Математические структуры и моделирование
2019. № 3(51). С. 88–96

УДК 519.237.7
DOI 10.25513/2222-8772.2019.3.88-96

ФАКТОРНЫЙ АНАЛИЗ КАЧЕСТВЕННЫХ ПОКАЗАТЕЛЕЙ

В.А. Шовин
научный сотрудник, e-mail: v.shovin@mail.ru

Институт математики им. С.Л. Соболева СО РАН (Омский филиал)

Аннотация. В работе предлагаются новые подходы к проведению факторного анализа для качественных показателей. Подход вычисления количественных показателей на базе многомерного шкалирования и матрицы взаимных расстояний объектов. И новый подход вычисления корреляционной матрицы. Оба метода могут быть предварительно использованы для проведения количественного факторного анализа. Также предлагается новый подход к восстановлению расстояний объектов и их расположения в многомерном пространстве. Основой метода является принцип выбора ближайшего объекта из двух других объектов. В результате расстояния между всеми парами объектов становятся определены. Такая матрица расстояний далее используется в методе многомерного шкалирования на базе метода Верле. Все объекты помещаются в многомерное пространство выбранной размерности. Такой метод восстановления расстояний между объектами и помещение их в многомерное пространство может быть использован предварительно также для классических методов классификации.

Ключевые слова: качественные показатели, факторный анализ, метод Верле, многомерное шкалирование, восстановление расстояний.

Введение

Большой потребностью анализа данных является обработка качественных показателей. Цель данной работы — осуществление возможности проведения факторного анализа качественных показателей.

Для реализации факторного анализа качественных показателей оказалось возможным использование количественного факторного анализа, когда качественные показатели преобразуются в количественные.

Классический количественный факторный анализ может использовать альтернативные исходные данные. Это матрица количественных показателей объектов или матрица корреляций показателей.

Расчёт коэффициента корреляции между двумя качественными показателями можно осуществить на базе информации о доле объектов, у которых эти качественные показатели одинаковы. Корреляция между такими показателями тем больше, чем больше представителей с одинаковыми значениями пары показателей.

Преобразование качественных показателей в количественные можно осуществить на базе метода многомерного шкалирования, предварительно вычислив матрицу взаимных расстояний.

Расстояние между двумя объектами с количественными показателями тем больше, чем больше различий в значениях одинаковых качественных показателей.

Чтобы была возможность использования смешанных данных (с качественными и количественными показателями), количественные показатели преобразуются в качественные. Для этого интервал каждого количественного показателя разбивается на небольшое число частей, и попадание показателя объекта в один из таких интервалов считается новым качественным значением.

Также предлагается новый подход восстановления расстояний объектов и их расположения в многомерном пространстве. Принцип выбора из двух объектов более близкого используется во многих алгоритмах. Например, в алгоритме сортировки и в классификации на базе FRiS–компактности. Такой принцип может быть также использован для восстановления расстояний между всеми парами объектов. Матрица взаимных расстояний является исходной для многих других алгоритмов. В частности, такая матрица может быть предварительно использована методом многомерного шкалирования и помещения объектов в многомерное пространство выбранной размерности. Известные координаты объектов в многомерном пространстве могут быть использованы во многих классических методах классификации. Такой принцип выбора ближайшего объекта, или предиката, может быть также использован для проведения социологического и психологического тестирования.

1. Вычисление корреляционной матрицы

Если в матрице исходных данных имеются количественные показатели, то их значения приводятся к качественным посредством приведения их значений в интервалы. Интервал их значений от min до max делится на 2 или 3 интервала.

Коэффициент корреляции между двумя качественными показателями может быть определён по формуле $r_{ij} = \sum_{k=1}^{n} \sum_{l=1}^{n} \delta_{ij}^{lk}$, где

$$\delta_{ij}^{lk} = \begin{cases} 1, & if \quad x_i^k = x_i^l \quad and \quad x_j^k = x_j^l \\ 0, & else \end{cases},$$

m – число исходных показателей, n – число объектов, x_i^k – i-ый качественный показатель у k-го объекта.

Далее в матрице элементов r_{ij} находятся минимальный (min) и максимальный (max) вне диагональные элементы. И интервал значений таких элементов переводится из $[\min, \max]$ в $[0, 1]$ по формуле:

$$r_{ij} = \frac{r_{ij} - \min}{\max - \min}.$$

2. Вычисление матрицы исходных данных

Для вычисления матрицы исходных количественных данных может быть использован алгоритм на базе многомерного шкалирования. Для этого необходимо подготовить матрицу взаимных расстояний объектов. Такую матрицу можно получить с помощью алгоритма Dist Redux или по следующей формуле:

$$d_{ij} = \sum_{k=1}^{m} \delta_k^{ij},$$

$$\delta_k^{lk} = \begin{cases} 0, & if \quad x_k^i = x_k^j. \\ 1, & else \end{cases}$$

3. Метод восстановления взаимных расстояний Dist Redux

Метод восстановления взаимных расстояний использует выбор ближайшего объекта из двух других или из двух предикатов.

Пусть z_i — i-ый объект из n множества объектов.

Путь p_i — i-ый предикат из m множества предикатов.

Пусть представлены все пары предикатов p_s,p_t, где $s \neq t$, s и $t = 1, \ldots, m$.

Все объекты проходят тест на выбор ближайшего предиката из пары p_s,p_t.

Все объекты, выбравшие предикат p_s, формируют свой класс объектов A_{st}. Аналогично для предиката p_t формируется свой класс объектов B_{st}. Для всех пар объектов из своих классов A_{st} или B_{st} взаимные расстояния не меняются. Для объектов из различных классов A_{st} и B_{st} взаимные расстояния увеличиваются на величину v, зависящую от p_s,p_t

$$d(z_i, z_j) := d(z_i, z_j) + v(p_s, p_t).$$

Далее перебираются все пары предикатов p_s,p_t и изменяются все взаимные расстояния $d(z_i, z_j)$.

4. Многомерное шкалирование

На базе метода Верле и полученной матрицы взаимных расстояний возможно провести многомерное шкалирование и поместить множество объектов в многомерное пространство выбранной размерности. Данный метод был предложен в статье [1].

5. Численный эксперимент

В качестве исходных данных были взяты данные, представленные в таблице 1.

Таблица 1. Исходные качественные показатели

Пол	муж	муж	муж	муж	муж	жен	жен	жен	жен	жен
Рост	выс	выс	сре	сре	низ	низ	сре	сре	сре	выс
Сила	сил	сил	сил	сла	сла	сла	сла	сил	сла	сла
Бюджет	бог	бог	бог	сре	бед	бед	бог	сре	сре	сре
Возраст	взо	сре	взр	мла	мла	мла	сре	сре	взр	сре
Ум	умн	умн	нет	нет	умн	нет	умн	умн	нет	умн
Кол. комп.	1	1	3	0	2	0	1	1	2	1

Таблица 2. Матрица корреляций

	Пол	Рост	Сила	Бюджет	Возраст	Ум	Кол. комп.
Пол	1	0,25	0,875	0,375	0,125	0,625	0,125
Рост	0,25	1	0,375	0,375	0,125	0,5	0,125
Сила	0,875	0,375	1	0,5	0,25	0,75	0,375
Бюджет	0,375	0,375	0,5	1	0	0,25	0,125
Возраст	0,125	0,125	0,25	0	1	0,625	0,5
Ум	0,625	0,5	0,75	0,25	0,625	1	1
Кол. комп.	0,125	0,125	0,375	0,125	0,5	1	1

Таблица 3. Матрица взаимных расстояний объектов

№ объекта	1	2	3	4	5	6	7	8	9	10
1	0	0,125	0,708	0,792	0,667	0,917	0,5	0,5	0,917	0,5
2	0,125	0	0,708	0,792	0,667	0,917	0,375	0,375	0,917	0,375
3	0,708	0,708	0	0,875	0,792	1,125	0,833	0,833	0,542	1,083
4	0,792	0,792	0,875	0	0,708	0,375	0,667	0,667	0,583	0,667
5	0,667	0,667	0,792	0,708	0	0,583	0,667	0,792	0,625	0,667
6	0,917	0,917	1,125	0,375	0,583	0	0,667	0,792	0,708	0,667
7	0,5	0,375	0,833	0,667	0,667	0,667	0	0,25	0,542	0,25
8	0,5	0,375	0,833	0,667	0,792	0,792	0,25	0	0,542	0,25
9	0,917	0,917	0,542	0,583	0,625	0,708	0,542	0,542	0	0,542
10	0,5	0,375	1,083	0,667	0,667	0,667	0,25	0,25	0,542	0

Результаты матрицы корреляций приведены в таблице 2.

Результат применения многомерного шкалирования к матрице дистанций представлен в таблице 4.

Таблица 4. Матрица исходных данных

Пол	-0,276	-0,198	-0,105	-0,081	0,08	0,22	0,508	-0,153	0,144	-0,025
Рост	0,231	0,375	0,178	-0,476	-0,48	-0,566	0,557	0,527	0,027	0,191
Сила	0,513	0,381	0,428	0,171	-0,402	-0,344	-0,172	-0,011	-0,348	-0,434
Бюджет	0,298	0,256	-0,456	0,013	0,293	0,298	0,022	-0,203	-0,59	0,235
Возраст	-0,566	-0,48	-0,11	0,436	-0,452	0,207	0,045	0,339	0,512	0,127
Ум	-0,436	-0,575	0,077	0,702	0,134	0,557	-0,377	-0,519	0,229	-0,498
Кол. комп.	-0,091	-0,222	0,74	-0,235	0,541	-0,267	-0,508	-0,523	0,441	-0,675

Матрица факторов, полученная методом Верле и вращением из матрицы исходных количественных данных, представлена в таблице 5.

Таблица 5. Матрица факторной структуры

Пол	**-1**	-0,009
Рост	**0,736**	0,676
Сила	**0,894**	0,449
Бюджет	0,431	**0,902**
Возраст	-0,362	**-0,932**
Ум	0,002	**1**
Кол. комп.	0,275	**0,962**

Матрица факторов, полученная многомерным шкалированием из корреляционной матрицы, представлена в таблице 6.

Таблица 6. Матрица факторной структуры

Пол	**0,956**	0,295
Рост	**-0,771**	0,637
Сила	**0,919**	0,395
Бюджет	0,605	**-0,796**
Возраст	-0,288	**-0,958**
Ум	0,06	**-0,998**
Кол. комп.	-0,154	**0,988**

В результате сравнения полученных факторных структур можно сделать вывод о взаимоподтверждаемости результатов двух подходов.

Такой подход для проведения факторного анализа можно рекомендовать, когда среди исходных качественных показателей есть показатели, не приводящиеся к количественным. Такие, например, как «пол», являющиеся исключительно качественными. Другие качественные показатели (все, кроме «пола» в данном примере) можно привести к количественным, используя их количественные характеристики («рост» — высота в сантиметрах, «сила» — жим в Ньютонах, «бюджет» — количество денег в рублях, «возраст» — в прожитых годах, «ум» — в величине коэффициента интеллекта IQ).

Стоить заметить, что в матрице исходных данных (см. табл. 4) 7 показателей являлись факторами, поставленными в соответствие исходным показателям. То есть были взяты взаимные расстояния объектов, и согласно им, в лучшем соответствии с ними, все объекты были помещены в 7–мерное факторное пространство. Поэтому в данной матрице строка с отметкой «пол» служит некоторому интегративному факторному показателю.

Также был проведён численный расчёт с данными артериальной гипертензии. По новой формуле коэффициента корреляции была рассчитана матрица корреляций, и на её основе была получена матрица факторной структуры (метод главных компонент), подвергнутая факторному вращению (критерий интерпретируемости).

Таблица 7. Новая факторная структура артериальной гипертензии

	*F*1	*F*2	*F*3	*F*4	*F*5
Вес	0,3269	**0,8957**	-0,0316	-0,0697	0,0254
ИМТ	0,2274	**0,8949**	-0,0567	-0,0955	0
ЧД	-0,4871	0,0494	-0,5293	**-0,5828**	0,272
С	-0,0852	**0,5525**	0,3681	-0,0568	-0,3236
Л	0	0,0181	**0,9143**	-0,1551	0,0404
КСР	**0,7028**	0,1688	-0,0152	0,1108	**0,5451**
КСО	0,3266	-0,0869	-0,3087	**0,8569**	0,0105
КДР	**0,902**	0,1849	0,0353	0,0112	0,3233
КДО	**0,9775**	-0,0178	-0,0445	0,0309	0,0401
УО	**0,9574**	0,0119	-0,0098	-0,147	-0,0825
МОС	**0,9355**	0,0186	-0,1139	-0,1231	-0,2582
ОПСС	**0,7399**	0,1605	-0,1475	-0,2109	-0,4773
ИХ	-0,0767	**0,8137**	-0,1215	-0,2146	-0,0238
ФВ	0,2578	**0,8169**	0,0595	0,1205	0,1596
ФУ	0,2132	**0,805**	0,0746	0,0264	0,0767

Полученные факторы артериальной гипертензии находятся в большом соответствии с факторами, полученными ранее в других работах [2]. Этот результат

также подтверждает работоспособность данного метода. Можно заметить, что в данной факторной структуре отдельные факторы из раннего исследования соединились. Это, видимо, произошло из-за большой дискретности количественных показателей, когда их значения были разделены всего на 2 интервала и большая часть информации о взаимоотношении показателей потерялась.

В качестве исходных данных для тестирования алгоритма Dist Redux были взяты 15 биофизических показателей для 131 лица с артериальной гипертензией начальной стадии [3]:

1) *вес,*
2) *индекс массы тела (ИМТ),*
3) *частота дыхания (ЧД),*
4) *сегментоядерные нейтрофилы (С),*
5) *лимфоциты (Л),*
6) *конечно-систолический размер левого желудочка (КСР),*
7) *конечно-систолический объём левого желудочка (КСО),*
8) *конечно-диастолический размер левого желудочка (КДР),*
9) *конечно-диастолический объём левого желудочка (КДО),*
10) *ударный объём (УО),*
11) *минутный объём сердца (МОС),*
12) *общее периферическое сосудистое сопротивление (ОПСС),*
13) *индекс Хильдебрандта (ИХ),*
14) *фракция выброса левого желудочка (ФВ),*
15) *фракция укорочения левого желудочка (ФУ).*

В качестве предикатов p_i использовались сами объекты z_i. В качестве принципа выбора ближайшего предиката — наименьшее расстояние $e(z_i, z_j)$ до объекта в евклидовом пространстве стандартизированных показателей. В качестве величины $v(p_s, p_t)$ используется расстояние $e(z_s, z_t)$.

В результате применения метода восстановления взаимных расстояний и многомерного шкалирования была получена структура объектов в многомерном пространстве показателей, подобная исходной структуре, но исключившая шум и разброс объектов около кривых зависимостей различных пар показателей.

6. Программная реализация

Методы обработки качественных показателей реализованы программно как web–приложение. Вычислительная часть приложения вынесена на сервер, написанный на языке PHP с использованием фреймворка Zend. Интерфейс приложения написан с использованием HTML, CSS, JavaScript, JQuery. Приложение многомерного шкалирования доступно по адресу: http://svlaboratory.org/application/multscalkind — после регистрации нового пользователя. Вычисление корреляционной матрицы реализовано в макросе Excel на языке VBA и доступно по адресу: http://svlaboratory.org/blog/blog-single/articleid/35.

Рис. 1. Результирующая структура объектов в плоскости первых двух показателей

Заключение

Предложены подходы обработки качественных показателей для проведения факторного анализа. По результату численного эксперимента подход вычисления корреляционной матрицы и подход вычисления матрицы исходных данных показали сходство факторных структур. Это подтверждает работоспособность таких методов. Предложен метод восстановления расстояний межу объектами. Данный метод позволяет вычислить матрицу взаимных расстояний между объектами и применить типичные методы классификации. Проведён численный эксперимент с количественными показателями объектов. В результате применения данного метода была получена структура объектов, лишённая шума и разброса объектов возле кривых зависимостей пар показателей.

Литература

1. Шовин В.А. Многомерное шкалирование на базе метода Верле // Математические структуры и моделирование. 2015. № 4(36). С. 117–122.
2. Шовин В.А., Гольтяпин В.В. Методы вращения факторных структур // Математические структуры и моделирование. 2015. № 2(34). С. 75–84.
3. Гольтяпин В.В., Шовин В.А. Косоугольная факторная модель артериальной гипертензии первой стадии // Вестник Омского университета. 2010. № 4. С. 120–128.

FACTOR ANALYSIS OF QUALITY INDICATORS

V.A. Shovin
Scientist Researcher, e-mail: v.shovin@mail.ru

S.L. Sobolev Institute of Mathematics (Omsk Branch), Siberian Branch of RAS

Abstract. The paper proposes new approaches for conducting factor analysis for qualitative indicators: the approach of calculating quantitative indicators on the basis of multidimensional scaling and matrix of mutual distances of objects; and a new approach to calculating the correlation matrix. Both methods can be preliminarily used for quantitative factor analysis. Also we propose a new approach to reconstructing the distances of objects and their location in a multidimensional space. The basis of the method is the principle of choosing the nearest object from two other objects. As a result, the distances between all pairs of objects are determined. Such a distance matrix is further used in the method of multidimensional scaling based on the Verlet method. All objects are placed in a multidimensional space of the selected dimension. This method of restoring distances between objects and placing them in a multidimensional space can also be used previously for classical classification methods.

Keywords: qualitative indicators, factor analysis, Verlet method, multidimensional scaling, distance reconstruction.

References

1. Shovin V.A. Mnogomernoe shkalirovanie na baze metoda Verle. Matematicheskie struktury i modelirovanie, 2015, no. 4(36), pp. 117–122. (in Russian)
2. Shovin V.A. and Gol'tyapin V.V. Metody vrashcheniya faktornykh struktur. Matematicheskie struktury i modelirovanie, 2015, no. 2(34), pp. 75–84. (in Russian)
3. Gol'tyapin V.V. and Shovin V.A. Kosougol'naya faktornaya model' arterial'noi gipertenzii pervoi stadii. Vestnik Omskogo universiteta, 2010, no. 4, pp. 120–128. (in Russian)

Дата поступления в редакцию: 01.07.2019