УДК 519.23

Неитерационный алгоритм визуализации многомерных данных

Калинкин А.А.
Алтайский государственный университет, г. Барнаул
kalinkin.7621@gmail.com

Аннотация

В связи с появившейся сегодня возможностью обрабатывать большие объемы данных особое значение приобретает задача первоначального грубого анализа этих данных с целью сформулировать предварительные направления исследования и сделать прикидочное заключение о возможных его результатах. Обычно подобный анализ проводится путём изучения некоторых изображений, но, если данные имеют достаточно большую размерность, то построение изображений, адекватно отображающих структуру этих данных, представляет собой серьезную математическую задачу. В статье представлен новый алгоритм построения неискаженных изображений многомерных данных в случае, когда подобные изображения возможны.

Ключевые слова: многомерные данные, визуализация статистических данных, неискаженное изображение.

1. Обычный недостаток методов визуализации многомерных данных

Визуализация (изображение) данных – представление в наглядной форме данных эксперимента или результатов некоторого исследования. Создание визуализаций востребовано в любом исследовании, в силу чего этой задаче посвящены многочисленные научные статьи, обзоры и монографии, см, например, [1–3] и библиографию там. Обычно каждый объект наблюдения задан собственным набором значений числовых показателей (первый вид данных). Количество показателей, значения которых предполагаются известными у каждого из объектов, называется размерностью задачи. Данные размерности 2 или 3 изобразить просто. Но, если данные имеют более высокую размерность, задача становится нетривиальной. Также возможны данные второго вида, когда сами объекты никак не заданы, но известны величины их попарных отличий друг от друга. Перевести данные из первого вида во второй несложно, поэтому без ограничения общности можно визуализировать данные именно второго вида. Существующие на сегодня методы решения такой задачи практически всегда искажают истинную картину данных. В частности, методы главных компонент [4, 5] и многомерного шкалирования [6, 7], используемые для визуализации многомерных данных, по-видимому, наиболее часто, искажают расстояния между объектами на результирующих рисунках. И, хотя эти искажения незначительны, но, при возможности избежать их, это стоит сделать.

2. Построение неискажённого изображения

Поскольку изображения на плоскости, вероятно, являются наиболее наглядными, сформулируем основную задачу настоящей работы следующим образом: по заданной таблице попарных расстояний $d_{i,j}$, $i, j = 1, \ldots, n$ между небольшим количеством объектов

$A_1, \ldots, A_n$ выяснить, возможно ли построить плоское изображение без искажений этих расстояний и, если это возможно, предложить алгоритм этого построения. Условимся не различать решений, которые можно перевести друг в друга сдвигами или симметриями.

Начнем с алгоритма для двух объектов и будем постепенно увеличивать их количество. Пусть объектов два. Для их визуализации возьмем произвольную точку A_1 и в произвольном направлении от нее построим отрезок длины $d_{1,2}$, на конце которого разместим точку A_2. Это построение возможно при произвольном неотрицательном значении $d_{1,2}$. Заметим, что, как бы мы не строили отрезок заданной длины, такой способ можно перевести в любой другой параллельным переносом или осевой симметрией, при необходимости, меняющей концы отрезка местами. Значит, решение здесь будет единственным.

Для трёх точек: сначала с помощью метода, описанного выше, построим отрезок, концы которого изображают первую и вторую точки A_1 и A_2. Далее строим окружности с центрами в этих точках радиусов $d_{1,3}$ и $d_{2,3}$ соответственно. Поскольку мы считаем, что известные нам различия между объектами даны точно, то эти две окружности обязательно пересекаются в силу справедливости неравенства треугольника. Их пересечение обычно представляет собой две точки. Эти точки симметричны относительно прямой A_1A_2. Поэтому, хотя в качестве изображения третьего объекта можно выбрать любую из них, решение будет также одно, и, как следует из описанного построения, такое решение всегда существует.

Для четырёх точек: для начала нужно проверить, возможно ли такое построение. Известно, что тетраэдр однозначно определяется заданными длинами своих ребер (см., например, [8]), следовательно, наши 4 точки обязательно будут являться вершинами этого тетраэдра. Таким образом, если все вершины этого тетраэдра не лежат в одной плоскости, то неискаженное построение невозможно.

Лемма 1. *[[8, с. 3]] Пусть известны все попарные расстояния $d_{i,j}$, $i, j = 1, \ldots, 4$ между четырьмя точками. Тогда объем тетраэдра с вершинами в этих точках независимо от конкретного его расположения в трехмерном пространстве, может быть найден из формулы:*

$$\begin{aligned} 144V^2 = [&d_{1,2}^2 d_{4,3}^2 \left(d_{1,4}^2 + d_{1,3}^2 + d_{2,4}^2 + d_{2,3}^2 - d_{1,2}^2 - d_{4,3}^2\right) + \\ &+ d_{1,4}^2 d_{2,3}^2 \left(d_{1,2}^2 + d_{1,3}^2 + d_{2,4}^2 + d_{4,3}^2 - d_{1,4}^2 - d_{2,3}^2\right) + \\ &+ d_{1,3}^2 d_{2,4}^2 \left(d_{1,2}^2 + d_{1,4}^2 + d_{4,3}^2 + d_{2,3}^2 - d_{1,3}^2 - d_{2,4}^2\right) - \\ &- d_{1,2}^2 d_{1,4}^2 d_{2,4}^2 - d_{1,4}^2 d_{1,3}^2 d_{4,3}^2 - d_{1,2}^2 d_{1,3}^2 d_{2,3}^2 - d_{2,4}^2 d_{4,3}^2 d_{2,3}^2]. \end{aligned} \tag{1}$$

Из этой леммы вытекает, что неискаженное изображение четырех точек на плоскости возможно тогда и только тогда, когда $V = 0$. Тогда строим первые три точки из наших четырех так, как это было описано выше. Затем построим возможные четвёртые точки $A_{4,1}$, $A_{4,2}$ пересечением окружностей в центрах A_1 и A_3 с радиусами $d_{1,4}$, $d_{3,4}$. Обозначим расстояния от точек A_1 и A_3 до точки $A_{4,1}$, как $d_{4,1;1}$ и $d_{4,1;3}$ соответственно. Аналогично для точки $A_{4,2}$. Нетрудно заметить, что $d_{4,1;1} = d_{4,2;1} = d_{1,4}$ и $d_{4,2;1} = d_{4,2;3} = d_{3,4}$. Таких точек A_4, для которых выполняются оба равенства, всего две. Вывод: если решение есть, то это одна из точек $A_{4,1}$ или $A_{4,2}$. Далее найдем расстояния от точки A_2 до $A_{4,1}$ и до $A_{4,2}$ ($d_{4,1;2}, d_{4,2;2}$). Точка, для которой полученное расстояние совпало с расстоянием $d_{4,2}$, и будет подходящей, ее мы и выберем для изображения четвертого объекта.

Описанный алгоритм построения для двух, трех и четырех точек нетрудно перевести с геометрического языка на язык формул, показывающих, как находить координаты очередной точки по координатам уже имеющихся.

Точка A_2 может быть расположена на оси абсцисс на расстоянии $d_{1,2}$ от точки A_1 с координатами $(0, d_{1,2})$. Для нахождения третьей точки нужно найти точку пересечения

окружностей с центрами в A_1 и A_2. Нужные координаты получаем, решая систему из двух уравнений относительно переменных x_3, y_3. Получим:

$$x_3 = \frac{-d_{2,3}^2 + d_{1,3}^2 + d_{1,2}^2}{2d_{1,2}}; \tag{2}$$

$$y_3 = \sqrt{d_{1,3}^2 - \left(\frac{-d_{2,3}^2 + d_{1,3}^2 + d_{1,2}^2}{2d_{1,2}}\right)^2}. \tag{3}$$

Для нахождения координат четвёртой точки решим систему относительно x_4, y_4:

$$\begin{cases} x_4^2 + y_4^2 = d_{1,4}^2, \\ (x_4 - x_3)^2 + (y_4 - y_3)^2 = d_{3,4}^2. \end{cases}$$

Полученные решения

$$y_{4,4'} = \frac{2y_3c \pm \sqrt{4c^2y_3^2 - 4(y_3^2 + x_3^2)(c^2 - d_{1,4}^2x_3^2)}}{2(y_3^2 + x_3^2)}, \text{ где } \quad c = \frac{d_{3,4}^2 - x_3^2 - y_3^2 - d_{1,4}^2}{-2}; \tag{4}$$

$$x_{4,4'} = \frac{c - y_{4,4'}y_3}{x_3}, \text{ где } \quad c = \frac{d_{3,4}^2 - x_3^2 - y_3^2 - d_{1,4}^2}{-2}. \tag{5}$$

нужно будет проанализировать описанным выше образом.

3. Основной алгоритм

Используя результаты предыдущего раздела, можно теперь сформулировать основной алгоритм работы.

На входе – число объектов и матрица попарных расстояний между ними.

Шаг 1. Проверка числа объектов. Если будет введены данные только одного объекта, выведется “Число объектов должно быть не менее 2” и повторяем просьбу о вводе. Если два или три – к шагу 3, иначе к шагу 2.

Шаг 2. Проверка возможности построения. Перебираем все возможные четвёртые точки и проверяем по формуле (1), что $V = 0$. Если хотя бы один раз это не выполняется, то выход с сообщением “Неискаженное изображение построить невозможно”. Иначе к шагу 3.

Шаг 3. Строим две точки с координатами $(0, 0)$ и $(0, d_{1,2})$. Все ли точки построены? Если да – выход из алгоритма, иначе к шагу 4.

Шаг 4. Строим третью точку, определяя её координаты по формулам (2) и (3). Все ли точки построены? Нет – запоминаем основную 3-конструкцию, т.е. координаты первых трех построенных точек, и к шагу 5.

Шаг 5. Берем очередную точку и строим её относительно основной 3-конструкции. Вычисляем координаты по формулам (4) и (5). Получаем две потенциально возможные точки $A_4(x_4, y_4)$ и $A'_4(x'_4, y'_4)$. Чтобы выбрать нужную подсчитаем расстояния между A_4 и A_2. Аналогично для A'_4. Нужная точка – та, у которой расстояние совпало с исходным $d_{2,4}$. Все ли точки построены? Если да – выход из алгоритма, иначе – повторяем шаг 5.

4. Пример работы алгоритма

Алгоритм из предыдущего раздела был реализован в виде компьютерной программы на языке Python. Для иллюстрации её работы возьмем 4 точки трехмерного пространства, лежащие в плоскости, заданной уравнением $2x - y - z + 1 = 0$. Тогда неискаженное их изображение будет заведомо возможно.

Таблица 1

Координаты точек в примере

	x	y	z
A_1	0	0	1
A_2	1	0	3
A_3	2	3	2
A_4	-2	-2	-1

Вычислим попарные расстояния:

Таблица 2

Таблица попарных расстояний для четырёх точек

	A_1	A_2	A_3	A_4
A_1	0	2.236	3.742	3.464
A_2	2.236	0	3.317	5.385
A_3	3.742	3.317	0	7.071
A_4	3.464	5.385	7.071	0

Запустим программу со этими же данными. На выходе получаем координаты точек и конечное изображение:

```
A1 = (0, 0)
A2 = (2.236, 0)
A3 = (1.789, 3.287)
A4 = (-2.684, -2.190)
```

Рисунок 1. Вывод в консоль результатов работы программы

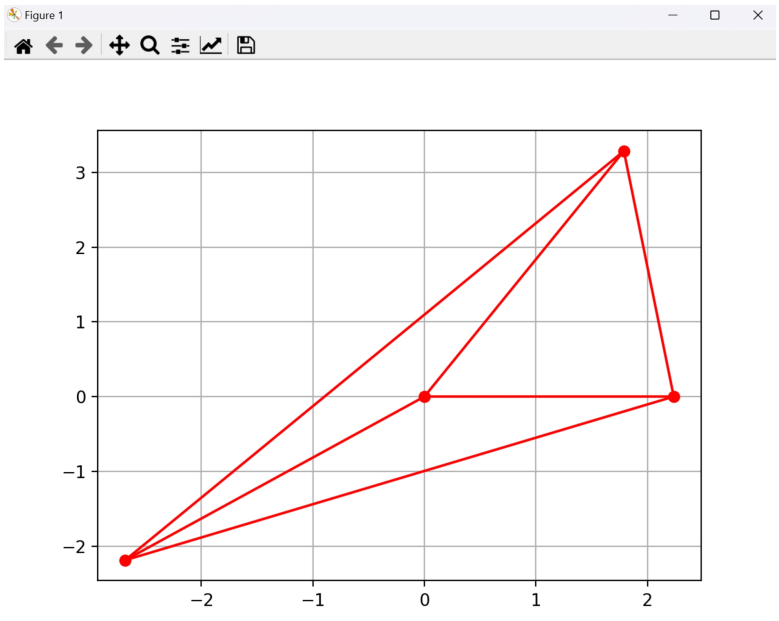

Рисунок 2. Итоговое изображение

5. Сравнение с классическими алгоритмами

К данным из предыдущего раздела применим метод многомерного шкалирования с помощью статистического процессора IBM SPSS 21. Здесь наши точки A_1, A_2, A_3, A_4 – это VAR00001, VAR00002, VAR00003, VAR00004 соответственно. Результаты приведены в таблице 3.

Таблица 3

Координаты в двумерном пространстве, полученные методом многомерного шкалирования

	Измерение	
	1	2
VAR00001	.104	-.060
VAR00002	-.280	-.343
VAR00003	-.680	.283
VAR00004	.856	.120

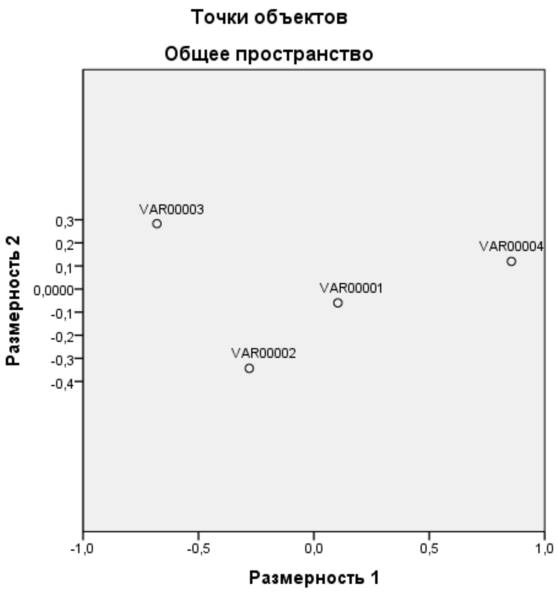

Рисунок 3. Изображение, полученное методом многомерного шкалирования

Вычислим попарные расстояния между точками, полученными методом многомерного шкалирования, после чего заполним таблицу 4.

Таблица 4

Попарные расстояния между точками, полученными методом многомерного шкалирования

	A_1	A_2	A_3	A_4
A_1	0	0.477	0.855	0.773
A_2	0.477	0	0.743	1.226
A_3	0.855	0.743	0	1.544
A_4	0.773	1.226	1.544	0

Можно увидеть, что если наше изображение (рисунок 2) отобразить симметрично от-

носительно вертикали и перевернуть влево на 45 градусов, то оно будет очень похоже на изображение, полученное методом многомерного шкалирования.

Применим метод главных компонент к трёхмерным координатам (собственные значения 2.638, 0.362, $1.22 \cdot 10^{-16}$) и получим координаты в двумерном пространстве.

Таблица 5

Координаты в двумерном пространстве, полученные методом главных компонент

	x	y
A_1	-0.147	-0.032
A_2	0.472	1.358
A_3	0.999	-1.034
A_4	-1.325	-0.291

Изобразим эти точки:

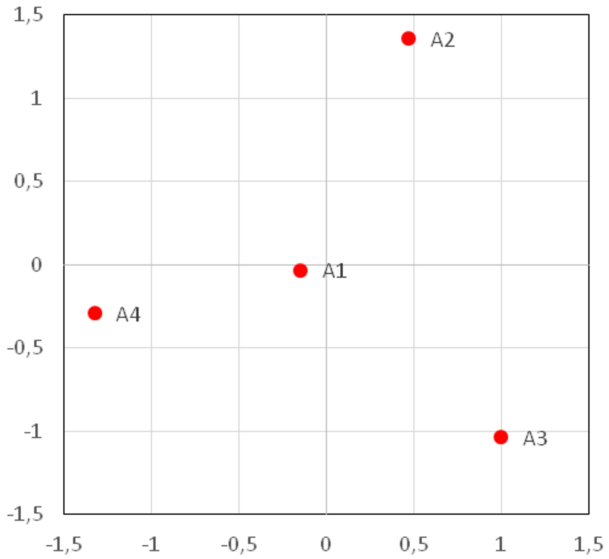

Рисунок 4. Изображение, полученное методом главных компонент

Вычислим попарные расстояния между точками, полученными с помощью метода главных компонент. Они приведены в таблице 6.

Таблица 6

Попарные расстояния между точками, полученными методом главных компонент

	A_1	A_2	A_3	A_4
A_1	0	1.522	1.522	1.206
A_2	1.522	0	2.449	2.440
A_3	1.522	2.449	0	2.440
A_4	1.206	2.440	2.440	0

Чтобы сравнить полученные при помощи различных методов результаты, для всех вариантов решения, вычислим значения так называемого стресс-критерия по формуле $S = \sum\limits_{i,j} \frac{(d_{i,j} - o_{i,j})^2}{d_{i,j}}$, где $d_{i,j}$ – реальное расстояние между i-м и j-м объектами, а $o_{i,j}$ – расстояние между их образами. Величина этого критерия позволяет оценить степень искаженности расстояний на изображении. Для рассматриваемого метода этот критерий, очевидно,

равен 0.

Для метода главных компонент стресс-критерий равен 15,77, а у метода многомерного шкалирования он составляет 30,46. Сравнив с реальными попарными расстояниями, можно заметить, что оба метода исказили расстояния между точками. Наш алгоритм эти расстояния изобразил точно.

6. Некоторые выводы

Методы многомерного шкалирования и главных компонент как методы визуализации многомерных данных многократно проверены и уверенно вошли в инструментарий современной статистики. Но, как всегда, их большая универсальность оставляет возможности для улучшений при решении конкретных задач. Одной из подобных задач является визуализация малого числа объектов.

При решении подобной задачи бывает нетрудно прежде, чем пытаться произвести визуализацию, проверить возможность построения неискаженного ее варианта. При наличии такой возможности предложенный в статье алгоритм позволяет получить результаты лучшие, чем дают упомянутые выше методы.

В развитие предлагаемого алгоритма автор планирует в дальнейшем предложить методику построения изображений небольшого числа объектов на плоскости также и в случае невозможности их неискаженного построения, все еще обладающих меньшими искажениями, чем классические методы. Этого, видимо, можно добиться за счет того, что даже полный перебор возможностей при небольшом числе исходных объектов оказывается вполне осуществимым.

Список литературы

1. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. — 2-е изд. — СПб. : Питер, 2012. — 704 с.

2. Mohammed L.T., Al Habshy A.A., El Dahshan K.A. Big Data Visualization: A Survey // 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). — Ankara, Turkey, 2022. — P. 1–12.

3. Bandalos D.L., Boehm-Kaufman M.R. Four common misconceptions in exploratory factor analysis // Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences. — London : Taylor & Francis, 2008. — P. 61–87.

4. Gorban A.N., Kegl B., Wunsch D., Zinovyev A.Y. Principal Manifolds for Data Visualisation and Dimension Reduction // Lecture Notes in Computational Science and Engineering, 58 / Ed. by Zinovyev A.Y. — New York : Springer, 2007. — 364 p.

5. Мокеев В.В. Метод главных компонент в задачах экономического анализа и прогнозирования: монография. — Челябинск : ЮУрГУ, 2009. — 168 с.

6. Torgerson W.S. Multidimensional scaling: I. Theory and method // Psychometrika. — 1952. — Т. 17. — С. 401–419.

7. Толстова Ю.Н. Основы многомерного шкалирования. — М. : КДУ, 2006. — 160 с.

8. Сабитов И.Х. Объёмы многогранников. — М. : МЦНМО, 2002. — 18 с.