Ранжирование объектов в разнородных шкалах

В. В. Стрижов, В. В. Шакин
Вычислительный Центр Российской Академии Наук
E-mail: strijov@cs.ru, shakin@cs.ru

1 Введение

Ранжированием объектов будем называть отображение $A \xrightarrow{f} Q$, где
$A \in \mathbb{A}$ — описание объектов в пространстве $\mathbb{A}$;
$Q = \{1, 2, 3, ..., m - d\}$ — конечное множество положительных целых чисел, или ранговая шкала; m — количество объектов, d — количество объектов, имеющих один и тот же ранг;
f — функция, выполняющая данное отображение. В дальнейшем мы будем описывать функцию f алгоритмически, и поэтому будем называть ее методом.
Тривиальное решение имеет задача по ранжированию объектов

$$A \xrightarrow{\rho} Q, \tag{1}$$

описываемых одним измеряемым параметром. Например, результаты спортивных соревнований однозначно отображаются в ранговую шкалу:

Таблица 1. Ранжирование однопараметрических объектов

Объекты	A	Q
Заповедники	**Площадь, га**	**Ранг**
Хинганский	93995	2
Хоперский	16178	4
Центрально-Лесной	24447	3
Центральноси-Сибирский	972017	1

При измерении нескольких параметров, произвольный объект описывается с помощью вектора-строки $a_{i\cdot} = \{a_{i1}, a_{i2}, ..., a_{in}\} \in \Re^n$, где элемент $a_{ij} \in \Re^1$ — значение j-го параметра для i-го объекта; $i = \overline{1, m}$, $j = \overline{1, n}$. Множество измерений представляется в виде матрицы A исходных данных

$$A = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \dots & a_{mn} \end{pmatrix},$$

обозначаемой $A = \{a_{i,j}\}_{i,j=1}^{m,n}$; $A \in \Re^{m \times n}$. Вектор-столбцы $a_{\cdot j} \in \Re^m$ матрицы A содержат измерения данного параметра для всех ранжируемых объектов.

Рассмотрим следующие способы представления исходных данных:

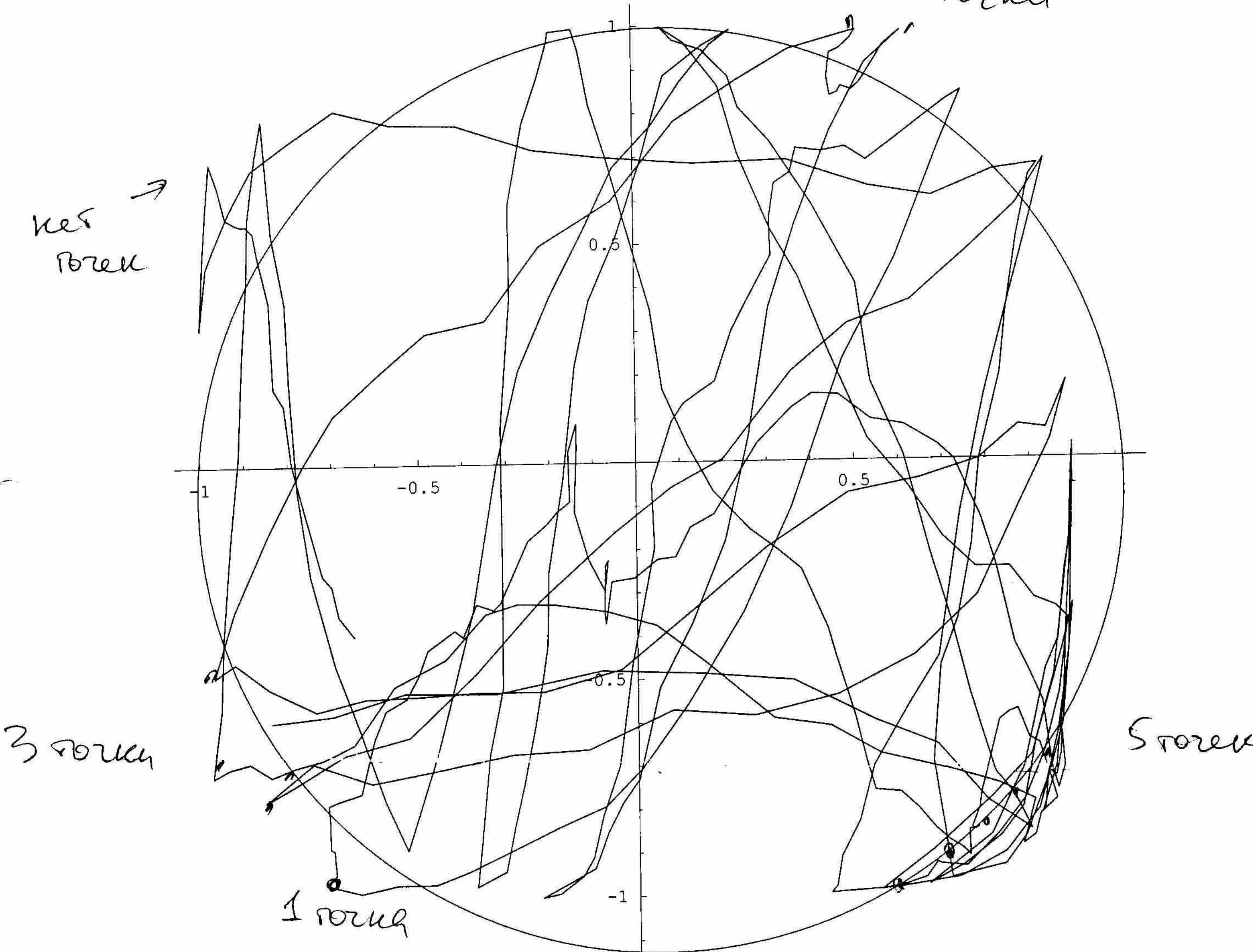
1. Матрица абсолютных значений, или матрица измерений m объектов в n-мерном пространстве параметров $A = \{a_{ij}\}_{i,j=1}^{m,n}$, $a_{ij} \in \Re^1$, $a_{i\cdot}$ — i-й вектор, описывающий произвольный объект в n-мерном пространстве параметров. Как частный случай можно рассмотреть $a_{ij} \in Z^1$ — объект принадлежит пространству порядковых шкал. Эксперты иногда используют упорядоченную шкалу индикаторов. Например, $a_{ij} \in \{0, 1, 2\}$, где 0 — плохо, 1 — удовлетворительно, 2 — хорошо.

2. Матрица попарных расстояний $A = \{a_{ij}\}_{i,j=1}^{m,m}$; $a_{ij} \in \Re^1$; $a_{ij} = a_{ji}$, $a_{ii} = 0$; a_{ij} — расстояние от i-го вектора до j-го вектора.

3. Матрица попарных предпочтений $A = \{a_{ij}\}_{i,j=1}^{m,m}$; $a_{ij} = -a_{ji}$, где $a_{ij} \in \Re^1$ — степень предпочтения объекта a_i объекту a_j. Как частный случай можно рассмотреть значения элемента a_{ij} матрицы A:

$$\begin{cases} a_{ij} = 1, & \text{если объект } a_i \text{ более предпочтителен, чем объект } a_j; \\ a_{ij} = -1, & \text{если объект } a_j \text{ более предпочтителен, чем объект } a_i; \\ a_{ij} = 0, & \text{если предпочтение не определено.} \end{cases}$$

Рис.2



Ссылки на работы, где настраивают электронику.

$$\frac{p}{1} = \frac{\gamma}{2} - 1$$

3 Два экспертных подхода к выбору методов ранжирования

С точки зрения эксперта, применение вышеприведенных методов на практике означает следующее. Возможны две подхода в ранжировании:

1. Наилучшим считается тот объект, который имеет один отличный показатель при прочих удовлетворительных.
2. Наилучшим считается тот объект, который имеет много достаточно хороших показателей.

Так, при существующем порядке подсчета среднего балла для абитуриента, некоторые выпускники имеют неплохой средний балл при наличии плохих оценок по некоторым предметам, компенсируя это отличными оценками по другим предметам. Или, при выборе работы, одни сотрудники предпочитают высокооплачиваемую, но тяжелую работу, в то время как другие предпочитают такую работу, которая имеет удовлетворительные условия труда и оплаты. Выбор одного из подходов остается за экспертом и может диктоваться неформальными причинами, математической моделью или моделью порождения исходных данных. Для первого подхода можно использовать метод суммирования (вычисление Манхэттенского расстояния) или расслоение Парето, для второго подхода можно применять метод вычисления Евклидова расстояния, нахождения главных компонент, или метод сингулярного разложения. Как вариант усугубления двух подходов в ранжировании для первого случая можно возвести значения всех показателей в квадрат или более высокую степень, а для второго случая — найти расстояние Минковского более высокой степени или предложить сингулярное разложение, в котором бы фигурировали собственные значения матрицы данных более высоких степеней.

4 Литература

References

[1] Айвазян С.А, Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика // Классификафия и снижение размерности. — М.: Финансы и статистика, 1989. — С. 334.

[2] Айвазян С.А Интегральные индикаторы качества жизни населения: их построение и использование в социально-экономическом управлении и межрегиональных сопоставлениях. — М.: ЦЭМИ РАН, 2000. — С. 56.

[3] Айвазян С.А., Мхитарян В. С. Прикладная статистика и основы эконометрики — М.: ЮНИТИ, 1998 — С. 111.

[4] Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P. Numerical Recipies in C: The Art of scientific Computing — NY: Cambridge University Press, 1992 — P. 456.

[5] Голуб Дж., Ван-Лоун Ч. Матричные вычисления —М.: Мир, 1999

[6] Шакин В. В. Простые алгоритмы классификации линий // в кн. Опознавание и описание линий — М.: Наука, 1972 — С. 40.

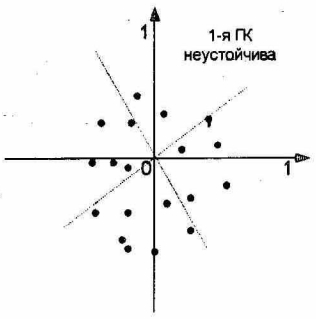


Рис. 4. Измерение первой главной компоненты при малом изменении исходных данных

2.5 Расслоение Парето

Метод расслоения Парето инвариантен к любым преобразованиям исходных данных, сохраняющих порядок значений объектов внутри данного параметра. Это дает возможность опустить предварительную обработку данных. Имеем исходные данные, представленные матрицей абсолютных значений $X = \{x_{ij}\}_{i,j=1}^{m,n}$, где $x_{i\cdot} \in \Re^n$ — вектор, описывающий i-й объект. Вектор $x_{i\cdot}$ называется недоминируемым, если не найдется ни одного вектора $x_{k\cdot}$, такого, что $x_{kj} > x_{ij}$ для всех значений $j = \overline{1,m}$. Для некоторого вектора $x_{i\cdot}$ пространство, в котором он находится, является объединением двух областей (см. рис. 5). В одной области заключены вектора, доминирующие над $x_{i\cdot}$, в другой области заключены вектора не доминирующие над $x_{i\cdot}$. Считается, что сам вектор $x_{i\cdot}$ находится в недоминирующей области.

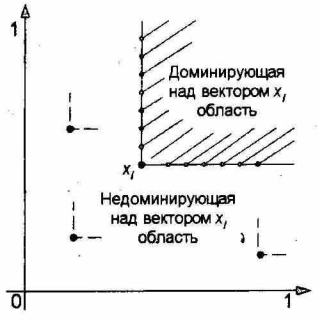


Рис. 5. Доминирующая и недоминирующая области при расслоении Парето

Для нахождения слоев Парето необходимо выполнить следующие шаги.

1. Находим все доминирующие, но недоминируемые вектора (см. рис. 6a).
2. Помечаем все найденные вектора как находящиеся в одном слое (см. рис. 6b).
3. Исключаем эти вектора из дальнейшего рассмотрения и повторяем процедуру до тех пор, пока не останется ни одного вектора (см. рис. 6c).

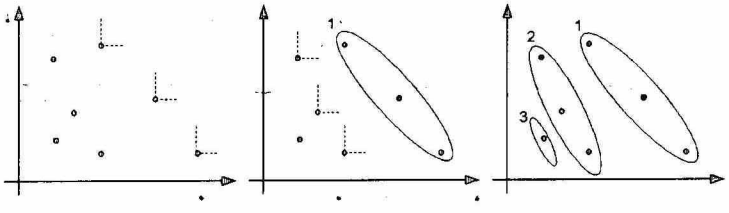


Рис. ??. Нахождение слоев Парето

Существенным недостатком расслоения Парето является то, что при большой размерности пространства параметром или при отрицательной корреляционной зависимости параметров (см. рис. 1b), все объекты оказываются в одном слое.

Упрощенно это можно представить как поворот осей координат таким образом, что проекция векторов на ось абсцисс обладала бы наибольшей дисперсией (см. рис. 3). Из рисунка видно, что проекции векторов на ось z_2 имеют большую дисперсию, чем проекции векторов на ось z_1.

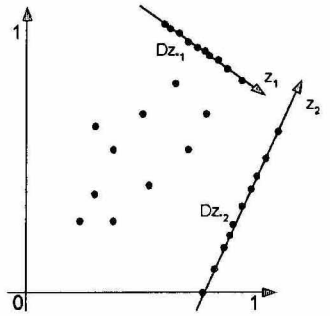


Рис. 3. Дисперсия векторов, проецированных на оси z_1 и z_2

В общих словах, процедура нахождения главных компонент [2] заключается в следующем:
1) Создается ковариационная матрица $\Sigma = \{\sigma_{jk}\}_{j,k=1}^{n,n}$, элементы которой находятся по формуле

$$\sigma_{jk} = \frac{1}{n-1}\Sigma_{i=1}^{n}(x_{ij} - \bar{x}_{\cdot j})(x_{ik} - \bar{x}_{\cdot k}).$$

2) Определяется наибольшее собственное значение λ_1 матрицы Σ, как наибольший по величине корень характеристического уравнения

$$|\Sigma - \lambda I_p| = 0.$$

Алгоритм нахождения собственных значений матриц описан в [4].
3) Решая уравнение

$$(\Sigma - \lambda_1 I)C = 0,$$

находим компоненты c_j собственного вектора $C = (c_1, c_2, \ldots, c_n)^T$ матрицы Σ.
4) Для каждого объекта подсчитываем значение первой главной компоненты

$$z_i = c_1(x_{i1} - \bar{x}_{\cdot 1}) + c_2(x_{i2} - \bar{x}_{\cdot 2}) + \ldots + c_n(x_{in} - \bar{x}_{\cdot n}).$$

В результате вышеприведенной процедуры для каждого объекта мы получаем интегральный показатель $z_i \in R^1$, по которому ранжируем все объекты. Более подробное описание метода главных компонент см. в [1]).

2.4 Ранжирование по сингулярному разложению

Процедура ранжирования по сингулярному разложению заключается в следующем:
1) Находим сингулярное разложение матрицы исходных данных. Известно, что любую матрицу $X \in \Re^{m \times n}$, в которой число строк m больше числа столбцов n можно представить в виде произведения ортогональной матрицы U размерности $m \times n$, диагональной матрицы W, размерности $n \times n$ и транспонированной ортогональной матрицы V, размерности $n \times n$.

$$X = UWV^T$$

Матрица W содержит на своей диагонали убывающие по значению сингулярные числа. Выполняется условие $W = \mathrm{diag}(w_1 \geq w_2 \geq \cdots w_r \geq \cdots w_n \geq 0)$, где индекс r элемента w_r — есть ранг матрицы X
2) Находим проекцию всех векторов X на вектор матрицы U, соответствующий сингулярному числу λ_1

$$Z = U\mathrm{diag}(w_1, 0, \cdots, 0)$$

Как и в предыдущем методе, вектор-столбец $Z = \{z_i\}_{i=1}^{m}$ является интегральным показателем ранга данных объектов.

Недостатком метод главных компонент и метод сингулярного разложения является то, что при n-мерном гауссовом распределеннии , когда исходные данные образуют гиперсферу, мы имеем кратные собственные значения матрицы X, и первая главная компонента становится неустойчивой: при малом изменении X первая главная компонента изменяется значительно (см. рис. 4).

В таблице 3 показано, какие методы ранжирования предполагают нормирование и центрирование матрицы исходных данных.

Таблица 3. Требования к предварительной обработке данных

Метод	Нормирование	Центрирование
Ранжирование с идеальным вектором	+	—
Ранжирование по первой главной компоненте	+	+
Ранжирование по сингулярному разложению	+	+
Расслоение Парето	—	—

2.2 Ранжирование с идеальным вектором

Для ранжирования результатов спортивных соревнований часто применяют бальную шкалу, а сам метод ранжирования спортсменов формулируется так: "По сумме набранных баллов спортсмен a_i занял место q". Здесь каждый объект a_i отображается на вещественную ось функцией $z_i = \sum_{j=1}^{n} x_{ij}$, после чего задача ранжирования сводится к (1). Величина z_i в данном случае называется Манхэттенским расстоянием. Второй частный способ — вычисление Евклидова расстояния от начала координат до объекта, представляя его параметры как элементы вектора $z_i = \sqrt{\sum_{j=1}^{n} x_{ij}^2}$. В общем случае ранжирование с идеальным вектором сводится к нахождению расстояния Минковского для каждого объекта:

$$z_i = \sqrt[k]{\sum_{j=1}^{n} x_{ij}^k}, \tag{2}$$

где значение $k \in Z^1$ определяется выбранным методом ранжирования.

В название метода включены слова "идеальный вектор", так как мы находим расстояние данного объекта либо

1. от наихудшего объекта с параметрами $x_i = \{0, 0, ..., 0\}$,
2. до идеального объекта с параметрами $x_i = \{1, 1, ..., 1\}$.

Во втором случае формула (2) ранжирования будет выглядеть так:

$$z_i = \sqrt[k]{\sum_{j=1}^{n} (1 - x_{ij})^k}. \tag{3}$$

Очевидно, что применение формул (2) и (3), отображающих пространство исходных данных в ранговую шкалу даст различные результаты.

2.3 Ранжирование по первой главной компоненте

Для нахождения первой главной компоненты нормированной и центрированной матрицы исходных данных необходимо найти такие коэффициенты c_j, что линейные комбинации векторов

$$z_{\cdot 1}(X) = c_{\cdot 1} x_{\cdot 1} + c_{\cdot 2} x_{\cdot 2} + ... + c_{\cdot n} x_{\cdot n}$$

обладали бы наибольшей дисперсией, т.е.

$$\frac{\mathrm{D}z_{\cdot 1} + \mathrm{D}z_{\cdot 2} + ... + \mathrm{D}z_{\cdot n}}{\mathrm{D}x_{\cdot 1} + \mathrm{D}x_{\cdot 2} + ... + \mathrm{D}x_{\cdot n}} \longrightarrow \max,$$

при ограничениях нормировки

$$\sum_{i=1}^{m} c_{ij}^2 = 1; j = \overline{1, n};$$

$$\sum_{i=1}^{m} c_{ij} c_{ik} = 0;\ j, k = \overline{1, n};\ j \neq k.$$

Здесь D — знак операции вычисления дисперсии соответсвующей случайной величины

$$\mathrm{D}x_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{\cdot j})^2.$$

Полученные коэффициенты можно представить в виде симметричной таблицы попарных коэффициентов корреляции:

$a_{\cdot j}$	$a_{\cdot 1}$	$a_{\cdot 2}$	$\dots$	$a_{\cdot n}$
$a_{\cdot 1}$	1	r_{21}	$\dots$	r_{n1}
$a_{\cdot 2}$	r_{12}	1	$\dots$	r_{n2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$a_{\cdot n}$	r_{1n}	r_{2n}	$\dots$	1

Если один или несколько коэффициентов корреляции окажутся отрицательными, то это будет рассматриваться как индикатор противоречивости данных. Чем больше отрицательных коэффициентов, тем меньше шансов достичь успеха в ранжировании объектов. Для улучшения результатов можно отказаться от первоначальных параметров и заменить некоторые из них производными параметрами. Например, для ранжирования объектов с параметрами "качество" и "стоимость" при наличии отрицательной корреляционной зависимости имеет смысл ввести новый параметр качество/стоимость, при условии, что оба параметра были измерены с достаточной точностью.

2.1 Нормирование

Многие методы ранжирования объектов в абсолютных шкалах предполагают сопоставимость шкал, или приведение к общей единице измерения. Нормированием матрицы A по столбцам называется отображение $a_{\cdot j} \longrightarrow x_{\cdot j}$, $j = \overline{1,n}$. В большинстве случаев для нормирования применяют линейное отображение: $(a_{i_{min}j}, a_{i_{max}j}) \longrightarrow (0,1)$. Если наибольшему значению объекта в ранговой шкале соответствует наибольшее значение измеряемого параметра, то нормирование выполняется по формуле:

$$x_{ij} = \frac{a_{i,j} - \min(a_{\cdot j})}{\max(a_{\cdot j}) - \min(a_{\cdot j})}.$$

Если наибольшему значению объекта в ранговой шкале соответствует наименьшее значение измеряемого параметра, то нормирование выполняется по формуле:

$$x_{ij} = 1 - \frac{a_{i,j} - \min(a_{\cdot j})}{\max(a_{\cdot j}) - \min(a_{\cdot j})}.$$

Иногда наибольшему значению объекта в ранговой шкале соответствует некоторое оптимальное значение $a_{\cdot j}^{(opt)}$ измеряемого параметра. Такие параметры называются немонотонными. Приведение значений параметра в соответствии с тезисом *tbtb* можно сделать по формуле:

$$x_{ij} = 1 - \frac{a_{i,j} - a_{\cdot j}^{(opt)}}{\max([a_{\cdot j}^{(opt)} - \min(a_{\cdot j})][a_{\cdot j}^{(opt)} - \min(a_{\cdot j})])}.$$

Приведенные выше способы нормирования, где выполняется условие $0 \leq a_{ij} \leq 1$, $a_{ij} \in \Re^1$, называются естественным нормированием [6]. Центрирование векторов — такой параллельный перенос векторов, что среднее арифметическое значение по всем параметрам находится в начале координат (см. рис. 2).

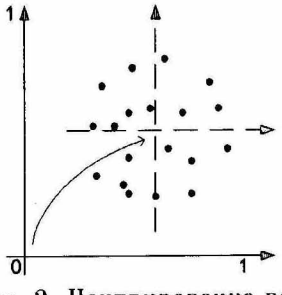


Рис. 2. Центрирование векторов

Значение центрированного вектора находится по формуле:

$$\tilde{x}_{ij} = x_{ij} - \frac{1}{m}\sum_{i=1}^{m} x_{ij}.$$

Для простоты описания методов ранжирования допустим следующие утверждения:
1. Все измерения сопоставимы, то есть приведены к какой-либо единой шкале, которая позволяет обрабатывать сразу всю матрицу исходных данных.
2. Все измерения сделаны с достаточно высокой точностью.
3. Матрица исходных данных не имеет пропущенных значений.

2 Ранжирование в абсолютных шкалах

Приведем пример исходной матрицы данных $A = \{a_{ij}\}_{i,j=1}^{m,n}$ с абсолютными значениями a_{ij} измеряемых параметров:

Таблица 2. Матрица исходных данных в абсолютных шкалах

Объект	Параметр 1	Параметр 2
a_1	0.5	0.2
a_2	0.3	0.4
a_3	1.0	0.1
a_4	0.1	0.3

Легко видеть, что для данной таблицы количество объектов $m = 4$, количество измеряемых параметров $n = 2$, в i-я строка матрицы A соответствует объекту a_i. (вектор-строка), а j-й столбец матрицы соответствует параметру $a_{\cdot j}$. Предполагается, что на каждой шкале измеряемых параметров задано отношение предпочтения: объект с большим значением по данному параметру предпочтительнее объекту с меньшим значением по этому же параметру.

Основная идея методов ранжирования в абсолютных шкалах заключается в том, что наилучшим считается i-й объект с максимальными значениями параметров (обозначим ее "*tbtb*" — the bigger the better). Объект самого высокого ранга имеет параметры: $a_i = \{1, 1, ..., 1\}$. Сильная сторона данной идеи в ее простоте и универсальности. Слабая сторона идеи заключается в том, что она предполагает зависимость между столбцами матрицы A. Приведем пример. Ранжируя объекты, показанные на рис. 1a, эксперт легко находит наилучший и наихудший элементы. Ранжирование объектов, показанных на рис. 1b без предварительной обработки затруднительно, и качество ранжирования будет невысоким. Эксперт, который ориентируется на гипотезу *tbtb*, скажет, что данные, показанные на рис. 1b противоречивы.

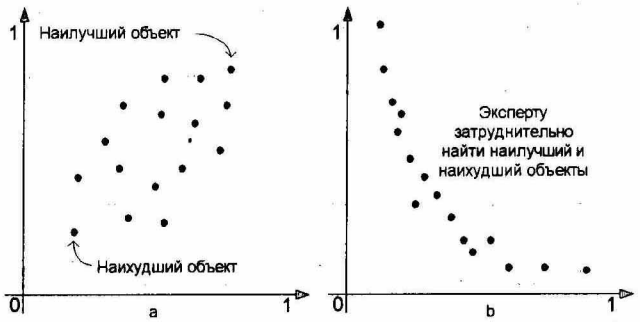


Рис. 1. Влияние зависимости между параметрами на результаты ранжирований

Чтобы оценить качество ранжирования многопараметрических объектов согласно *tbtb*, найдем матрицу попарной корреляции [3] вектор-столбцов $a_{\cdot j}$. Для этого разобьем столбцы матрицы A на пары, обозначая произвольную пару $(a_{\cdot j}, a_{\cdot k})$, и для каждой пары найдем коэффициент корреляции:

$$r_{jk} = \frac{\sum_{i=1}^{m}(a_{ij} - \bar{a}_{\cdot j})(a_{ik} - \bar{a}_{\cdot k})}{\sqrt{\sum_{i=1}^{m}(a_{ij} - \bar{a}_{\cdot j})^2 \sum_{i=1}^{m}(a_{ik} - \bar{a}_{\cdot k})^2}},$$

где $\bar{a}_{\cdot j}$ — среднее арифметическое значение j-го столбца $a_{\cdot j}$:

$$\bar{a}_{\cdot j} = \frac{1}{m}\sum_{i=1}^{m} a_{ij}.$$