Математические структуры и моделирование 2018. № 4(48). С. 78–84

УДК 519.237.7
DOI: 10.25513/2222-8772.2018.4.78-84

ФАКТОРНЫЙ АНАЛИЗ НА БАЗЕ МЕТОДА K-СРЕДНИХ

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

Институт математики им. С.Л. Соболева Сибирского отделения РАН, Омск, Россия

Аннотация. Актуальной проблемой медицинских и математических исследований является анализ и поиск скрытых зависимостей в экспериментальных данных. Определение таких зависимостей позволяет построить модель явления или объекта, которая бы наиболее соответствовала экспериментальным данным и при этом обладала минимальной сложной структурой. Известным математическим и программным инструментом для автоматического построения таких моделей является факторный анализ. Представляются востребованными различные обобщения и модернизации методов факторного анализа. В статье предлагается новый подход к проведению факторного анализа на базе метода кластеризации данных k-средних и последующего факторного вращения. Факторный анализ выделяет из множества исходных показателей — k главных компонент или факторов — с наибольшей точностью аппроксимирующих разброс и распределение исходных данных. Такие главные компоненты формируют факторную структуру исходных данных. В качестве направлений и положений главных компонент могут быть использованы различные характеристические точки исходной структуры данных. В данной работе предлагается использовать центры кластеров исходных данных. Для разделения точек данных на классы существует большое число методов кластеризации. Наиболее популярным является метод k-средних. В результате метод k-средних позволяет найти факторную структуру в исходном многомерном пространстве данных из положений k центров выделенных кластеров. Последующее факторное вращение по оригинальному критерию интерпретируемости позволяет найти простую факторную структуру. Проведение численных экспериментов показало хорошее соответствие результатов данного метода факторного анализа с ранее известными методами. Предлагаемый метод факторного анализа обладает хорошей и превосходящей эффективностью по сравнению с другими методами факторного анализа. Он обладает меньшей сложностью и количеством необходимых действий для определения факторной структуры.

Ключевые слова: метод k-средних, факторный анализ, факторное вращение.

Введение

Факторное моделирование является одним из наиболее востребованных инструментов для изучения скрытых закономерностей в экспериментальных данных. Методы факторного анализа позволяют автоматически построить простую факторную модель данных и произвести диагностику новых объектов. Факторная структура данных отражает основные направления разброса данных в многомерном пространстве исходных показателей объектов. Суть факторной структуры данных — это каркас данных, с нужной точностью покрывающий исходные данные. Математическая постановка задач в области факторного анализа — это разработка подходов к вычислению и построению минимальных факторных структур в различных случаях распределений данных.

Факторный и кластерный анализ являются одними из самых популярных методов анализа данных и математической статистики. Кластерный анализ позволяет автоматически найти классы объектов, используя только информацию о количественных показателях объектов (обучение без учителя). Каждый такой класс может задаваться одним самым характерным для него объектом, например, средним по показателям. Существует большое число методов и подходов для классификации данных.

Факторный анализ позволяет найти факторную структуру показателей объектов такую, которая гипотетически объясняла бы значение экспериментальных данных и находилось в тесной математической связи с ними. Для получения первичного факторного решения на данный момент существует большое число методов. Например, итеративный метод и метод Якоби расчёта собственных векторов и собственных значений, центройдный метод, метод минимальных остатков, метод максимального правдоподобия [1]. Такие алгоритмы обладают большой вычислительной сложностью. Поэтому в работе предлагается подход, уменьшающий сложность алгоритма вычисления факторной структуры.

В данной статье предлагается использовать метод k-средних для определения центров кластеров и определения с помощью них факторной структуры, определяющей каркас данных.

Такая факторная структура может быть подвергнута факторному вращению для получения её простоты с точки зрения интерпретации. Существует большое количество методов факторного вращения. В работе предлагается использовать авторский критерий интерпретируемости, позволяющий упростить интерпретацию факторной структуры.

Современные исследования в области факторного анализа проводятся с целью нелинейных обобщений факторных структур [2,3], а также совершенствования эффективности методов в случае больших данных. Нелинейные факторные структуры позволяют с большей точностью аппроксимировать разброс данных, а также учитывать сложную топологию моделей данных. В таких методах часто задействуется инструмент нейронных сетей и их обучение. Огромным коммерческим спросом пользуется изучение больших данных, хранящихся в объёмных базах данных. Поэтому актуальными являются разработка и совершенствование методов, позволяющих быстро обрабатывать такие массивы

данных. Целью подобных исследований является снижение порядка сложности и количества действий алгоритмов.

В предлагаемом в работе подходе для расчёта факторной структуры замена метода классификации на другой метод позволит дополнительно улучшить скорость работы алгоритма и снизить его сложность.

1. Метод k-средних

Метод k-средних является одним из наиболее популярных методов кластеризации. Его целью является получение таких центров данных, которые соответствовали бы гипотезе компактности классов данных при их симметричном радиальном распределении. Одним из способов определить положение таких центров, при заданном их количестве k, является EM–подход.

В данном методе выполняются последовательно две процедуры.

1. Определение для каждого объекта данных X_i ближайшего центра C_j, и назначение данному объекту метки класса X_i^j. Далее для всех объектов определяется их принадлежность различным классам.

2. Вычисление нового положения центров всех классов.

Итеративно повторяя эти две процедуры из начального случайного положения центров k классов, можно добиться разделения объектов на классы, которые максимально соответствовали бы гипотезе радиальной компактности классов.

2. Факторный анализ

В линейной модели факторного анализа становится понятна матрица факторной структуры, определяемой из следующей формулы линейной связи значений факторов и исходных переменных [1,4]:

$$\begin{cases} a_{11}p_{1t} + a_{12}p_{2t} + \ldots + a_{1g}p_{gt} \to z_{1t}, \\ a_{21}p_{1t} + a_{22}p_{2t} + \ldots + a_{2g}p_{gt} \to z_{2t}, \\ \qquad\qquad\qquad \vdots \\ a_{m1}p_{1t} + a_{m2}p_{2t} + \ldots + a_{mg}p_{gt} \to z_{mt}, \end{cases} \tag{1}$$

где матрица $A \underset{m\times g}{\leftrightarrow} a_{ij}$ называется матрицей факторной структуры размерности $m \times g$ весовых коэффициентов и где m — число изучаемых параметров, g — число общих факторов.

Матрица $Z \underset{m\times n}{\leftrightarrow} z_{ij}$ — матрица значений измеряемых переменных у исследуемых объектов или состояний объекта размерности $m \times n$, где m — число измеряемых параметров, n — число объектов или состояний объекта (объём выборки).

Матрица $P \underset{g\times n}{\leftrightarrow} p_{ij}$ — матрица значений латентных переменных (факторов) объектов размерности $g \times n$, где g — число латентных параметров.

На вид факторной структуры A налагаются дополнительные ограничения: общности переменных факторной структуры должны быть не больше 1, а также не меньше определённого порога значимости:

$$h_i = \sqrt{\sum_{k=1}^{g} a_{ik}^2} \leqslant 1.$$

Результатом факторного анализа является определение двух матриц *A* и *P*.

3. Гипотеза соответствия

Можно предположить, что матрица координат центров g классов данных может быть использована как матрица факторной структуры A. В то время как матрица вероятностного отношения объектов к различным классам (матрица дистанций объектов до центров классов) — как матрица значений факторов *P*.

Для полного соответствия координаты центров g классов приводятся в диапазон [-1, 1] с приведением общностей $h_i = 1$ по формулам:

$$\min_i = \min_{j=1\ldots g} a_{ij};$$

$$\max_i = \max_{j=1\ldots g} a_{ij};$$

$$a_{ij} := \alpha a_{ij} + \beta.$$

$$\alpha = \frac{2}{\max - \min}, \beta = -\frac{\max + \min}{\max - \min}.$$

$$a_{ij} := \frac{2a_{ij} - \max_i - \min_i}{\max_i - \min_i};$$

$$a_{ij} := \frac{a_{ij}}{\sqrt{\sum_{k=1}^{g} a_{ik}^2}}.$$

Численный эксперимент

В качестве исходных данных были взяты 15 биофизических показателей для 131 лица с артериальной гипертензией начальной стадии [5]:

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 таблице 1 приведена матрица факторной структуры, выделенная по методу главных компонент и подвергнутая факторному вращению по критерию интерпретируемости [6].

В таблице 2 приведена матрица факторной структуры, выделенная по гипотезе соответствия кластерного анализа и факторного анализа и подвергнутая факторному вращению по критерию интерпретируемости.

Незначительное несоответствие факторных структур двух методов могло быть результатом применения к матрице данных другого метода извлечения грубых ошибок. В целом по результату сравнения значимых факторных нагрузок можно утверждать, что гипотеза соответствия факторного и кластерного анализа оказалась правдивой на данных артериальной гипертензии.

Сложность алгоритма факторного анализа на базе метода k-средних $\sim m^2$ и $\sim n$. В то время как, например, подход на базе метода Якоби для вычисления собственных векторов и собственных значений имеет сложность $\sim m^4$ и $\sim n$.

4. Программная реализация

Метод кластеризации k-средних был реализован 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 доступно по адресу: http://svlaboratory.org/application/klaster — после регистрации нового пользователя. Приложение позволяет визуализировать принадлежность объектов различным кластерам в заданной плоскости координат.

Заключение

Предложен метод получения факторной структуры данных из матрицы координат центров k классов, выделяемых методом k-средних. Такой подход к извлечению факторной структуры является более эффективным в случае больших данных, т. к. требует меньшего количества действий, чем, например, метод Якоби для вычисления главных компонент.

Таблица 1. Факторная структура по критерию интерпретируемости, полученная из матрицы главных факторов

	*F*1	*F*2	*F*3	*F*4	*F*5
Вес	0,1949	0,0000	-0,021	-0,0062	**0,7783**
ИМТ	0,16	-0,0815	0,0000	0,0000	**0,7679**
ЧД	0,2165	-0,0211	0,01	**-0,8305**	0,0261
С	-0,0598	-0,0382	**-0,888**	0,0225	0,1504
Л	0,0000	0,0000	**0,9059**	0,0205	-0,19
КСР	**0,8631**	-0,4849	0,008	-0,0288	0,0149
КСО	**0,8502**	-0,4783	-0,0095	-0,0214	0,0097
КДР	**0,9721**	-0,0817	0,0251	0,0187	-0,0001
КДО	**0,9734**	-0,1221	-0,0066	0,0000	0,0178
УО	**0,9085**	0,2377	0,0152	0,0023	0,0202
МОС	**0,8934**	0,2084	-0,0105	0,0107	-0,0357
ОПСС	**-0,7374**	-0,2506	0,0000	0,0829	0,0839
ИХ	-0,1191	0,0000	0,0003	**0,8462**	0,0000
ФВ	-0,2005	**0,8173**	-0,0213	-0,0535	-0,0217
ФУ	-0,1863	**0,7167**	0,0441	0,0000	0,0094

Таблица 2. Факторная структура по критерию интерпретируемости, полученная из кластерного анализа

	*F*1	*F*2	*F*3	*F*4	*F*5
Вес	-0,3482	0,0147	-0,3964	0,1242	**0,774**
ИМТ	-0,2878	-0,0165	-0,3872	0,3025	**0,8224**
ЧД	-0,3177	-0,0757	-0,4894	**0,7997**	-0,0493
С	0,3827	-0,0557	**-0,8118**	-0,1787	0,2269
Л	-0,3032	-0,0097	**0,797**	0,0141	-0,4382
КСР	**-0,9656**	-0,046	-0,103	0,0591	0,0971
КСО	**-0,9596**	-0,0315	-0,152	0	-0,0075
КДР	**-0,9651**	-0,0355	0,0096	0,0537	0,1883
КДО	**-0,978**	-0,02	-0,065	-0,0082	0,0537
УО	**-0,983**	-0,0036	0,032	-0,0195	0,1101
МОС	**-0,9776**	-0,0148	0,0174	-0,0557	0,1226
ОПСС	**0,8594**	0,0181	-0,2225	-0,1846	-0,3785
ИХ	0,1003	-0,0087	0,6124	**-0,7831**	0
ФВ	**0,9086**	-0,0241	0,1947	0,1185	0,2376
ФУ	**0,9093**	0,3854	-0,1322	0,0765	0,0402

Литература

1. Харман Г. Современный факторный анализ / пер. с англ. В.Я. Лумельского. М. : Статистика, 1972.
2. Gorban A., Kegl B., Wunsch D., Zinovyev A. Principal Manifolds for Data Visualization and Dimension Reduction / Springer-Verlag Berlin Heidelberg. 2008. V. 58. P. 340.
3. Шовин В.А. Факторное моделирование артериальной гипертензии на базе метода Верле / Междисциплинарные исследования в области математического моделирования и информатики // Материалы 7-й научно-практической internet-конференции. 2016. С. 190–198.
4. Иберла К. Факторный анализ / пер. с нем. В.М. Ивановой. М. : Статистика, 1980.
5. Гольтяпин В.В., Шовин В.А. Косоугольная факторная модель артериальной гипертензии первой стадии // Вестник Омского университета. 2010. № 4. С. 120–128.
6. Шовин В.А., Гольтяпин В.В. Методы вращения факторных структур // Математические структуры и моделирование. 2015. № 2. С. 75–84.

FACTOR ANALYSIS BASED ON THE K-MEANS METHOD

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

The Federal State Budgetary Institution of Science Sobolev Institute of Mathematics of the Siberian Branch of RAS (Omsk Branch)

Abstract. Actual problems and research methods are analysis and search for hidden dependencies in experimental data. The definition of such dependencies allows them to construct a model of phenomenon or object that would best fit the experimental data and at the same time have a minimally complex version. Known mathematical and software tools for the automatic construction of such models is a factor analysis. Various generalizations and modernizations of the methods of factor analysis are in demand. The article proposes a new approach to factor analysis based on the k-means data clustering method and subsequent factor rotation. Factor analysis identifies from a set of initial indicators k main components or factors — with the greatest accuracy approximating the scatter and distribution of the initial data. These main components form the factorial structure of the source data. Various characteristic points of the original data structure can be used as the directions and aspects of the main components. In this paper, we propose to use centers of raw data clusters. There is a large number of clustering methods for dividing data points into classes. The most popular is the k-means method. As a result, the k-means method allows finding the factor structure in the original multidimensional data space from the positions of the k centers of the selected clusters. The subsequent factor rotation according to the original criterion of interpretability allows us to find a simple factor structure. Conducting numerical experiments showed good agreement with the results of this method of factor analysis with previously known methods. The proposed method of factor analysis has good and superior efficiency compared with other methods of factor analysis. It has limited capabilities and resources needed to determine the factor structure.

Keywords: k-means method, factor analysis, factor rotation.

Дата поступления в редакцию: 07.11.2018