УДК 519.7

МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ В СЛУЧАЕ МАТРИЦ ПОПАРНЫХ РАССТОЯНИЙ С ЭЛЕМЕНТАМИ ИЗ КОНЕЧНОГО МНОЖЕСТВА

© Иофина Г.В.

Московский физико-технический институт (государственный университет)
факультет управления и прикладной математики
РФ, г. Долгопрудный, Институтский переулок, 9, 141700

e-mail: *giofina@gmail.com*

Abstract. The paper considers multidimensional scaling problem with proximity matrix consisting of two different non-diagonal elements. Elements of the superdiagonal matrix don't decrease along the string and don't increase along the column. The structure of considered matrix is described. In one-, two- and three-dimensional spaces all available proximity matrices are listed. Matrices that can be considered as proximity matrices in t-dimensional euclidian space have their dimensions in a definite interval. The exact bounds of this interval are also found.

Введение

Многомерное шкалирование – один из методов анализа данных, основанный на исследовании близостей [1]. Под понятием близости часто понимаются сходства или различия между объектами в евклидовом пространстве. В простейшем случае в качестве близости между объектами берется расстояние между ними. Задачей многомерного шкалирования является поиск представления матрицы близости системой точек в евклидовом пространстве некоторой размерности.

В некоторых случаях можно считать, что расстояния между объектами принимают одно из двух значений. Это можно сделать, например, когда важен факт близости или дальности объектов, а не их точные значения. Тогда средние расстояния между близкими объектами можно считать равным 1, а далекими – M.

Известно, что в данном случае максимальное число точек, которое можно расположить наилучшим образом в пространстве размерности $t = 1$, равно трем, для $t = 2$ – пяти (точки располагаются в вершинах правильного пятиугольника), для $t = 3$ — шести (точки находятся в вершинах правильного октаэдра или правильной треугольной призмы, или правильного пятиугольника и соответствующей точке на его оси). Точных значений для $t \geq 4$ неизвестно, но для них известны довольно точные границы: $t(t+1)/2 \leq N \leq (t+1)(t+2)/2$. Для случая расположения точек на сфере в пространстве размерности $t \in \{2, 6, 22\}$ Дельсартом было показано, что можно разместить $t(t+3)/2$ точек [2].

Может оказаться, что объекты можно линейно упорядочить по убыванию по удаленности вначале от первого объекта, потом от второго и т. д. Это накладывает на матрицы попарных расстояний условие порядка – неубывания по строкам и невозрастания по столбцам элементов матрицы.

Очевидно, что условию порядка подчиняются не все матрицы попарных расстояний, удовлетворяющие аксиомам метрик. В первой части работы находится общий

вид матриц попарных расстояний, элементы которых удовлетворяют условию порядка. Во второй части описаны все возможные матрицы попарных расстояний в евклидовых пространствах размерностей $t = 1, 2, 3$. В заключительной части работы находятся ограничения на количество объектов, которые можно разместить в пространстве размерности t.

1. Структура матриц

Определение 1. Будем говорить, что элементы матрицы $A = \{a_{ij}\}$ размерности $N \times N$ удовлетворяют аксиомам метрики, если они удовлетворяют следующим условиям:

1. $a_{ij} = 0 \Leftrightarrow i = j$,
2. $a_{ij} = a_{ji}$,
3. $a_{ij} + a_{jk} \geq a_{ik}$, $\forall i, j, k = 0, \ldots, N-1$.

Рассмотрим матрицу $A = \{a_{ij}\}$, элементы которой удовлетворяют условию порядка и первым двум аксиоам метрики, то есть она симметричная, с нулевой диагональю, и для элементов, находящихся выше главной диагонали, выполнены условия

$$a_{ij} < a_{ik} \forall i, j, k : i > j, j < k;$$
$$a_{ij} > a_{kj} \forall i, j, k : i > j, i < k.$$

Далее будем рассматривать элементы, находящиеся выше главной диагонали. Элементы, находящиеся ниже главной диагонали, определяются автоматически из-за симметричности матрицы.

Условия означают, что в каждой строке после диагонального нуля стоит какое-то количество единиц, а потом идут элементы, равные M. Причем, если в строке i первая тройка стоит на l-ом месте, то в строке $j > i$ первая тройка будет стоять на месте с номером $k \geq l$.

Наложим на элементы матрицы выполнение третьей аксиомы метрики – неравенство треугольника.

Первый случай. $M > 2$.

Пусть в первой строке матрицы первые l элементов равны единицам, а остальные M, то есть $a_{0i} = 1, \forall i \leq l$ и $a_{0i} = M, \forall i > l$.

Так как $a_{0i} = a_{0j} = 1$, $\forall i, j \leq l$, то для выполнения неравенств треугольника должно быть выполнено $a_{0i} + a_{0j} = 1 + 1 \geq a_{ij} \in \{1, M\}$. Поэтому $a_{ij} = 1$.

Если $i \leq l$, $j > l$, то есть $a_{0i} = 1$, $a_{0j} = M$, то для выполнения неравенств треугольника необходимо, чтобы $a_{0i} + a_{ij} = 1 + a_{ij} \geq a_{0j} = M$. Так как $a_{ij} \in \{1, M\}$, то $a_{ij} = M$.

Поэтому, если в строке есть единица, то она полностью определена. Следовательно, определена первая $l + 1$ строка . В них $\forall i \leq l$ $a_{ji} = 1$ и $\forall i > l$ $a_{ji} = M$. По симметрии определяются первые $l + 1$ столбцов.

Остальные элементы матрицы можно определять независимо от первых $l + 1$ строк (столбцов). Действительно, $\forall i > l + 1$ для выполнения неравенств треугольника с элементами из первых строк необходимо, чтобы

$a_{ki} + a_{kj} = M + M = 2M \geq a_{ij} \in \{1, M\}$, что будет выполняться для любых a_{ij} (так как $a_{ki} + a_{ij} = M + a_{ij} \geq a_{kj} = M$ верно для любых $a_{ij} \in \{1, M\}$).

Следовательно, можно выкинуть определенные строки и столбцы и рассматривать матрицу размером $N - l - 1$. Таким образом можно определить всю матрицу.

Причем число матриц размерности $M \geq 3$ может быть определено, как $f(M) = \sum\limits_{k=0}^{M-1} f(M-1-k) = \sum\limits_{k=0}^{M-1} f(k)$, где $f(0) = f(1) = 0$, $f(2) = 2$, $f(3) = 4$. Или в явном виде: $f(M) = 6 \cdot 2^{M-3}$, $\forall M \geq 4$.

Второй случай. $M \leq 2$. Единственный невырожденный случай это $M = 2$. В данном случае $a_{ij} \in \{1, 2\}$. Любая комбинация трех элементов из этого множества не будет нарушать неравенство треугольника.

Число матриц в данном случае $f(2) = \frac{C_{2\cdot 3}^{3}}{3+1} = 5$.

2. Матрицы, задающие конфигурации в пространствах размерности $t = 1, 2, 3$

Итак, рассматриваемые матрицы имеют структуру, состоящую из единичных блоков, расположенных по диагонали, которые отделены друг от друга строками, состоящими из элементов, равных M (кроме нулевых диагональных элементов). Такая структура дает в евклидовом пространстве кластеризацию близких точек, имеющих структуру правильных симплексов. Точки из различных кластеров находятся на расстояниях равных M.

Найдем всевозможные матрицы, удовлетворяющие рассматриваемым условиям, которые задают расположения объектов на прямой, плоскости и в пространстве. Вначале заметим, что для пространства произвольной размерности выполнена

Теорема 1. *Если в пространстве размерности $t > 1$ есть правильный симплекс, состоящий из $t+1$ точки, то в пространство можно добавить только одну точку, отстоящую от других на равных расстояниях, и эта точка – центр симметрии симплекса.*

Доказательство. Действительно, точка должна отстоять ото всех вершин фигуры на одинаковых расстояниях, следовательно, находиться в центре описанной сферы (окружности в двумерном случае) данного симплекса. Из геометрии известно, что, если около фигуры можно описать сферу, то это можно сделать единственным образом. Также известно, что около правильного многогранника, которым является правильный симплекс, можно описать сферу. □

2.1. **Пространство размерности $t = 1$.** Так как максимальный симплекс в пространстве размерности $t = 1$ состоит из двух точек, то возможны только следующие случаи.

- **Максимальный симплекс в рассматриваемой конфигурации состоит из 2 точек.**
  - Два максимальных симплекса. На прямой это возможно только когда одна точка находится от других на расстояниях равных 1, то есть симплексы пересекаются. Тогда расстояние между оставшимися точками $M = 2$.

Действительно, если бы симплексы не пересекались, на прямой располагалось бы 4 точки. Их можно пронумеровать слева направо. Все расстояния от первой точки до остальных трех различны, что не удовлетворяют условию.
Отсюда также следует, что число точек, которое можно расположить на прямой не превышает 3.

– Один максимальный симплекс. Две точки находятся на расстоянии 1. К ним можно добавить только одну точку, находящуюся на одинаковых расстоянии от двух других. Это единственная точка, находящаяся в середине отрезка. В этом случае расстояние $M = \frac{1}{2}$.

- **Максимальный симплекс в рассматриваемой конфигурации состоит из 1 точки.** Все точки находятся на расстояниях равных M. Всего на прямой можно расположить только две таких точки.

2.2. **Пространство размерности $t = 2$.** Так как максимальный симплекс в пространстве размерности $t = 2$ состоит из 3 точек, то возможны только следующие случаи.

- **Максимальный симплекс в рассматриваемой конфигурации состоит из 3 точек.** Максимальный симплекс единственнен по теореме 1. К нему можно добавить единственную точку, находящуюся на одинаковых расстояниях от других точек. Это центр описанной около треугольника окружности, отстоящий на расстоянии $M = \frac{\sqrt{3}}{3}$ от вершин.
- **Максимальный симплекс в рассматриваемой конфигурации состоит из 2 точек.**

  – Два максимальных симплекса. Расположим две точки на расстоянии 1 друг от друга. Остальные точки находятся на одинаковых расстояниях от данных точек, то есть на серединном перпендикуляре отрезка, соединяющего эти точки. На нем можно расположить только 2 точки, отстоящие на расстоянии $M \neq 1$, симметрично отрезку-симплексу (если бы расстояние равнялось 1, то максимальный симплекс состоял бы из 3 точек). Добавление еще одной точки даст минимум одно дополнительное расстояние, что не будет согласовано с условием. Если эти две точки составляют симплекс, то из простых геометрических соображений можно получить, что $M = \frac{\sqrt{2}}{2}$.
  – Один максимальный симплекс. Точки, находящиеся на серединном перпендикуляре, находятся на расстоянии 1 от вершин первого симплекса и на расстоянии $M = \frac{\sqrt{3}}{3}$ друг от друга.

- **Максимальный симплекс в рассматриваемой конфигурации состоит из 1 точки.** Все точки находятся на расстоянии M. Всего на плоскости можно расположить три таких точки, находящихся в вершинах равностороннего треугольника.

2.3. **Пространство размерности** $t = 3$**.** Так как максимальный симплекс в пространстве размерности $t = 3$ состоит из 4 точек, то возможны только следующие случаи.

- **Максимальный симплекс в рассматриваемой конфигурации состоит из 4 точек.** Максимальный симплекс единственнен по теореме 1. К нему можно добавить единственную точку, отстоящую на одинаковых расстояниях от других точек. Это центр сферы, описанной около тетраэдра и отстоящий на расстоянии $M = \frac{\sqrt{6}}{4}$ от вершин.
- **Максимальный симплекс в рассматриваемой конфигурации состоит из 3 точек.** Три точки, находящиеся на расстоянии 1, образуют правильный треугольник. Остальные точки могут находиться только на прямой, проходящей через его центр и перпендикулярной плоскости треугольника. Поэтому можно добавить только две точки, симметричные относительно плоскости треугольника. Добавление еще одной точки даст минимум одно дополнительное расстояние, что не будет согласовано с условием. Эти точки могут находиться на расстоянии 1, тогда они находятся на расстоянии $M = \sqrt{\frac{7}{12}}$ от вершин первого симплекса. Или они могут находиться на расстоянии M друг от друга и от первого симплекса, тогда из геометрических соображений $M = \frac{2}{3}$.
- **Максимальный симплекс в рассматриваемой конфигурации состоит из 2 точек.** Две точки находятся на расстоянии 1 друг от друга. Остальные точки находятся в плоскости, перпендикулярной данному отрезку на окружности радиуса $\sqrt{M^2 - 1/4}$.
  Так как максимальная размерность симплекса в расссматриваемом случае равна 2, то в плоскость можно поместить дополнительный максимальный симплекс-отрезок. Это отрезок, чьи вершины лежат на окружности радиуса $\frac{1}{2}$, чей центр проходит через серидину первого симплекса. Второе расстояние определяется автоматически и равно $M = \frac{\sqrt{2}}{2}$. Более того, заметим, что, если ко второму максимальному симплексу добавить две точки, они будут лежать на той же окружности и автоматически окажутся на расстоянии 1 друг от друга, то есть составят еще один максимальный симплекс-отрезок.
  Если в плоскость не помещать симплекс-отрезок, то точки, которые можно добавить к первому симплеску, будут находиться на равных расстояниях друг от друга, не равных 1. Максимальное число точек, которое можно поместить таким образом на окружность, равно трем и при этом $M = \sqrt{\frac{7}{12}}$.
- **Максимальный симплекс в рассматриваемой конфигурации состоит из 1 точки.** Все точки находятся на расстоянии M. Всего в пространстве можно расположить четыре таких точки, находящихся в вершинах правильного тетраэдра.

3. Максимальное число объектов в евклидовом пространстве размерности t

Из рассмотренных случаев видно, что число объектов, которые можно поместить в пространство конечной размерности так, чтобы попарные расстояния задавались матрицей близости, удовлетворяющей рассматриваемым условиям, конечно, а расстояния M иногда могут быть не произвольными, а строго определенными. Для максимального числа объектов, помещенных в пространство конечной размерности, справедлива следующая теорема.

Теорема 2. *Число точек, чья матрица попарных расстояний состоит из недиагональных элементов, принимающих значения из множества* $\{1, M\}$ *и удовлетворяющих аксиомам метрики и условию порядка, которые можно поместить в пространство размерности* t *не превышает* $2t$.

Доказательство. Для случаев $t = 1, 2, 3$ утверждение проверено (см. выше). Пусть утверждение доказано для случая $k \leq t - 1$, докажем для случая $k = t$.

I. Пусть максимальная размерность единичного блока равна $d > 1$. Тогда построим в пространстве этот симплекс. Это можно сделать в пространстве размерности $d-1 \geq 1$. Тогда остальные точки находятся на одинаковом расстоянии от точек этого блока, то есть в гиперплоскости размерности $t - (d - 1)$.

По предположению индукции, в гиперплоскости размерности $t - (d - 1)$ число точек $k \leq 2(t-(d-1))$. Тогда общее число точек в пространстве размерности t будет $k \leq d + (2(t - (d - 1)) = 2t - d + 2 \leq 2t$ при $d \geq 2$.

II. Пусть максимальная размерность единичного блока равна $d = 1$. Это означает, что в пространство нужно поместить какое-то число пар объектов, находящихся на расстоянии друг от друга, равном 1, и какое-то число объектов, находящихся на расстоянии M от всех остальных объектов.

Построим конфигурацию, состоящую из пар точек, находящихся на расстоянии 1. Расположим первую пару произвольным образом. Следующие две точки должны находиться на одинаковом расстоянии от уже построенных точек, следовательно, они должны находиться в гиперплоскости, проходящей через середину отрезка, соединяющем первую пару точек. Точки должны располагаться симметрично относительно этого отрезка и находиться друг от друга на расстоянии 1. Заметим, что размерность пространства, в котором расположены точки, при таком построении увеличилась с 1 до 2. Далее добавляем еще две точки симметрично относительно построенной конфигурации так, чтобы расстояние между ними было равно 1, и отрезок, соединяющий их был перпендикулярен двум предыдущим отрезкам, то есть их содержащей плоскости. Это, очевидно, можно сделать в пространстве размерности 3. При добавлении нового отрезка, размерность пространства увеличивается на 1. Поэтому число точек, которое можно разместить в пространстве размерности t таким образом будет равно $2t$.

Это число будет максимальным. Действительно, добавим к получившейся конфигурации одну точку, равноотстоящую от других точек. Размерность пространства увеличилась на 1, число точек также увеличилось на 1. Следующую точку можно

добавить симметрично последней точке относительно конфигурации перпендикулярных отрезков, не увеличив при этом размерности пространства. Добавление третьей точки ведет к тому, что 3 точки будут находиться на одинаковых расстояниях равных M друг от друга. Эти точки находятся в вершинах правильного симплекса в пространстве размерности 2. Для его построения требуется 3 точки, в то время как конфигурация с перпендикулярными отрезками обеспечила бы в плоскости расположение 4-ех точек. Следовательно, данная конфигурация не будет оптимальной.

□

Итак, если в евклидовом пространстве размерности t задана конфигурация точек, попарные расстояния между точками которой принимают одно из двух значений, то число точек в этой конфигурации не превышает $2t$. Если данную конфигурацию точек нельзя поместить в пространство меньшей размерности, то число точек в этой конфигурации не может быть меньше $t+1$, что соответствует случаю t-мерного симплекса. Поэтому верно следующее утверждение.

Следствие 1. *Размерность пространства t, требуемая для размещения N произвольных точек, находится в интервале* $[\lceil N/2 \rceil, N-1]$.

Заключение

В работе найден общий вид матриц, для элементов которых выполняются аксиомы метрики и условие порядка, которые могут являться матрицами близости для объектов из евклидова пространства. В пространствах размерностей $t = 1, 2, 3$ описанны все возможные матрицы близости, удовлетворяющие рассматриваемым условиям. Доказана теорема о том, что в евклидово пространство размерности t, можно поместить $N \leq 2t$ объектов, матрица попарных расстояний которых удовлетворяет рассматриваемым условиям. Получены нижняя и верхняя оценки для размерности пространства t, требуемой для размещения N точек.

В дальнейшем исследовании предполагается более глубокое изучение структуры матриц попарных расстояний для размещения объектов в пространствах размерности $t \geq 4$ и нахождение условия, когда произвольные матрицы задают конфигурации в евклидовом пространстве. Также планируется найти новые критерии для размещения объектов в евклидовых пространствах по матрице попарных расстояний, удовлетворяющей рассматриваемым свойствам, при неточном соответствии элементов матрицы расстояниям между объектами.

Список литературы

1. *Дэйвисон М.* Многомерное шкалирование. — М.: Финансы и статистика, 1988.— 254 с.
2. *Hallard T. Croft, K. J. Falconer , Richard K. Guy* Unsolved problems in Geometry. — Springer-Verlang, 1991. — 198 p.

Статья поступила в редакцию 27.04.2008