О геометрических подходах к решению некоторых задач кластеризации данных

Игорь Валерьевич Зиновеев
http://orcid.org/0000-0002-7392-2327

Аннотация Рассматривается геометрический подход к упорядочиванию данных при решении задач кластерного анализа. В ходе решения задачи упорядочивания данные интерпретируются как множество точек декартовой плоскости с введенной Евклидовой метрикой.

Ключевые слова Кластерный анализ, упорядочивание данных, геометрический подход, Евклидова метрика.

УДК 51-74: 519.65

1 Введение

Достаточно часто в различных областях науки для понимания сути явлений, разработки объясняющих их принципов, возникает необходимость проведения классификации объектов, для чего необходимо их предварительно упорядочить, сгруппировать. Вопросами разработки типологии или классификации, исследования концептуальных схем группировки объектов, формирования гипотез на основе исследования данных занимается кластерный анализ. При этом для выделения кластеров, т. е. групп близких по некоторому критерию объектов, в кластерном анализе используются различные подходы и методы, например методы линейной и нелинейной регрессии, метод деревьев решений, метод нейронных сетей и т. д. (на рис. 1 приведены примеры разбиений множества объектов базы данных в двухмерной интерпретации на кластеры).

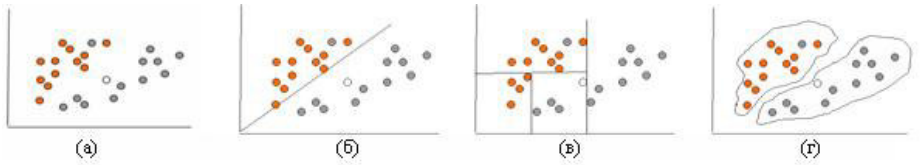


Рис. 1 Примеры кластеризации множества объектов базы данных в двухмерной интерпретации: а) множество объектов базы данных; б) классификаця методом линейной регрессии; в) классификация методом деревьев решений; г) классификация методом нейронных сетей.

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы: отбор выборки количественных данных для кластеризации; определение признакового пространства; вычисление значений меры сходства (различия) между объектами; применение одного или нескольких методов кластерного анализа (достаточно полный обзор методов кластеризации приведен, например, в работах A. Jain, M. Murty, P. Flynn [1], Берикова В. С [2]) для создания групп сходных объектов; проверка достоверности результатов кластерного решения.

Во многих задачах кластерного анализа помимо необходимости сгруппировать данные возникает еще задача эти данные упорядочить, ввести на этом множестве отношение порядка. Данная статья посвящена изучению вопроса введения на заданном множестве сгруппированных данных отношения порядка с помощью геометрического подхода.

2 Постановка задачи

Рассмотрим задачу анализа исходного конечного набора данных, каждый объект которого задается двумя числовыми характеристками (например, географическое расположение хостов компьтерной сети). Понятно, что в этом случае можем рассмотреть геометрическую интерпретацию исходных данных, представив их как набор геометрических точек, заданных своими декартовыми координатами (числовыми характеристиками) на Евклидовой плоскости с введенной Евклидовой метрикой (для произвольных точек A и B заданного множества в качестве метрики $d(A, B)$ выбирается евклидово расстояния между ними). Тогда отношение порядка на заданном множестве точек может быть сформулировано следующим образом: для заданного конечного набора n точек $(B_1,\ B_2,\ ...,B_n)$ требуется найти такую перестановку $(B'_1,\ B'_2,\ ...,B'_n)$ этих точек, чтобы замкнутая ломаная $B'_1\,B'_2\ ...\ B'_n\ B'_1$ имела минимальную длину, т. е. была минимальной сумма

$$S = d\,(B_1,\ B_2) + d\,(B_2,\ B_3) + \ldots d\,(B_{n-1},\ B_n) + d\,(B_n,\ B_1)\,, \qquad (1)$$

где *d(A, B)* – евклидово расстояние между точками *A* и *B*. Порядок следования точек в найденной перестановке определит порядок данных в заданном кластере и может определить, например, архитектуру компьютерной сети с наименьшим (в локальном или глобальном понимании) временем прохождением сигнала по сети.

Заметим, что в терминах теории взвешенных графов (вершины – заданные точки, веса ребер – евклидово расстояние между соответствующими вершинами) задача может быть сформулирована как задача определения порядка вершин гамильтонова цикла наименьшего суммарного веса.

Таким образом, рассматриваются две задачи: задача введения порядка на конечном множестве данных и задача введения порядка на множестве кортежей данных $(B'_1,\ B'_2,\ ..., B'_n)$, ... , $(B''_1,\ B''_2,\ ..., B''_n)$ по признаку минимальности суммы (1).

3 Обсуждение и результаты

Решение задачи упорядочивания множества данных в кластере.

Воспользуемся подходами, описанными в работах [3], [4], в которых рассматриваются принципы разработки алгоритмов построения приближённых решений задачи комивояжера. Эти принципы основаны на применений аналогов конечноэлементных аппроксимаций уравнений одномерных краевых задач при построении кусочно гладких кривых, аппроксимирующих в пространстве высокой размерности путь коммивояжера.

Для нахождения требуемой перестановки $(B'_1,\ B'_2,\ ..., B'_n)$ строится гладкая вспомогательная кривая. Для некоторой перестановки $(B_1,\ B_2,\ ..., B_n)$ заданного набора точек на плоскости Oxy будем рассматривать функционал от *2n* переменных, представляющий собой длину замкнутой ломаной, определяемой заданным набором точек:

$$F(x_1, ..x_n, y_1, ..y_n) = \sum_{i=1}^{n-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} + \\ + \sqrt{(x_1 - x_n)^2 + (y_1 - y_n)^2}. \qquad (2)$$

Задача о нахождении замкнутой ломаной минимальной длины сводится к задаче минимизации значения этого функционала на фиксированном наборе упорядоченных пар чисел – координат вершин ломаной.

Применяя к (2) необходимое условием локального екстремума функции многих переменных

$$\frac{\partial F\left(x_1,\ \ldots,\ x_n,\ y_1,\ \ldots,\ y_n\right)}{\partial x_i}=0,\ \frac{\partial F\left(x_1,\ \ldots,\ x_n,\ y_1,\ \ldots,\ y_n\right)}{\partial y_i}=0,\quad i=1,...,n,$$

получаем однородную систему $2n$ квазилинейных алгебраических уравнений с $2n$ неизвестными с ленточной матрицей $A_{n\times n}$ (лента шириной три элемента), элементы которой определяются соотношениями (3) :

$$A_{kk}=\frac{1}{\sqrt{(x_{k-1}-x_k)^2+(y_{k-1}-y_k)^2}}+\frac{1}{\sqrt{(x_{k+1}-x_k)^2+(y_{k+1}-y_k)^2}},$$

$$A_{kk\pm1}=\frac{-1}{\sqrt{(x_{(k\pm1}-x_k)^2+(y_{k\pm1}-y_k)^2}}. \tag{3}$$

В работе [4] приведено обоснование того, что ленточная матрица, элементы которой определяются выражениями (3), соответствует системе квазилинейных алгебраических уравнений, которая может быть рассмотрена как численная аппроксимация оператора краевой задачи на неравномерной сетке:

$$\begin{cases} -\tilde{X}''(t)=0, \\ -\tilde{Y}''(t)=0, \end{cases}\ \tilde{X}(0)=\tilde{X}(2\pi),\tilde{Y}(0)=\tilde{Y}(2\pi)\ , \tag{4}$$

где $\tilde{X}''\left(t\right)$ и $\tilde{Y}''\left(t\right)$ являются некоторыми гладкими периодическими функциями, которые в узлах сетки аппроксимации принимают заданные значения в порядке их следования.

Воспользуемся этим подходом для нашей задачи, считая узлами аппроксимации заданные точки, а порядком следования порядок точек в наборе $(B_1,\ B_2,\ ...,B_n)$.

Будем искать $\tilde{X}\left(t\right)$ и $\tilde{Y}\left(t\right)$ в виде

$$\tilde{X}(t)=\sum_{k=1}^{n}\alpha_k v_k(t),\qquad \tilde{Y}(t)=\sum_{k=1}^{n}\beta_k v_k(t), \tag{5}$$

где $v_k(t)$ выбираются в виде некоторой периодической функции с периодом 2π, например, в виде $v_k(t)=e^{ikt}=\cos(kt)+i\cdot\sin(kt)$. Тогда минимум левых частей получившихся в (4) выражений достигается на некотором наборе коэффициентов $\left(\alpha_k,\ \beta_k\right),\ k=\overline{0,n}$, который можно сопоставить с одним из возможных вариантов замкнутой ломаной, при этом длина этой ломаной существенно зависит от скорости убывания этих коэффициентов.

Таким образом, требуется найти функции $\tilde{X}(t)$ и $\tilde{Y}(t)$ в виде (5) такие, что в узлах сетки аппроксимации они принимают значения координат заданных точек, причём порядок, в котором точки будут пройдены, заранее не известен, и коэффициенты из (5) должны обеспечивать минимум левым частям выражений в системе (4).

На основе приведённых рассуждений строится итеррационный процесс решения поставленной задачи. Предположим, что на некотором шаге вычислены k коэффициентов для (5) и известен некоторый порядок обхода точек $(B'_1,\ B'_2,\ ...,B'_n)$. На следующем шаге итерации требуется скорректировать этот порядок, т. е. определить новый порядок $(B''_1,\ B''_2,\ ...,B''_n)$, уточнить коэффициенты с учётом нового порядка обхода и вычислить $(k+1)$-е коэффициенты в (5), при этом значение функции (2) на кортеже $(B''_1,\ B''_2,\ ...,B''_n)$ должно быть не больше, чем на кортеже $(B'_1,\ B'_2,\ ...,B'_n)$.

Назовём вспомогательной кривой упорядочивания (ВКУ) параметризованную кривую, координаты точек которой – это значения функций $\tilde{X}(t)$ и $\tilde{Y}(t)$. Опустим из заданных точек перпендикуляры на ВКУ и для каждой точки выберем наименьший перпендикуляр. Основания, полученных таким образом перпендикуляров, однозначно сопоставляются заданным точкам, поэтому порядок их следования по ВКУ задаёт новый порядок обхода точек. Этим обеспечивается решение первой из поставленных задач – введение отношения порядка на заданном множестве данных.

Заметим, что описанный подход может быть применен к задачам проектирования архитектур компьютерных сетей. Если на множестве компьютеров разрабатываемой сети провести разбиение на кластеры, то порядок, введенный на множестве объектов (компьютеров) каждого полученного кластера может служить основой для принятия решения о топологии (архитектуре) компьютерной сети. В частности, описанный далее подход позволяет использовать порядок данных для конструирования сетей кольцевой и шинной топологий.

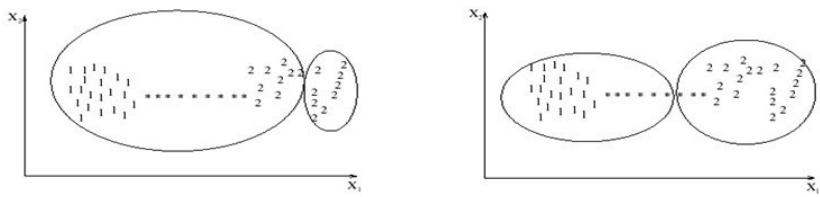
Построение вспомогательной кривой упорядочивания данных.

Рассмотрим разности между координатами заданных точек кластера и координатами их проекций на ВКУ. По ним с помощью метода линейной интерполяции можно построить функции невязки $\Delta X(t)$ и $\Delta Y(t)$, которые в узлах сетки аппроксимации характеризуют, насколько значения функций $\tilde{X}(t)$ и $\tilde{Y}(t)$ (координаты точек-проекций) отличаются от соответствующих координат исходных точек. Используя стандартные процедуры численного анализа (например метода быстрого преобразования Фурье), для функций

$\Delta X(t)$ и $\Delta Y(t)$ могут быть получены уточнения для уже вычисленных коэффициентов и найдены $(k+1)$-е коэффициенты.

Таким образом, определив ВКУ, соответствующую найденному набору коэффициентов, определяется порядок следования точек и строится соответствующая ломаная.

В зависимости от вида расположения данных в качестве ВКУ может быть подобрана соответствующая замкнутая линия, удовлетворяющая описанным выше условиям, однако в большинстве случаев в качестве ВКУ можно выбирать линию второго порядка – эллипс, например на рис.2 приведены два варианта такого представления.



а) б)

Рис. 2 Примеры кластеризации данных.

Пусть в качестве ВКУ выбран эллипс, заданный параметрическими уравнениями $x = p\cos(t)$, $y = q\sin(t)$, что соответствует первым слагаемым в (5) (тем самым будут определены первые два коэффициента из (5)), параметры p и q которого необходимо определить. Положим, что одна из осей этого эллипса лежит на прямой, обладающей тем свойством, что сумма квадратов расстояний от заданных точек до этой прямой минимальна, а координаты центра эллипса являются средними арифметическими координат проекций точек на эту прямую. Тогда задача построения эллипса сводится к построению указанной прямой и нахождению полуосей эллипса.

Будем искать уравнение этой прямой в виде $\mathrm{ax} + \mathrm{by} + 1 = 0$. Значения коэффициентов a и b найдем из условия минимума суммы квадратов расстояний от заданных точек до прямой

$$S(a,b) = \frac{\sum_{k=o}^{n-1}(ax_k + by_k + 1)^2}{a^2 + b^2} \longrightarrow \min, \tag{6}$$

то есть, как решение системы уравнений $\frac{\partial S}{\partial a} = 0$, $\frac{\partial S}{\partial b} = 0$

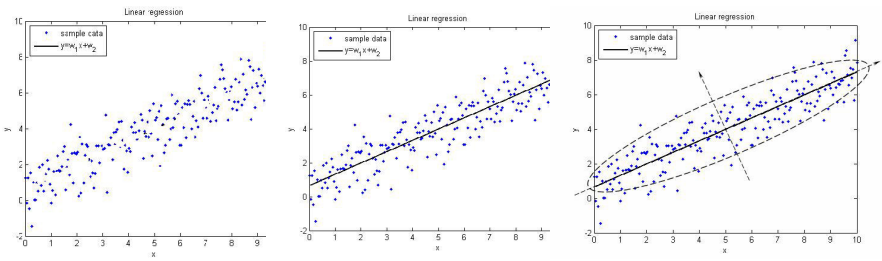
$$\frac{\partial S}{\partial a} = \frac{2(b(b^2 - a^2)\overline{XY} + ab^2(\bar{X}_2 - \bar{Y}_2) - 2ab\bar{Y} + (b^2 - a^2)\bar{X} - an)}{a^2 + b^2} = 0,$$

$$\frac{\partial S}{\partial b} = \frac{2(a(a^2-b^2)\overline{XY} + ba^2(\bar{Y}_2 - \bar{X}_2) - 2ab\bar{X} + (a^2-b^2)\bar{Y} - bn)}{a^2+b^2} = 0, \quad (7)$$

где $\overline{XY} = \sum_{k=0}^{n-1} x_k y_k$, $\bar{X} = \sum_{k=0}^{n-1} x_k$, $\bar{Y} = \sum_{k=0}^{n-1} y_k$, $\bar{X}_2 = \sum_{k=0}^{n-1} x_k^2$, $\bar{Y}_2 = \sum_{k=0}^{n-1} y_k^2$.

Значение b находится из уравнения $b \cdot \left(a_1 \cdot b^2 + a_2 \cdot b + a_3\right) = 0$ – следствия системы (7). Здесь $a_1 = \overline{XY}\left(1 - \left(\frac{\bar{Y}}{\bar{X}}\right)^2\right) - \frac{\bar{Y}}{\bar{X}}(\bar{X}_2 - \bar{Y}_2)$, $a_2 = -\frac{2\bar{Y}^2}{\bar{X}} + \bar{X}\left(1 - \left(\frac{\bar{Y}}{\bar{X}}\right)^2\right) - \frac{2n\bar{Y}\cdot\overline{XY}}{\bar{X}^2} - \frac{n(\bar{X}_2 - \bar{Y}_2)}{\bar{X}}$, $a_3 = \frac{n\bar{Y}}{\bar{X}} - \frac{n^2\overline{XY}}{\bar{X}^2}$. Выбор необходимого корня b делается на основании сравнения значений $S(a,b)$. для каждой пары (a,b).

Альтернативным вариантом выбора этой прямой может быть прямая, полученная в результате линейного регрессионного анализа (рис.3). Центром эллипса выбирается точка этой прямой, координаты которой являются средними арифметическими координат проекций заданных точек на эту прямую.



а) б) в)

Рис. 3 Изображение ВКУ, действительная ось которой – прямая линейной регрессии.

Для нахождения полуосей p и q эллипса можно применить один из методов градиентного спуска в соответствии с такой итеррационной схемой:

$$p_{i+1} = p_i - \alpha\frac{L(p_i+\Delta, q_i) - L(p_i,q_i)}{\Delta}, q_{i+1} = q_i - \alpha\frac{L(p_i, q_i+\Delta) - L(p_i,q_i)}{\Delta}, \quad i = 0, ..., m,$$

где p_0, q_0, α, Δ, m - рабочие параметры, которые подбираются для каждого кластера отдельно, $L(p,q)$ – сумма квадратов расстояний от точек до эллипса с полуосями pи q.

Заметим, что если, отталкиваясь от некоторой начальной ВКУ (например, эллипса), сразу вычислить (подобрать) все коэффициенты вспомогательной кривой (5), то на следующем этапе итеррационного процесса при

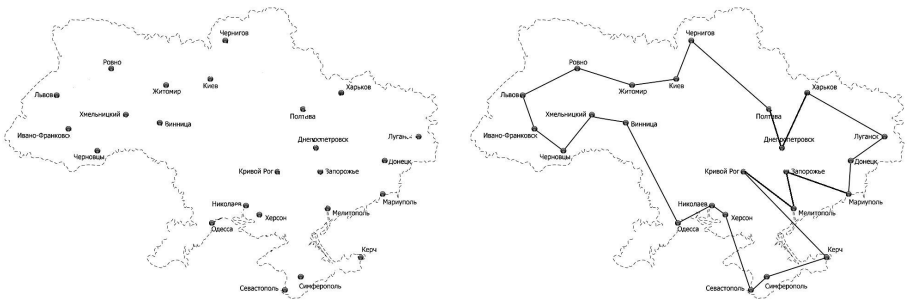
помощи некоторой вспомогательной функции, определенной на дискретном множестве точек контура ВКУ–эллипса, строится новая кривая, которая может быть построена по правилу: точки ВКУ, в которых эта вспомогательная функция положительна, перемещаются вдоль соотвествующих им внешних нормалей к ВКУ, а точки, в которых функция отрицательная, перемещаются вдоль соотвествующих им внутренних нормалей к ВКУ. Новая ВКУ определит новый порядок данных и дальнейшее преобразование, при этом значение функции (2) не будет увеличиваться, что обеспечит решение второй задачи.

Пример практического применения подхода

Для реализации описанного подхода была разработана компьютерная программа на языке системы аналитических вычислений MAPLE. В качестве тестовой была рассмотрена задача упорядочивания данных – городов Украины, заданных своими географическими координатами (рис.4а). В качестве признака формирования кластера выступил признак расположения в городе представительства одной из страховых компаний. В результате в кластер попало 24 города, алфавитный список и координаты которых приведены в таблице (данные взяты на сайте http://www.city-maps.ru/city-coords.php).

№	Город	Широта/Долгота	№	Город	Широта/Долгота
1	Chernigov	51.50N / 31.30E	**13**	Lvov	49.83N / 24.00E
2	Chernovtsy	48.32N / 25.87E	**14**	Mariupol	47.08N / 37.57E
3	Dnepropetrovsk	48.48N / 35.00E	**15**	Melitopol	46.85N / 35.37E
4	Donetsk	48.00N / 37.83E	**16**	Nikolayev	46.95N / 32.00E
5	Ivano-Frankovsk	48.92N / 24.70E	**17**	Odessa	46.50N / 30.77E
6	Kerch	45.37N / 36.45E	**18**	Poltava	49.58N / 34.58E
7	Kharkov	50.00N / 36.25E	**19**	Rovno	50.65N / 26.17E
8	Kherson	46.65N / 32.63E	**20**	Sevastopol	44.60N / 33.52E
9	Khmelnitskiy	49.42N / 26.82E	**21**	Simferopol	44.95N / 34.08E
10	Kiyev	50.47N / 30.48E	**22**	Vinnitsa	49.18N / 28.50E
11	Krivoy Rog	47.92N / 33.40E	**23**	Zaporozhye	47.83N / 35.17E
12	Lugansk	48.58N / 39.33E	**24**	Zhitomir	50.30N / 28.67E

В качестве начального кортежа взят набор, соответствующий порядку городов в таблице (1,2,3,...,22,23,24). В результате применения разработанной программы расчета получен порядок следования, который на рисунке 4(б) проиллюстрирован ломаной.



а) б)

Рис. 4 Иллюстрация результатов применения алгоритма упорядочивания.

Подводя итог, отметим, что алгоритм решения задачи упорядочивания данных кластера состоит в такой последовательности действий: задаются координаты исходных точек; выбирается начальная последовательность следования всех точек и задается соответствующий кортеж; строится ВКУ – эллипс; находятся проекции заданных точек на ВКУ и определяется порядок их следования вдоль ВКУ; в соответствии с порядком проекций определяется перестановка исходных точек и новый кортеж, определяющий порядок данных в кластере.

Выводы

В данной статье рассмотрен геометрический подход к задаче введения на заданном множестве сгруппированных данных отношения порядка, введенного по принципу следования проекций геометрических точек, аналогов этих данных, на вспомогательную кривую упорядочивания. Дальнейшим направлением исследований может быть исследование в направлении определения архитектур компьютерных сетей с некоторыми заданными оптимизационными характеристиками.

Список литературы

1. Jain A.K. Data clustering: a review [Text] / A.K. Jain, M.N. Murty, P.J. Flynn // ACM Computing Surveys/ - Vol. 31. - No. 3. - P. 264-323.
2. Бериков В.С. Современные тенденции в кластерном анализе [Текст] / В.С. Бериков, Г.С. Лбов // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы». - 2008. — 26 с.
3. Данилова О.Т. Приближенный глобальный алгоритм решения задачи коммивояжера с неравенством треугольника [Текст] / О.Т. Данилова, Р.Т. Файзуллин // Матема-

тические структуры и моделирование : [сб. науч. тр.] / Ом. гос. ун-т. - Омск, 2003. - Вып. 11. - С. 5-9.
4. Файзуллин Р.Т. Гладкие приближения в задаче коммивояжёра [Текст] / Р.Т. Файзуллин, Р.Р. Файзуллин // Таврический вестник информатики и математики. - 2004. - №27. - С.72-76.

Игорь Валерьевич Зиновеев
http://orcid.org/0000-0002-7392-2327
Запорожский национальный университет,Запорожье, Украина
E-mail: zinoveyev@mail.ru

Igor V. Zinoveyev

On the geometrical approaches to the solving of the data clusterization problems

The geometrical approach to the ordering of the data in solving cluster analysis problems has been considered. In the course of the solving the problem of the organizing the data are interpreted as a set of the points in a Cartesian plane with the Euclidean metric.

Keywords Cluster analysis, ordering of the data, geometrical approach, Euclidean metric