Применение анализа соответствий в обработке нечисловой информации

Ю.Н.Клишина

(Москва)

Статья знакомит в общих чертах с малоизвестным в отечественной практике исследований методом обработки нечисловой информации - анализом соответствий и демонстрирует его применение на конкретном социологическом примере.

Ключевые слова: нечисловая информация, метод анализа, анализ соответствий, ценностные ориентации.

Для анализа количественных (числовых) данных, заданных шкалой отношений либо интервальной, у социолога 'имеется богатый арсенал статистических методов. Однако часто ему приходится работать с признаками нечисловой природы, измеренными номинальной или порядковой шкалой. К ним нельзя применить многие классические методы математической статистики, что существенно затрудняет исследования.

Между тем, вне рамок классического подход существует апробированный аппарат, предназначенный для анализа подобной информации. В частности, следует указе на так называемый анализ соответствий. Он широко используется за рубежом начиная с 60-х годов, однако в отечественную практику был внедрен сравнительно недавно. ЦЭМИ АН СССР создан пакет прикладных программ «САНИ», который реализует этот метод наряду с другими приемами обработки нечисловой информации[1] .

Цель статьи популяризировать идеи, предложенные в [1,2,3,4], опираясь на результаты работы с пакетом «САНИ».

Существуют два подхода к анализу соответствий. При первом устанавливается взаимное соответствие града! пары признаков; при втором объекты и категории неколичественных признаков представляются в виде точек на плоскости, что позволяет выделить аномальные наблюдения и возможные группировки, строить гипотезы о взаимосвязях.

Первый подход к анализу соответствий: условный пример [5]

Три преподавателя были аттестованы десятью студентами по шкале «хороший», средний», «плохой» (см. табл.1).

Табл. 1 обобщает результаты опроса, но не позволяв сделать какой-то объективный вывод относительно деятельности преподавателей, поскольку, во-первых, получение распределение мнений нельзя интерпретировать однозначно во-вторых, процедура оценки носит довольно приблизительный характер: каждый студент имеет свои представления о «хорошем», «среднем» и «плохом» преподавателе, поэтому преподаватели, по сути дела, оцениваются по 10 различным,

[1] См. статью С.Ю.Адамова в настоящем номере журнала.

Таблица 1

Распределение оценок преподавателей группой студентов

Порядковый номер	«хороший»	«средний»	«плохой»	Сумма
1	1	3	6	10
2	3	5	2	10
3	6	3	0	9
Сумма	10	11	8	29

более или менее совпадающим, экземплярам предложенной шкалы. Устранить такие эффекты можно разными способами, например, давая каждому варианту шкалы подробный комментарий. Однако представляется более интересным другой путь: обобщая мнения студентов, получить числовые выражения для этих вариантов, а затем на их основе посчитать средний балл каждого преподавателя. Таким образом можно установить соответствие между градациями первого и второго признака: между порядковыми номерами преподавателей и характеристиками «хороший», «средний», «плохой». Другими словами, значения первого и второго признака можно представить в виде точек на числовой прямой и рассмотреть их взаимное расположение.

Предложенный метод позволяет параллельно с основной задачей (подбором каждому преподавателю соответствующей оценки) решать ряд дополнительных, д том числе: получить числовой эталон шкалы оценок Для группы респондентов; определить расстояние между вариантами шкалы, т.е. насколько «хороший» лучше «среднего», а «средний» - «плохого»; на основе вычисленных средних баллов провести сравнение преподавателей, т. е. определить, во сколько раз и на какие величины различаются их рейтинги среди студентов.

Первый подход: математический алгоритм

Продемонстрируем его на нашем примере. Припишем метку X_1 - «хорошему», X_2 - «среднему», X_3 - «плохому ». Тогда средние баллы 1-го, 2-го и 3-го преподавателей соответственно можно выразить

$$Y_1(X) = \frac{X_1 + 3X_2 + 6X_3}{10}, \qquad (1)$$

$$Y_2(X) = \frac{3X_1 + 5X_2 + 2X_3}{10}, \qquad (2)$$

$$Y_3(X) = \frac{6X_1 + 3X_2 + 6X_3}{9}. \qquad (3)$$

Будем искать числовые значения X_1, X_2, X_3 наилучшим образом представляющие взаимосвязь признаков, таким способом, чтобы максимизировать показатель тесноты связи η^2 между двумя признаками [б]. Способ нахождения X_1, X_2, X_3 излагается ниже мелким шрифтом. Представим решение в общем виде. Рассмотрим таблицу сопряженности признаков X и Y, имеющих соответственно n и m градаций,

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{bmatrix}$$

Обозначим $s_{i.}$ - сумму строки i, а $s_{.j}$ - сумму столбца j. Построим по $s_{j.}$ диагональную матрицу

$$D_n = \begin{bmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_n \end{bmatrix}$$

и аналогичную матрицу с элементами s_i.

$$D_n = \begin{bmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_m \end{bmatrix}.$$

Ищем собственные значения $A = F^T D_n^{-1} F D_m^{-1}$ матрицы т.е. корни уравнения

$$\det|A - \lambda I| = 0 \qquad (4)$$

Верно следующее: число собственных значений матрицы равно ее рангу; эти значения можно упорядочить по возрастанию (убыванию). Если *q=rang A, mo* для множества решений ^ уравнения (4) получаем

$$\lambda_1 > \lambda_2 > \dots > \lambda_n.$$

В [1] показано, что искомыми метками являются координаты собственного вектора $X(X_1, X_2, X_3)$ матрицы А, которые находятся из системы линейных уравнений $Ax = \lambda x$. Однако (4) порождает q собственных значений матрицы А и, следовательно, q различных решений задачи. В качестве основного выбираем собственный вектор

$$\lambda_1 = \lambda_{max} = \max_i \lambda_i,\ i = 1, \dots, q.$$

Насколько удачно полученное решение устанавливает соответствие между признаками, можно судить по отношению

$$\lambda_1/S_pA, \qquad (5)$$

где $S_pA = \sum\lambda_i = \sum a_{ij}$ - след матрицы А.

Чем ближе (5) к единице, тем точнее полученное решение. Для нашего примера исходная матрица

$$F = \begin{bmatrix} 1 & 3 & 6 \\ 3 & 5 & 2 \\ 6 & 3 & 0 \end{bmatrix},$$

а построенные на ее основе матрицы имеют следующий вид

$$F^t = \begin{bmatrix} 1 & 3 & 6 \\ 3 & 5 & 3 \\ 6 & 2 & 0 \end{bmatrix}, \quad D_n = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 11 & 0 \\ 0 & 0 & 8 \end{bmatrix}, \quad D_m = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 9 \end{bmatrix},$$

$$D_n^{-1} = \begin{bmatrix} \frac{5}{44} & 0 & 0 \\ 0 & \frac{11}{80} & 0 \\ 0 & 0 & \frac{4}{55} \end{bmatrix}, \quad D_m^{-1} = \begin{bmatrix} \frac{1}{9} & 0 & 0 \\ 0 & \frac{1}{9} & 0 \\ 0 & 0 & \frac{9}{100} \end{bmatrix}.$$

Перемножим их по известной формуле: $A = F^t D_n^{-1} F D_m^{-1}$

Найдем максимальное собственное значение матрицы $A - \lambda I$. Для этого решим уравнение (4), которое в нашем случае является уравнением третьей степени относительно λ..

Его максимальный корень: λ_1 примерно равен 0, 368**.** Для этого значения λ_1 построим собственный вектор. Решая систему линейных уравнений

$$0{,}381\ X_1 + 0{,}357\ X_2 + 0{,}015\ X_3 = 0{,}368\ X_1$$
$$0.357\ X_1 + 0.0496\ X_2 + 0{,}024\ X_3 = 0{,}368\ X_2$$
$$0{,}15\ X_1 + 0.246\ X_2 + 0{,}048\ X_3 = 0{,}868\ X_3,$$

Находим: $X_1 \approx 1{,}0761$; $X_2 \approx 0.092$; $X_3 \approx 1{,}4717$.

Для порядковой шкалы «хороший», «средний», «плохой» получены такие числовые значения: «хороший» приблизительно равен 1,0761; «средний» - 0,092; «плохой» -1,4717. Расстояния между «хорошим» и «средним», «средним» и «плохим» соответственно составляют 0,984 (или 1,0761-0,092) и 1,379 (или 1,4717-0,092) и различаются на 0,395, т.е. предложенная шкала неравномерна. Происходит смещение в сторону положительных оценок. Это требует дополнительного анализа как самой шкалы (возможно, она неполна и нуждается в разукрупнении), так и выборочной совокупности (может быть, опрашиваемые более склонны давать положительные оценки, что искажает объективную картину).

Далее вычисляются по (1)-(3) средние баллы преподавателей. Они соответственно равны: -1,2322, 0,1227, 1,2326. Сравнение полученных преподавателями баллов с числовыми значениями шкалы однозначно говорит о том, что, по мнению опрошенных, 1-й преподаватель является «плохим», 2-й - «средним», а 3-й - «хорошим». Можно сделать также вывод, что самый высокий рейтинг у 3-го преподавателя, причем, сравнивая баллы его и 1-го, можно предположить, что эти преподаватели являются антиподами с точки зрения их профессиональных характеристик.

Выясним, насколько хорошо представленное соответствие отражает реальную ситуацию. Вычислим отношение (5)

$$\frac{\lambda_1}{S_p A} = 0{,}368/0{,}47 \approx 0{,}78\ .$$

Это означает, что в 78 случаях из 100 верно установленное соответствие между вариантами рассматриваемых признаков.

Предложенный способ анализа нечисловой информации позволяет решать ряд социологических задач, связанных с использованием порядковых шкал (например задачи аттестации, из года в год встающие перед социологом, работающим на предприятии), обобщением мнений и др.

Второй подход к анализу соответствии: математический алгоритм

Он применяется при работе с большим и сложным признаковым пространством, в котором соответствие можно установить только выходя за рамки прямой. Этот подход заключается в представлении градаций обоих признаков в виде точек на плоскости. Таким образом, результат представляется не в виде чисел, которые исследователь сравнивает между собой, а графически. Это существенно облегчает восприятие и анализ материала, выдвижение и проверку гипотез.

Как и в предыдущем случае, ищется такое представление признаков, которое наиболее точно отражает их взаимосвязь. Изложим математическую суть метода, сохраняя все введенные выше обозначения.

Градации признака *X* представляются в одномерном пространстве в виде *n* точек с координатами

$$X_1\left(\frac{f_{11}}{s_{\cdot 1}};\ldots;\frac{f_{m1}}{s_{\cdot 1}}\right);\quad \ldots;\quad X_n\left(\frac{f_{1n}}{s_{\cdot n}};\ldots;\frac{f_{mn}}{s_{\cdot n}}\right),$$

а градации признака Y в виде *m* точек n-мерного пространства

$$Y_1\left(\frac{f_{11}}{s_{1\cdot}};\ldots;\frac{f_{1n}}{s_{1\cdot}}\right);\quad \ldots;\quad Y_m\left(\frac{f_{m1}}{s_{m\cdot}};\ldots;\frac{f_{mn}}{s_{m\cdot}}\right),$$

(Относительные частоты используются для нивелирования различий в маргинальных частотах). Цель построений - графическое представление градаций признаков на одной плоскости. Для этого сначала решается задача отображения на плоскости каждого множества точек.

В [7] показано, что точки градаций признаков X и Y представляются на плоскости с осями, являющимися собственными векторами соответственно матриц $A_x = F^T D_n^{-1} F D_m^{-1}$ и $A_y = F D_m^{-1} F^T D_n^{-1}$

Легко показать, что собственные вектора этих матриц ортогональны. В [7] показано, что собственные значения *A-y* и *Ay* совпадают. Если иь и *Vy* собственные вектора матриц *A-y* и *Ay* соответственно, отвечающие fe-му собственному значению, то имеют место соотношения

$$v_k = \frac{1}{\sqrt{\lambda_k}} F u_k D_m^{-1}; \quad u_k = \frac{1}{\sqrt{\lambda_k}} F^T v_k D_n^{-1}.$$

Это дает основание для совмещения представлений точек *X* и Y на одной плоскости. Причем поиск координат точки $Y_i(Y_i(u_1), Y_i(u_2))$ на плоскости (u_1, u_2) сводится к решению задачи предыдущего случая. Для отыскания координат $X_j(X_j(v_1), X_j(v_2)$на плоскости (v_1, v_2) решается симметричная задача. Доказывается, что выбранные главные оси наилучшим образом представляют взаимосвязь между исследуемыми признаками.

Если описывать алгоритм, используя решение предыдущей задачи, то он становится чрезвычайно наглядным:

1) ищем два наибольших собственных значения λ_1, λ_2 матрицы A_x ;

2) вычисляем собственные вектора u_1, u_2, соответствующие этим значениям;

3) для каждого значения признака *X* по найденным собственным векторам определяем два средние балла, которые и есть его координаты на искомой плоскости;

4) координаты значений признака Y находим, повторяя операции 1)-3) для матрицы A_y.

Второй подход: иллюстративный пример

Проводилось исследование по выяснению ценностных ориентации студентов в возрасте от 18 до 22 лет. Опрашиваемым предлагалось для рассмотрения 18 жизненных ценностей (см. табл.2).[2]

Таблица 2

Жизненные ценности, фигурирующие в опросе

№ п/п	Название ценности	Пояснения
1	Активная жизнь	
2	Жизненная мудрость	зрелость суждений и здравый смысл, достигаемый жизненным опытом.
3	Интересная работа	
4	Искусство и красота природы	переживание прекрасного в природе и искусстве
5	Любовь	духовная и физическая близость с любимым человеком
6	Материально обеспеченная жизнь .	отсутствие материальных затруднений
7	Дружба	наличие хороших и верных друзей
8	Мир, обстановка в Стране	общая хорошая обстановка в стране, в обществе, сохранение мира между народами, как условие благополучия каждого
9	Общественное признание	уважение окружающих, товарищей по работе
10	Познание	возможность расширения образования, кругозора
11	Равенство	братство, равные возможности для всех
12	Самостоятельность	независимость в суждениях и оценках
13	Свобода	независимость в поступках и действиях
14	Счастливая семейная жизнь	

[2] Адаптированная методика Рокича [8]. Исследование проводилось кафедрой физвоспитания МГУ под руководством Б.И.Новикова

15	творчество	возможность Творческой деятельности
16	Уверенность в Себе	Свобода от внутренних противоречий, сомнений
17	Удовольствия	жизнь, полная удовольствий, развлечений, приятного проведения времени.
18	Здоровье	физическое и психическое здоровье

Используя метод парных сравнений, для каждого из опрошенных удалось проранжировать предложенные ценности в порядке убывания важности (1 - наиболее важная, 18 -наименее важная). Общие результаты сведены в табл. 3.

Таблица 3

Ранжировка 18 жизненных ценностей респондентами

Порядковый номер ценности	Ранг («место»)																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	∑
1	6	4	5	5	3	6	2	4	2	3	2	2	3	3	1	1	2	0	54
2	5	6	6	3	4	3	5	1	2	2	4	0	2	4	2	2	1	2	54
3	2	4	7	2	2	4	3	4	1	4	4	2	4	3	3	2	3	0	54
4	2	5	8	5	3	7	6	3	3	4	3	1	1	1	0	1	1	0	54
5	2	5	2	7	4	3	2	4	2	3	1	4	4	2	3	2	2	1	53
6	3	1	4	4	8	6	6	6	7	2	1	0	1	3	2	0	0	0	54
7	1	0	2	0	5	1	3	3	2	2	3	7	2	5	4	5	1	8	54
8	1	1	2	2	5	1	5	3	7	3	3	1	0	3	3	6	3	5	54
9	0	0	0	4	2	4	3	3	2	3	5	3	2	2	7	2	8	4	54
10	3	2	1	2	5	4	2	1	2	2	2	7	6	6	3	1	3	2	54
11	0	1	1	1	0	1	0	4	3	1	5	5	5	3	2	6	7	9	54
12	0	0	0	3	4	3	1	1	8	4	4	2	4	2	5	6	2	5	54
13	5	10	1	2	0	1	2	3	2	3	2	1	4	2	3	7	3	3	54
14	3	4	1	5	2	4	3	4	2	3	4	3	5	1	5	2	2	1	54
15	3	5	4	3	2	1	4	4	1	3	5	2	3	9	1	2	0	2	54
16	2	1	7	3	3	2	3	4	1	2	1	6	2	2	2	4	4	5	54
17	15	4	0	2	2	3	2	1	5	3	2	4	0	1	0	1	7	2	54
18	1	1	3	1	0	0	2	1	2	7	3	4	6	2	8	4	4	5	54
∑	54	54	54	54	54	54	54	54	54	54	54	54	54	54	54	54	54	53	

Число 6, стоящее в первой строке и первом столбце, показывает, что шесть человек из опрошенных пятидесяти четырех считают для себя самым важным активную жизнь. «7» в четвертой строке и шестом столбце означает, что для семерых студентов искусство и красота природы занимают шестое место в ряду предложенных ценностей, и т.д.

Таким образом, получена таблица сопряженности признаков - «жизненная ценность» и «место» — имеющих по 18 градаций. Проанализировать ее и сделать обобщающие выводы весьма затруднительно. Однако анализ соответствий позволяет это сделать. Результатом его работы в данном случае является графическое описание таблицы сопряженности, представленное на рисунке.

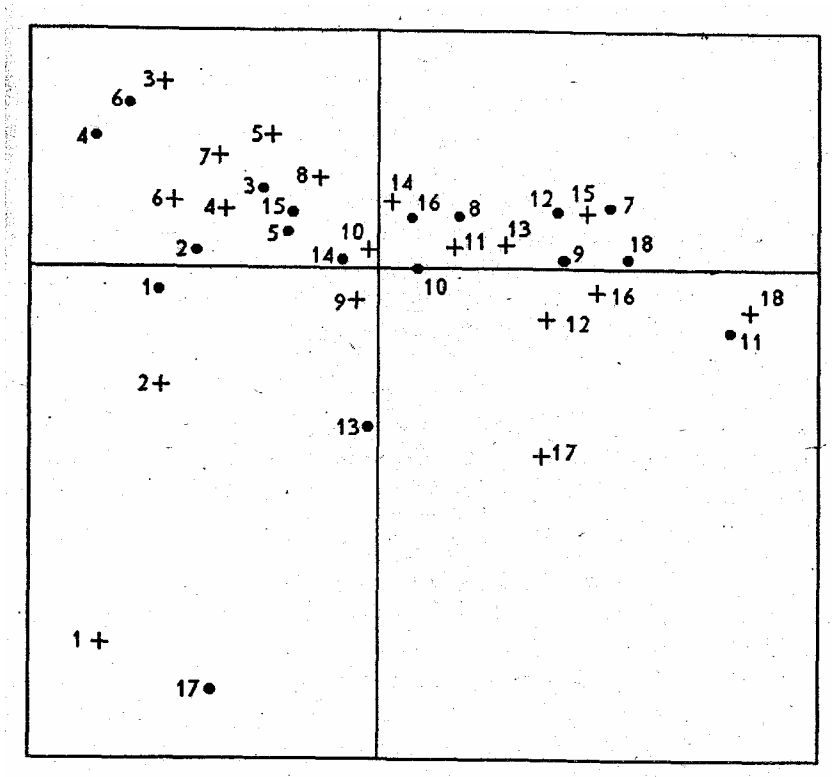
Изучение взаимного расположения точек и крестов дает следующие результаты.

1. На первое место, безусловно, выходит «ценность» ,№ 17 - «удовольствия», а к последнему, восемнадцатому, месту ближе всего находится № 11 - «равенство». Третье место занимает «материально обеспеченная жизнь», а «красота природы и искусство» располагается между третьим и шестым. Перечисление можно продолжать, но читатель легко сможет сделать это самостоятельно. Отметим, что в первой по важности половине оказываются такие ценности, как «удовольствия», «активная жизнь», «жизненная мудрость», «интересная работа», «искусство и красота природы», «любовь», ^материально обеспеченная жизнь», «счастливая семейная жизнь» и «творчество».

2. Когда для какой-то градации признака «ценность» нельзя явно выделить градацию признака «место», следует говорить о том, что мнения о важности данной категории у опрошенных сильно расходятся. Примером может служить .№ 13 («свобода»), которая занимает промежуточную позицию между вторым и семнадцатым местом.

3. В работе [7] показано, что варианты признаков, симметричные относительно оси и находящиеся от нее на достаточно большом расстоянии, можно интерпретировать как противоположные по смыслу. В частности, это относится к №№ 2, 9. Варианты признаков, расположенные рядом (такие как №.№ 3,5,15), можно считать близкими по смыслу.

Рисунок



Графическое описание таблицы сопряженности - табл. **3:**
• - признак «жизненная ценность», + - «место»; градации первого признака обозначены их порядковыми номерами в табл. 2.

Отметим, что анализ соответствий лучше всего использовать для предварительного изучения данных, формирования рабочих гипотез. Особенно удачным может быть его применение при пилотажных исследованиях. В заключение [следует сказать, что данный метод трудно реализуем без средств вычислительной техники, т.к. даже для таблицы

сопряженности размером 2x2 уже требуется большой объем вычислений.

Литература

1. *Адамов СЈО., Енюков И..С.* Методы обработки неколичественной информации, реализованные в пакете программ по прикладному статистическому анализу (ППСА) //Программно-алгоритмическое обеспечение анализа данных в медико-биологичееких исследованиях. М., Пущи-но, 1967.

2. *Адамов СЈО.* Предельные свойства некоторых методов обработки нечисловой информации // III школа-семинар «Программно-алгоритмическое обеспечение прикладного статистического анализа». Тез.докл. Цахкадзор: ЦЭМИ АН СССР, 1987.

3. *Адамов СЈО.* Визуализация неколичествеиных данных //Многомерный статистический анализ и вероятностное моделирование реальных процессов. М.: Наука, 1990.

4. *Адамов С.Ю.* Предельные свойства некоторых методов обработки нечисловой информации //Многомерный статистический анализ и вероятностное моделирование реальных процессов. М.: Наука, 1990.

5. *Nishisato 8.* Analysis of Categorical Data: Dual Scaling and Its Application. Toronto, 1980.

6. Статистические методы анализа информации в социологических исследованиях. М.: Наука, 1979.

7. *Lebart L., Mori.ne.au A., Warwick K.* Multivariate Descriptive Statistical Analysis. N.Y., 1984.

8. *Гоштаутас А.. Семенов А.А.. Ядов ВА.* Адаптированный вариант методики М.Рокича //Саморегуляция и прогнозирование социального поведения личности. Л.: Наука, 1979.