum_{a=1}^{l}(a+p)$, где p - количество объектов, предшествовавших серии. После такой канонизации $d^{П}_{ij}$ находится по правилу $d^{П}_{ij} = |x'_i - x'_j|$.

В том, что эта мера $d^{П}_{ij}$ тождественно равна мере, найденной ранее по матрице $M^{П}$, легко убедиться на примере, приведенном на рис.1.

	$x_{П}$
1	12
⋮	6,3
i	8
⋮	
j	12
⋮	5
m	123

а)

	$x^{1}_{П}$
1	4,5
⋮	2
i	3
⋮	
j	4,5
⋮	1
m	6

б)

$x_{П}$	12	6 3	8	12	5	123
12	=	<	>	=	>	<
6 3	>	=	>	>	>	<
8	<	<	=	<	>	<
12	=	<	>	=	>	<
5	<	<	<	<	=	<
123	>	>	>	>	>	=

в)

Рис.1. Правила определения расстояния в шкалах порядка: а) протокол в шкале порядка, б) протокол в шкале нормированных рангов, в) инвариантная матрица отношений $M^{П}$.

Перейдем теперь к шкале наименований. Допустимые преобразования φ^{M} для шкал этого типа всегда сохраняют отношения равенства, так что при всех возможных переименованиях в мат-

рице M^H размера $m \times m$ будут сохраняться значения отношений между всеми парами объектов из выборки A в виде символов $(=, \neq)$.

Расстояние d^H_{ij} между i-м и j-м объектами можно найти по матрице M^H, используя следующее правило:

$$d^H_{ij} = \frac{1}{m} \sum_{k=1}^{m} d^k_{ij},$$

где

$$d^k_{ij}=0, \text{ если } \begin{cases} x_i = x_k \text{ и } x_j = x_k; \\ x_i \neq x_k \text{ и } x_j \neq x_k; \end{cases}$$

$$d^k_{ij}=1, \text{если } \begin{cases} x_i = x_k, \text{ а } x_j \neq x_k; \\ x_i \neq x_k, \text{ а } x_j = x_k. \end{cases}$$

Эта мера расстояния в шкале наименований удовлетворяет требованиям "а" - "д".

Как отмечено в [5], мера d^H_{ij} может быть найдена и без построения матрицы M^H, а прямо через числа m_i и m_j, указывающие частость встречаемости объектов, имеющих имена, одинаковые с i-м и j-м объектами, соответственно (см. рис.2):

$$d^H_{\substack{ij \\ (i \neq j)}} = \frac{m_i + m_j}{m}; \qquad d^H_{\substack{ij \\ (i=j)}} = 0.$$

x^H
a
a
a
b
b
c

а)

x^H	a	a	a	b	b	c
a	=	=	=	≠	≠	≠
a	=	=	=	≠	≠	≠
a	=	=	=	≠	≠	≠
b	≠	≠	≠	=	=	≠
b	≠	≠	≠	=	=	≠
c	≠	≠	≠	≠	≠	=

б)

$m_a=3$, $m_b=2$, $m_c=1$; $d^H_{ab}=\frac{5}{6}$,

$d^H_{ac}=\frac{4}{6}$, $d^H_{bc}=\frac{3}{6}$.

Рис.2. Пример определения меры близости в шкале наименований: а) протокол в шкале наименований, б) инвариантная матрица отношений M^H.

Рассмотрим некоторые свойства этой меры. Если признак использует только два имени, тогда $m_i + m_j = m$ и $d^H_{ij} = 1$. Так что для бинарных признаков эта мера однозначно соответствует мере близости, вычисляемой через хэммингово расстояние.

Если признак $x^{н}$ содержит более двух имен, то две большие группы объектов с именами i и j будут считаться более "далекими", чем две группы с малым числом одноименных объектов. Это свойство хорошо согласуется со стремлением минимизировать ошибку, возникающую при перепутывании имен: при равной цене каждой ошибки суммарные потери будут пропорциональны числу перепутываемых объектов, так что две крупные группы одноименных объектов следует различать более уверенно, при таксономии объединять их в общий таксон в последнюю очередь и т.п. Перепутывание мелких групп, ошибочное объединение их в один таксон приведет к меньшим суммарным потерям.

Мерами $d^{с}$, $d^{п}$, $d^{н}$ теперь можно пользоваться в многомерном случае, определяя расстояние $d^{р}$ в пространстве разнотипных признаков по правилам следующего вида (для трех разнотипных признаков):

$$d^{р} = \frac{\alpha^{с} \cdot d^{с} + \alpha^{п} \cdot d^{п} + \alpha^{н} \cdot d^{н}}{\alpha^{с} + \alpha^{п} + \alpha^{н}};$$

$$d^{р} = \sqrt[\beta]{\frac{(\alpha^{с} d^{с})^{\beta} + (\alpha^{п} d^{п})^{\beta} + (\alpha^{н} d^{н})^{\beta}}{(\alpha^{с})^{\beta} + (\alpha^{п})^{\beta} + (\alpha^{н})^{\beta}}};$$

и т.п. Меры такого типа также удовлетворяют требованиям "а"-"д".

Весовые положительные коэффициенты α следовало бы задавать как функцию от априорно известной информативности признака, однако информативность признака зависит от решаемой задачи и становится известной обычно уже после того, как задача решена. В условиях такой неопределенности полагают все $\alpha = 1$. Но в случае разнотипных признаков можно, по-видимому, задавать α пропорционально п о т е н ц и а л ь н о й информативности шкалы. Ведь известно, что один и тот же признак несет больше информации, если его измерять, например, в шкале отношений, чем в шкале порядка, и тем более в шкале наименований. Так что при отсутствии априорных данных об актуальной информативности каждого признака можно было бы считать, что $\alpha^{с} > \alpha^{п} > \alpha^{н}$.

Количественные соотношения потенциальной информативности шкал разного типа еще, к сожалению, не изучены. Так что никаких конкретных рекомендаций (кроме тривиальной $\alpha = 1$) дать пока нельзя.

§3. Мера расстояния между разнотипными признаками

При корреляционном, регрессионном анализе, при обработке групповых экспертных оценок и в других задачах анализа данных нужно измерять расстояния между признаками. В литературе известны разные меры, применяемые для пар однотипных признаков.

Здесь мы попытаемся сконструировать меру, которая была бы применима для пар как однотипных, так и разнотипных признаков и при этом удовлетворяла бы всем вышеприведенным требованиям "а" - "д".

Начнем с однотипных. Если признаки x_1 и x_2 измерены в шкалах, более с и л ь н ы х, чем шкала порядка, то всем требованиям "а"-"д" удовлетворяет мера $d^{cc} = 1 - |r|$, где r - коэффициент линейной корреляции.

Среди многочисленных мер расстояния двух признаков, измеренных в шкале п о р я д к а [6], своей простотой и естественностью отличается мера Кенделла-Кемени [7,8]. Для ее определения нужно перебрать все C_m^2 парных сочетаний из m объектов и для каждой пары (a,b) сравнить порядковое отношение по признаку x_1 и x_2. Если порядковые отношения одинаковы, т.е. если $x_1(a) > x_1(b)$ и $x_2(a) > x_2(b)$ или $x_1(a) < x_1(b)$ и $x_2(a) < x_2(b)$ или $x_1(a) = x_1(b)$ и $x_2(a) = x_2(b)$, то $d_{ab} = 0$. Если отношения порядка на объектах (a,b) по признаку x_1 прямо противоположны порядку по x_2 (т.е. при $x_1(a) > x_1(b)$ и $x_2(a) < x_2(b)$ или $x_1(a) < x_1(b)$ и $x_2(a) > x_2(b)$), то $d_{ab} = 1$. В промежуточном случае, когда по одному признаку имеет место отношение (>) или (<), а по другому (=), считается, что $d_{ab} = 0{,}5$.

Общее расстояние определяется как средняя мера "несогласия" двух признаков на всех парах объектов

$$d^{пп} = \frac{\sum_{a,b=1}^{m} d_{ab}}{C_m^2}.$$

Если упорядочивания на всех объектах совпадают, то $d^{пп} = 0$, если они на всех объектах противоположны, то $d^{пп} = 1$.

Если один признак ("эксперт" x_1) в порядковой шкале все объекты различает, а другой эти объекты считает одинаковыми (т.е. если x_2 выдает "серию" длиной m), то мера $d^{пп} = 0{,}5$, что вполне естественно. Уместно отметить, что рекомендуемая во многих пособиях мера Спирмэна [5,7,8] в этом случае дает $d^{пп} = \infty$.

Мера расстояния между двумя признаками, измеренными в шкале наименований, определяется по правилу, аналогичному предыдущему: перебираются все пары объектов (a, b), и если отношение по признаку x_1 совпадает с отношением по признаку x_2, т.е. если $x_1(a) = x_1(b)$ и $x_2(a) = x_2(b)$ или $x_1(a) \neq x_1(b)$ и $x_2(a) \neq x_2(b)$, то $d_{ab} = 0$. Если же эти отношения по двум признакам не совпадают, т.е. при $x_1(a) = x_1(b)$ и $x_2(a) \neq x_2(b)$ или $x_1(a) \neq x_1(b)$ и $x_2(a) = x_2(b)$, то $d_{ab} = 1$. В итоге

$$d^{нн} = \frac{\sum_{a,b=1}^{m} d_{ab}}{C_m^2}.$$

Величина $d^{нн}$ в точности равна величине хэммингова расстояния между двоичными матрицами смежности, одна из которых построена по признаку x_1, а вторая - по признаку x_2.

Перейдем теперь к разнотипным парам признаков. Оба признака можно сделать однотипными, если один из них "обеднить" до более слабого или второй "усилить" ("оцифровать" [9]) до более сильного.

Сделаем по очереди то и другое, для каждого случая найдем свою меру расстояния. Общую меру расстояния будем определять как средневзвешенную величину этих двух частных расстояний. Рассмотрим пару $x^c, x^п$ (см. пример на рис. 3).

а)

x^c	$x^п$
3,5	12
2,3	6
7,0	8
2,1	1
11,0	50
15,6	153
13,2	50
10	10

ослабление

x^c до $\overset{*}{x}{}^п$

б)

$\overset{*}{x}{}^п$	$x^п$
3	5
2	2
4	3
1	1
6	6,5
8	8
7	6,5
5	4

$$\overset{*}{d}{}^{пп} = \frac{2,5}{28} = 0,0893.$$

Рис.3. Пример ослабления сильной шкалы x^c до шкалы порядка $\overset{*}{x}{}^п$: а) протокол в разнотипных шкалах, б) протокол в шкале порядка .

Ослабление сильной шкалы (признак x^c) до шкалы порядка ($\overset{*}{x}{}^п$) состоит в том, что на числах x^c и $x^п$ учитываются только отношения порядка и находится $\overset{*}{d}{}^{пп}$ по методу, изложенному выше.

Усиление ("оцифровка") шкалы порядка ($x^{П}$) до сильной ($\overset{*}{x}^{c}$) делается так, чтобы: а) значение порядка объектов по признаку $\overset{*}{x}^{c}$ совпадало с тем, который указал признак $x^{П}$, и б) числовые значения признака $\overset{*}{x}^{c}$ были максимально коррелированны со значениями признака x^{c}. Достигается это способом, показанным на рис.4.Объек-

x^{c}	$x^{П}$
3,3	12
2,3	6
7,0	8
2,1	1
11,0	50
15,6	153
13,8	50
10	10

а)

x^{c}	$x^{П}$
2,1	1
7,0	8
3,5	12
2,4	16
10	10
13,2	50
11,0	50
15,6	153

б)

x^{c}	$\overset{*}{x}^{c}$	
2,1	2,1	
7,0	4,2	$\tilde{\overset{*}{x}}^{c}_{1} = 4,3$
3,5	4,3	
2,4	4,4	
10	10	
11,0	12,0	$\tilde{\overset{*}{x}}^{c}_{2} = 12,1$
8,8	12,2	
15,6	15,6	

в)

Рис.4. Пример усиления шкалы порядка $x^{П}$ до сильной шкалы $\overset{*}{x}^{c}$: а) протокол в разнотипных шкалах, б) протокол, упорядоченный по $x^{П}$, в) протокол в сильной шкале, $\overset{*}{d}^{cc} = 0,074$.

ты упорядочиваются по возрастанию значений признака $x^{П}$.Если встречается "серия" (т.е. $l>1$ объектов с одинаковым порядковым номером), то они выписываются в порядке убывания их признака x^{c}. Затем, начиная с первого объекта, по значениям признака x^{c} выделяются "блоки инверсий", т.е. такие последовательности объектов,которые начинаются объектом i и заканчиваются самым далеким от него объектом j, такими, что $x^{c}_{i} \geq x^{c}_{j}$.

Если отношение порядка по $x^{П}$ и x^{c} совпадает, то в каждом блоке инверсии будет по одному объекту и усиление $x^{П}$ до $\overset{*}{x}^{c}$ делается приписыванием i-му объекту значения $\overset{*}{x}^{c}_{i} = x^{c}_{i}$.

Если порядки по $x^{П}$ и x^{c} не совпадают, то некоторые блоки инверсии (S) будут содержать по $t>1$ объектов. Требование "б" для этих объектов будет выполнено, если приписать каждому из них их среднее значение по признаку x^{c}: $\tilde{x}^{c}_{s} = \frac{1}{t}\sum_{k=1}^{t} x^{c}_{k}$.Для удовлетворения требования "а" объекты из S, различающиеся по $x^{П}$, нужно "раздвинуть" на величину Δ, так чтобы они равномерно заняли диапазон от

$x^c_s - \frac{t\cdot\Delta}{2}$ до $x^c_s + \frac{t\cdot\Delta}{2}$, где Δ – разрешающая способность признака x^c (на рис. 4 $\Delta = 0{,}1$). Затем вычислим коэффициент линейной корреляции r и через его модуль |r| найдем величину $d^{cc} = 1-|r|$. Средняя мера близости определяется так:

$$d^{cn} = \frac{\gamma^n \cdot \overset{*}{d}^{nn} + \gamma^c \cdot \overset{*}{d}^{cc}}{\gamma^n + \gamma^c} .$$

Здесь γ^n и γ^c – весовые коэффициенты, показывающие величину нашего доверия к каждой из составляющих величин ($\overset{*}{d}^{nn}$ и $\overset{*}{d}^{cc}$). При одинаковом доверии к ним можно взять $\gamma^n = \gamma^c = 1$, и тогда в нашем примере $d^{cn} = \frac{0{,}0893 + 0{,}074}{2} = 0{,}0816$.

Для пары x^c x^n (см. рис. 5) ослабление с и л ь н о й шкалы x^c до шкалы н а и м е н о в а н и й ($\overset{*}{x}^n$) состоит в том, что

а)

x^c	x^n
3,5	b
2,6	c
3,5	a
8,3	c
1,4	c
-3,7	a

ослабление →

б)

x^c	$\overset{*}{x}^c$
k	k
c	c
k	a
a	c
p	c
n	a

усиление →

в)

$\overset{*}{x}^n$	x^n
3,5	3,5
2,6	4,1
3,5	-0,1
8,3	4,1
1,4	4,1
-3,7	-0,1

Рис.5. Пример усиления и ослабления для шкал наименований x^n и сильных x^c: а) протокол в разных шкалах, б) протокол в шкале наименований, в) протокол в сильной шкале.

всем различным значениям x^c приписываются разные имена. Затем вычисляется мера $\overset{*}{d}^{nn}$ по правилу, описанному для d^{nn}. При усилении ("оцифровке") признака x^n до $\overset{*}{x}^c$ объекту a приписывается значение $\overset{*}{x}^c(a) = x^c(a)$. Если одинаковое имя b имеет "серия" из l объектов, то всем им приписывается

$$\overset{*}{x}^c(b) = \frac{1}{l} \sum_{i=1}^{l} x^c_i(b) .$$

По полученным числовым значениям вычисляется мера $\overset{*}{d}^{cc}$ по правилу d^{cc} (через корреляцию). Средняя мера

$$d^{cH} = \frac{\gamma^{H} \cdot \overset{*}{d}^{HH} + \gamma^{c} \cdot \overset{*}{d}^{cc}}{\gamma^{H} + \gamma^{c}} .$$

Здесь γ^{H} и γ^{c} имеют тот же смысл, что и выше, и при $\gamma^{H} = \gamma^{c} = 1$ в нашем примере $\overset{*}{d}^{HH} = 0{,}33$, $\overset{*}{d}^{cc} = 0{,}34$, $d^{cH} = 0{,}335$.

Наконец, для пары x^{Π} и x^{H} (см. рис.6) обеднение, как и в предыдущем случае, сводится к приписыванию разных имен объектам, имеющим разные порядковые номера. После чего находится $\overset{*}{d}^{HH}$. При усилении x^{H} до $\overset{*}{x}^{\Pi}$ признак x^{Π} канонизируется по нормированных рангов, а затем вместо имен x^{H} объектам ставятся в соответствие порядковые номера так же, как и в предыдущем случае: $\overset{*}{x}^{\Pi}(a) = x^{\Pi}(a)$ для одиночных объектов и $\overset{*}{x}^{\Pi}(b) = \frac{1}{l}\sum_{i=1}^{l} x_i^{\Pi}(b)$ для серии из одинаково поименованных объектов.

а)

x^{Π}	x^{H}
1	a
6	b
6	d
6	b
8	d

ослабление →

б)

$\overset{*}{x}^{H}$	x^{H}
a	a
c	b
c	d
c	b
p	d

усиление →

в)

x^{Π}	$\overset{*}{x}^{\Pi}$
1	1
3	3
3	4
3	3
5	4

Рис.6. Пример усиления и ослабления для шкал наименований x^{H} и порядка x^{Π}: а) протокол в разных шкалах, б) протокол в шкале наименований, в) протокол в шкале порядка .

После нахождения $\overset{*}{d}^{\Pi\Pi}$ (по правилу $d^{\Pi\Pi}$), можно определить

$$d^{\Pi H} = \frac{\gamma^{\Pi} \cdot \overset{*}{d}^{\Pi\Pi} + \gamma^{H} \cdot \overset{*}{d}^{HH}}{\gamma^{\Pi} + \gamma^{H}} .$$

При $\gamma^{\Pi} = \gamma^{H} = 1$ в примере на рис.6 $\overset{*}{d}^{HH} = 0{,}3$, $\overset{*}{d}^{\Pi\Pi} = 0{,}15$, $d^{\Pi H} = 0{,}225$.

Все промежуточные преобразования шкал входят в группу допустимых для своих шкал. Так что все приведенные меры также инвариантны к допустимым преобразованиям оцениваемых признаков. Усиление и ослабление шкал в некотором смысле вносят симметричную помеху (добавление и потерю информации), так что усреднение после этих процедур будет оправданным.

Применение указанных мер расстояния между объектами и между признаками позволяет использовать все то богатство математического обеспечения, которое имеется сейчас для анализа таблиц данных, измеренных в сильных шкалах. При этом нужно в программах добавить семантический блок, указывающий тип шкалы данного признака и заменить блок определения "расстояния" на блок вычисления соответствующей меры из набора, описанного выше.

Литература

1. ВОРОНИН Ю.А. Введение мер сходства и связи для решения геолого-географических задач. -Докл. АН СССР, 1971, т. 199, № 5, с. 1011-1015.

2. ШУСТОРОВИЧ А.М. Об адекватных парных мерах сходства в задачах распознавания образов с разнородными признаками. -В кн.: Вопросы обработки информации при проектировании систем (Вычислительные системы, вып. 69). Новосибирск, 1977, с. 147-152.

3. ЗАГОРУЙКО Н.Г. Таксономия в анизотропном пространстве. -В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып. 76). Новосибирск, 1978, с. 26-34.

4. СУППЕС П., ЗИНЕС Дж. Основы теории измерений.-В кн.: Психологические измерения. -М., Мир, 1976, с. 120.

5. ЧЕРНЫЙ Л.Б. Порождение мер связи между объектами с помощью мер связи между признаками. -В кн.: Проблемы анализа дискретной информации. Новосибирск, 1975, с. 167-174.

6. ВОРОНИН Ю.А., ГРАДОВА Т.А. Регион-комплекс программ для постановки и решения задач районирования. Препринт № 258 ВЦ СО АН СССР, Новосибирск, 1980.

7. КЕНДЕЛ М. Ранговые корреляции.-М.: Статистика,1975.-350 с.

8. КЕМЕНИ Дж., СНЕЛЛ Дж. Кибернетическое моделирование. -М.: Сов. радио, 1972. - 320 с.

9. ЕНЮКОВ И.С., КУЛАКОВА Е.П. Числовые метки для качественных признаков в дискриминантном анализе. -В кн.: Прикладной многомерный статистический анализ. Ученые записки по статистике, т.33. М., 1978, с. 353-358.

Поступила в ред.-изд.отд.
12 декабря 1982 года