УДК 519.25

Решение задачи многомерного анфолдинга для случая единственной цели

Дронов С.В., Леонгардт К.А.
Алтайский государственный университет
dsv@math.asu.ru

Аннотация

Задача многомерного анфолдинга (статистического развертывания, prefscal) представляет собой задачу расположения на одной карте одновременно двух множеств объектов – множества наблюдателей и множества целей по набору расстояний между каждым из наблюдателей и каждой из целей. Существующие сегодня приближенные методы ее решения предполагают, что в каждом из этих множеств не менее двух элементов. В работе предлагается метод точного решения этой задачи, когда цель имеется только одна. Алгоритмы легко переносятся на случай лишь одного наблюдателя.

1. Постановка задачи

Решение задач визуализации разного рода статистических данных сегодня весьма востребовано. Имеется довольно много разного рода алгоритмов, позволяющих изображать объекты наблюдения точками в пространстве небольшого количества измерений (чаще всего на плоскости). Имея в виду современные сферы применения методов обработки данных, в которых данные часто не имеют числового характера, но зато позволяют каким-то образом оценить степень различия изучаемых объектов, можно предполагать, что методы визуализации, использующие способ задания объектов в виде таблиц их попарных различий, будут приобретать все большую популярность.

Самым известным таким методом на сегодня, по-видимому, является метод многомерного шкалирования [1, 2]. При решении задач визуализации разного рода опросов в социологии [3], психологии [4] и т.п. применяется метод того же класса, но с неполными данными, который получил название метода многомерного анфолдинга.

Он предназначен для решения задачи в следующей постановке. Пусть имеется два множества объектов, одно из которых называют множеством целей, а другое – множеством наблюдателей. Каждый из наблюдателей сообщает расстояния от себя до каждой из целей. Информация о расстояниях между объектами, когда они оба принадлежат одному множеству, отсутствует. Требуется изобразить объекты обоих множеств на одной карте.

Имеются разные способы решения такой задачи. Например, [5] предлагает сводить эту задачу к задаче многомерного шкалирования путем заполнения пропущенных клеток в матрице попарных различий объектов разными способами, а в [6] описывается применение к проблеме анфолдинга двухэтапного метода наименьших квадратов. Тем не менее, при всех упомянутых подходах и во всех известных нам реализациях метода в статистических компьютерных пакетах требуется, чтобы каждое из двух множеств (наблюдателей и целей) состояло бы, по крайней мере, из двух элементов.

Предложим алгоритм, который позволит решать задачу, когда одно из этих множеств одноэлементно. Пусть у нас имеется одна цель и n наблюдателей, каждый из которых должен быть удален от цели на известное расстояние d_i, $i = 1, \ldots, n$. Без ограничения общности будем считать, что эти расстояния упорядочены по возрастанию, – точнее, неубыванию. Поскольку, как всегда в задачах визуализации по матрицам различий, нас интересует

лишь взаимное расположение объектов, можно поместить цель в начало координат. Требуется расположить наблюдателей так, чтобы все они располагались от цели на известных расстояниях.

Очевидно, без дополнительных ограничений задача имеет, вообще говоря, бесконечно много решений. Поэтому потребуем, чтобы точки, изображающие наблюдателей, были бы наиболее сильно разбросаны, точнее, чтобы минимальное из попарных расстояний между ними было бы максимально возможным. Такое ограничение обычно объясняют необходимостью изобразить наши объекты с наибольшей наглядностью. Можно дать этому и иное объяснение, – если целью обработки данных является, например, обнаружение всех наблюдателей на реальной местности, когда расположение цели известно, то предлагаемый подход описывает "наиболее неприятную" ситуацию, что позволяет оценить сверху время и затраты на их поиск.

2. Одномерное решение

Сначала предположим, что нам нужно разместить точки, изображающие наблюдателей, на прямой линии, где в точке с координатой 0 размещается цель. Оказывается, решение может быть получено с помощью следующего простого алгоритма:

Алгоритм T1-line. Разместим наблюдателей в точках с координатами d_i, $i = 1, \ldots, n$, а затем все точки с четными номерами отразим симметрично относительно начала координат.

Все наблюдатели, таким образом, будут расположены в точках с координатами

$$(-1)^{i-1} d_i, \quad i = 1, \ldots, n. \tag{1}$$

Отметим, что здесь имеется два симметричных относительно начала координат полностью равноценных решения.

Теорема 1. *Алгоритм T1-line дает расположение точек-наблюдателей Ob_i, $i = 1, \ldots, n$ на прямой, максимизирующее минимальное расстояние между ними.*

Доказательство. Доказательство теоремы проведем индукцией по числу наблюдателей n. При n =2 с точностью до симметрии имеется два возможных расположения наблюдателей, – по одну сторону от начала координат и по разные стороны. Ясно, что максимальное значение расстояния между ними (оно же и минимальное, а значит максимальное из возможных минимальных в этом случае) достигается при втором расположении. Таких расположений два, причем расстояние между точками в них одно и то же. Поэтому можно выбрать то из них, которое удовлетворяет (1).

Предположим, что для k наблюдателей алгоритм обоснован. Это означает, что среди всех возможных расположений k точек на прямой расположение (1) обладает максимальным минимальным расстоянием между точками. Возьмем произвольное расположение $k+1$ точки, назовем его текущим. При этом нам известно, что $(k+1)$-я точка обладает наибольшим расстоянием от начала координат. Следовательно, как бы мы не располагали точки, она будет располагаться вне отрезка, образованного точками Ob_k, Ob_{k-1}, а остальные – внутри этого отрезка. Поэтому отрезок минимальной длины можно искать в два этапа – сначала найти минимальный отрезок в облаке точек с номерами $1, \ldots, k$, а затем сравнить его с минимальным из расстояний $|Ob_{k+1}Ob_j|$, $j = 1, \ldots, k$. Понятно, что последний минимум должен выбираться из двух расстояний – до k-й и до $(k-1)$-й точки. Если разместить $(k+1)$-ю точку по ту же сторону от начала координат, что и k-ю, то этот минимум будет равен $d_{k+1} - d_k$, иначе $d_{k+1} - d_{k-1}$. Но, в силу того, что расстояния упорядочены, первое по крайней мере не больше, а значит, располагая новую точку

с противоположной стороны от начала координат, чем была расположена k-я точка, мы получаем возможность увеличить длину минимального отрезка.

Теперь изменим расположение первых k точек, расположив их согласно (1). По индукционному предположению в этом случае минимальная длина отрезка между ними заведомо не меньше, чем в текущем их расположении. Если при этом k-я точка окажется на том же месте, что и в текущем расположении, то в новой конструкции минимальная длина отрезка не меньше, чем в текущем, а координаты всех $(k+1)$ точек удовлетворяют (1). Если же k-я точка в новом расположении оказалась в положении, противоположному текущему, то отразим $(k+1)$-ю точку симметрично относительно начала координат, и вновь получим расположение, заведомо не хуже, чем текущее, координаты которого удовлетворяют (1).

Поскольку текущее расположение было произвольным, то теорема доказана. □

3. Визуализация в пространстве большего числа измерений

Начнем с задачи расположения точек Ob_i, $i = 1, \ldots, n$ на плоскости, т.е. определения для каждой из них пары координат. Критерием оптимальности их расположения по-прежнему является максимальность минимального расстояния между любыми двумя из них. Цель расположена в начале координат.

Предположим, что на плоскости заданы произвольные k точек $A_1, \ldots, A_k$. Для фиксированного $d > 0$ назовем системой d-окружностей множество окружностей равных радиусов d с центрами в этих точках. Следующее утверждение представляется очевидным.

Лемма 1. *Геометрическое место точек, для которых минимальное расстояние до заданных k точек равно d, есть объединение тех частей семейства d-окружностей, которые не лежат внутри ни одной из них.*

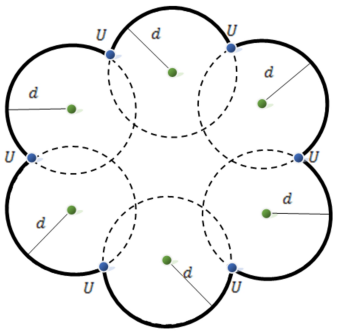

Рисунок 1. Пример d-оболочки для шести точек

Такое геометрическое место (пример его приведен на рисунке 1) условимся далее называть d-оболочкой точек $A_1, \ldots, A_k$. Те точки d-оболочки, которые принадлежат сразу двум из системы d-окружностей, будем называть углубленными точками, – смысл названия ясен из рисунка, где углубленные точки обозначены буквами U.

Перейдем к идее алгоритма. Пусть несколько точек Ob_i, $i = 1, \ldots, k$ уже построены. Опишем окружность с центром в начале координат радиуса d_{k+1} (“большая окружность”). Согласно сделанному ранее предположению, все построенные точки лежат внутри нее, а новую точку надо искать на этой окружности. Взяв d-оболочку уже построенных точек при достаточно малом d, начнем увеличивать это число до тех пор, пока у d-оболочки и “большой окружности” появятся общие точки. Затем продолжим его увеличивать, пока множество общих точек остается непустым. В качестве Ob_{k+1} выберем любую из общих точек при максимально возможном d.

Отметим, что из предложенной идеи вытекает, что каждая из точек Ob_i, $i = 1, \ldots, n$ в итоге будет выбираться из множества углубленных точек d-оболочки точек с меньшими номерами при d, подбираемого на каждом следующем шаге. Поскольку каждая углубленная точка лежит на срединном перпендикуляре, соединяющем центры некоторых двух из системы d-окружностей, то высказанная идея может быть реализована так.

Алгоритм T1-plane.

• Шаг 0. Точку Ob_1 разместим на оси абсцисс, в точке $(d_1, 0)$. К шагу 1.

• Шаг 1. Пусть k точек уже построены. Опишем "большую окружность" радиуса d_{k+1} с центром в начале координат. Найдем выпуклый многоугольник с вершинами в уже построенных точках такой, что он содержит их все. Восстановим срединные перпендикуляры к каждой из сторон этого многоугольника до пересечения с "большой окружностью". Ту из точек пересечения, длина отрезка срединного перпендикуляра у которой окажется наибольшей, объявим точкой Ob_{k+1}. Если таких точек несколько, выберем произвольную из них. К шагу 2.

• Шаг 2. Все ли точки построены? Если нет, к шагу 1. Иначе конец алгоритма.

Следующее утверждение немедленно вытекает из леммы 1.

Теорема 2. *Алгоритм T1-plane дает расположение точек-наблюдателей на плоскости, максимизирующее минимальное расстояние между ними.*

Нетрудно заметить, что алгоритм T1-line является следствием и частным случаем алгоритма T1-plane, когда все "большие окружности" вырождаются в две точки, лежащие на оси абсцисс и удаленные от начала координат на расстояния d_i, $i = 1, \ldots, n$. Ясно также, что алгоритм может быть обобщен на расположение наблюдателей в пространстве любой размерности, только вместо системы окружностей должны использоваться гиперсферы соответствующих размерностей, а срединные перпендикуляры следует восстанавливать в серединах ребер выпуклого многогранника, содержащего в себе все облако предварительно построенных точек. Назовем эту процедуру алгоритмом T1. Отсюда немедленно получается

Следствие 1. *При увеличении размерности визуализирующего пространства максимальное из минимальных расстояний между построенными с помощью алгоритма T1 точками, не уменьшается.*

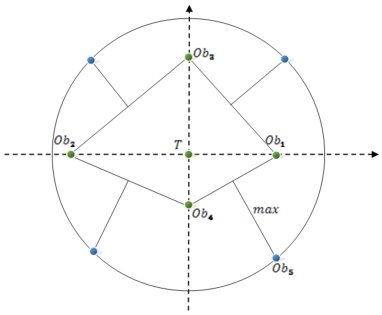

Рисунок 2. Иллюстрация работы алгоритма T1-plane

Пример работы алгоритма T1-plane при построении пятой точки приведен на рисунке 2. Конечно же, вторая точка при этом располагается на оси абсцисс по другую сторону от начала координат, чем первая, третья лежит на срединном перпендикуляре между первой и второй, четвертая – между ними же, но по другую сторону от оси.

Список литературы

1. Дейвисон М. Многомерное шкалирование: методы наглядного представления данных. — М. : Финансы и статистика, 1988.

2. Дронов С.В. Методы и задачи многомерной статистики. — Барнаул : Изд-во Алт. ун-та, 2015.

3. Price R.H., Bouffard D.L. Behavioral appropriateness and situational constraints as dimensions of social behavior // Journal of Personality and Social Psychology. — 1974. — Vol. 30. — P. 579–586.

4. Nesselroade John R., Cattell Raymond B. Handbook of Multivariate Experimental Psychology. — Springer Science & Business Media, 2013.

5. Jan de Leeuv. Multidimensional Unfolding // The Encyclopedia of Statistics in Behavioral Science. — Wiley, 2005.

6. Heiser W.J. Joint ordination of species and sites: The unfolding technique // Developments in numerical ecology / Ed. by P. Legendre and L. Legendre. — Berlin, Heidelberg : Springer-Verlag, 1987. — P. 189–221.