УДК 519.254

А. Н. Четырбоцкий, д-р физ.-мат. наук, вед. науч. сотр., e-mail: Chetyrbotsky@yandex.ru
Дальневосточный геологический институт ДВО РАН, г. Владивосток
Дальневосточный федеральный университет, г. Владивосток

Автоматическая классификация объектов многомерной выборки рекурсивными методами построения кривых Пеано (на примере выборки Р. Фишера)

На основании рекурсивных методов построения кривых Пеано разработан алгоритм автоматической классификации объектов многомерной выборки. Сущность разработки состоит в согласовании между собой соответствующих распределений образов объектов на отрезке вещественной оси. Для оценки работоспособности алгоритма используется общедоступная для использования выборка 150 четырехмерных наблюдений трех разновидностей цветков ириса (выборка Р. Фишера). Приведены результаты работы алгоритма и выполнен их анализ.

Ключевые слова: *многомерная выборка, кластерный анализ, заполняющие пространство кривые*

Введение

Многообразие современных явлений и систем в различных дисциплинах актуализирует проблемы их автоматической количественной систематизации (без вмешательства человека и отсутствии обучающих выборок) многомерных объектов. Такая ситуация обусловлена гигантским ростом объемов разнородной эмпирической информации. Для их согласования и углубленной обработки требуются значительные вычислительные ресурсы. Так, в некоторых задачах астрофизики число объектов (планет, метеоритов и других небесных тел) может превышать 10^{10} соответствующих единиц [1]. Значимость результатов классификации здесь обусловлена еще и тем, что ассоциации звезд в нашей Галактике обычно незаметны на фотографиях. Они выделяются из фона лишь как сгущения/кластеры звезд определенного типа [2]. В этой ситуации под кластером/классом объектов обычно понимают изолированную в исходном признаковом пространстве компактную совокупность объектов выборки [3, 4].

Эффективный инструментарий анализа выборочного материала основан на применении соответствующих так называемых кривых Пеано, под которыми понимается любое непрерывное отображение числового отрезка на плоский квадрат [4] (в связи с ее другим определением как непрерывного образа отрезка, полностью заполняющего квадрат, в работе [5] соответствующая кривая именуется "заполняющая пространство кривая" или ЗПК). Такая кривая была впервые построена Дж. Пеано [6], а существенно более простой ее вариант — Дж. Гилбертом [7]. Р. Г. Стронгин разработал оригинальный алгоритм ее построения [8], а В. В. Александров и Н. Д. Горский выполнили авторскую компьютерную реализацию построения [9]. Кривые Пеано имеют широкую область практического применения: распознавание образов [6, 10]; поиск глобального оптимума многоэкстремальных функций [11]; работа с базами данных [12] и т. д. Интересно заметить, отмеченный в работе [13], такой факт: согласно современным исследованиям, ДНК заполняет каждую клетку таким образом, что ее пространственная конструкция приближается к структуре кривой Пеано.

Широкие возможности кривых Пеано для решения многообразных практических задач в меньшей степени адаптированы для решения задач автоматической классификации. Целью работы является разработка алгоритма решения этой задачи на основании рекурсивных методов построения кривых Пеано. На примере общедоступной выборки Р. Фишера [14—16] разработан алгоритм автоматического выделения в признаковом пространстве многомерной выборки сравнительно таких кластеров ее объектов. Сущность предлагаемого алгоритма состоит в согласовании между собой результатов нелинейных отображений выборочных точек признакового пространства (объектов выборки) на отрезок вещественной оси с помощью рекурсивного построения кривых Пеано.

Выборка данных

Поскольку выборка Р. Фишера уже стала традиционным полигоном для разработок и тестирования различных методов анализа данных (в системе MATLAB эта выборка представлена набором iris.dat, а демопрограмма их анализа — модулем irisfcm.m [16]), то представляется полезным детальное описание ее структуры.

В работе [14] приведены результаты линейного дискриминантного анализа выборки, объектами которой выступают 150 четырехмерных наблюдений трех разновидностей цветков ириса (*Setosa*, *Versicolor* и *Virginica*), где каждый из них представлен 50 наблюдениями. Отдельное наблюдение ха-

рактеризуется четырьмя количественными признаками: длиной X_1 и шириной X_2 чашелистиков, длиной X_3 и шириной X_4 лепестков. Дескриптивные статистики, матрица корреляции признаков, коэффициенты корреляции переменных и главных компонент (ГК) на переменные и их значимость (графа %) приведены в табл. 1 (вычисления ГК в MATLAB выполняет программа princomp.m).

Наименьшая вариабельность X_2 отражается в ее низком влиянии на кластерную структуру выборки. Для X_3 имеет место обратная ситуация, иллюстрацией чему являются распределения на рис. 1 (см. третью сторону обложки).

Распределение выборочных точек на рис. 1, *а* показывает, что в пространстве $\{X_1, X_2, X_4\}$ имеет место значительное перекрытие выборочных объектов 2-го и 3-го видов ириса. На рис. 1, *б* оно практически отсутствует. В обоих случаях объекты 1-го вида образуют их компактное изолированное сгущение.

Вследствие высокой вариации X_3 она вносит наибольший вклад в неравномерность распределения

Таблица 1

Дескриптивные статистики выборки

	X_1	X_2	X_3	X_4	%
$\bar{X}$	5,833	3,061	3,752	1,197	—
σ_X	0,825	0,437	1,761	0,759	—
v_X	0,681	0,191	3,099	0,577	—
X_1	1	−0,108	0,861	0,803	—
X_2	−0,108	1	−0,429	−0,364	—
X_3	0,861	−0,429	1	0,958	—
X_4	0,803	−0,364	0,958	1	—
C_1	0,887	−0,399	0,997	0,963	92,1
C_2	0,410	0,813	−0,051	−0,055	5,5
C_3	−0,196	0,405	0,011	0,209	2
C_4	0,063	−0,126	−0,045	0,163	0,4

Примечание: первые три строки характеризуют средние значения (измерения в см) X, стандартные отклонения σ_X и коэффициенты вариации v_X; последующие четыре строки — матрица корреляции; последние четыре строки — коэффициенты корреляции $r(X_i, C_k)$ переменной X_i и главной компоненты C_k, последний столбец — значимость ГК (отношение собственного значения соответствующей матрицы алгоритма нахождения ГК к их сумме).

Таблица 2

Диапазоны изменения переменных, см

	1	2	3
X_1	5,006 ± 0,598	5,936 ± 0,876	6,558 ± 1,118
X_2	3,428 ± 0,643	2,770 ± 0,533	2,982 ± 0,557
X_3	1,460 ± 0,301	4,260 ± 0,797	5,536 ± 0,930
X_4	0,246 ± 0,179	1,326 ± 0,336	2,018 ± 0,467

Примечание: в столбцах характеристики отдельных видов цветков.

объектов выборки. В табл. 2 представлены диапазоны изменения переменных в формате $\langle X_{ik}\rangle \pm t_\alpha \sigma_{ik}$, где первый член среднее значение i-й переменной k-го вида, t_α—t-статистика Стьюдента уровня значимости α (следуя [17], при таком объеме выборки и $\alpha = 0{,}005$ она принимается равной 1,96) и σ_{ik} — соответствующее среднее квадратичное отклонение.

Сопоставление элементов столбцов показывает перекрытие диапазона изменения X_1 для всех разновидностей. Такая же ситуация отмечается и для X_2. Вследствие чего в рамках этих переменных цветки не разделяются по их разновидностям. В переменных X_3 и X_4 значимо выделяются цветки 1-го вида ирисов, что обусловлено значимыми отличиями диапазонов их изменений.

Переменная X_2 имеет наименьшие корреляционные связи с остальными переменными, а переменная X_3 — наибольшие. При этом только она коррелирует со второй ГК (коэффициент корреляции 0,883). Такая ситуация отражается в их корреляционных связях с ГК. В частности, среди всех переменных только X_2 коррелирует со второй ГК: коэффициент корреляции $r(X_2, C_2) = 0{,}833$, а $(X_2, C_1) = -0{,}399$. Коэффициенты корреляции C_2 с остальными переменными не превышают 0,410. Высокие значения $r(X_1, C_1) = 0{,}889$, $r(X_3, C_1) = 0{,}998$ и $r(X_4, C_1) = 0{,}963$ указывают на характер изменения этой ГК: ее рост обусловлен ростом этих переменных. На рис. 2 (см. третью сторону обложки) представлена конфигурация проекций объектов выборки на плоскость первых двух ГК.

Распределение маркеров объектов выборки на плоскости первых двух ГК указывает существенное различие в рассматриваемом признаковом пространстве ирисов 1-го вида (*Setosa*). Ирисы других видов (*Versicolor* и *Virginica*) сосредоточены в отдельном эллипсоидальном вытянутом достаточно плотном облаке. Конфигурация такого распределения указывает, что исходная выборка определяется двумя кластерами: первый кластер составляют ирисы *Setosa*, а второй кластер — ирисы совокупности *Versicolor* и *Virginica*.

Автоматическая классификация методами кривых Пеано

Автоматическая классификация объектов выборки есть процедура разбиения выборки на отдельные достаточно однородные (в смысле заранее заданного критерия) кластеры/классы (группы), выявлении ее структуры при отсутствии обучающих выборок [3]. Представляется, что эффективный инструментарий решения этой проблемы можно построить на основании последовательности построений кривых Пеано. С их помощью выполняется отображение исходных выборочных точек n-мерного гиперкуба на отрезок вещественной оси. Суть построений кривых состоит в равномерном регулярном разбиении гиперкуба выборочных точек

на составляющие (элементы разбиения или гиперкванты в терминологии работы [5]) и проведение определенным образом через их центры кусочно-ломаной без самопересечений кривой. Выборочные точки элемента разбиения предписываются его центру. Поскольку число разбиений является входным параметром такой процедуры, то с помощью кривых Пеано устанавливается иерархия распределения объектов выборки. Для случая $n = 2$ приближения к так называемым разверткам Пеано (кривым Пеано) представлены на рис. 3.

При таком отображении близкие на отрезке вещественной оси образы выборочных точек близки и в исходном пространстве. Однако близким в исходном пространстве точкам могут соответствовать

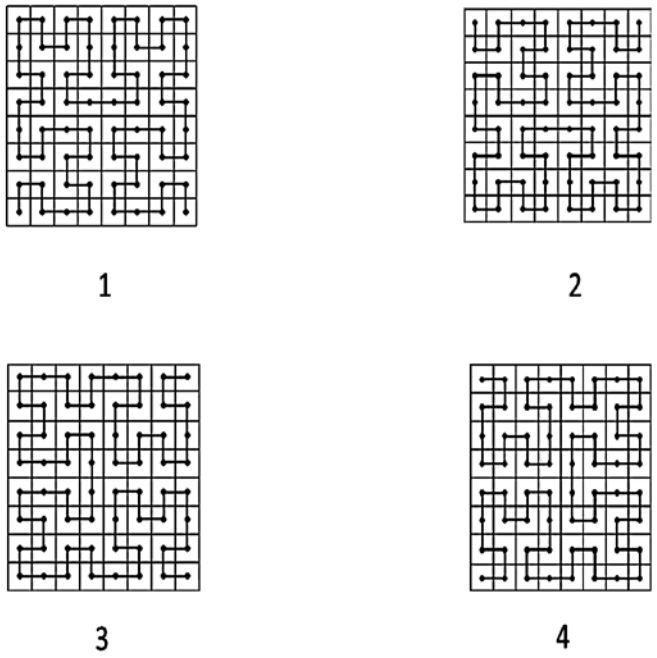

Рис. 3. Приближения к кривым Пеано (черными кружками отмечены узлы сетки, цифрами — варианты вращений)

Таблица 3

Распределение общего числа случаев смежности гиперквантов

\	1	2	3	4	5	6	7	8	9	10	11	12
1	1	6	0	6	0	0	0	0	0	0	0	0
2	6	1	6	0	0	0	0	0	0	0	0	4
3	0	6	1	6	0	0	0	0	0	0	2	0
4	6	0	6	2	4	0	0	0	0	0	0	0
5	0	0	0	4	2	6	0	0	0	0	0	0
6	0	0	0	0	6	2	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	6	0	2	0
8	0	0	0	0	0	0	0	1	6	0	0	0
9	0	0	0	0	0	0	6	6	1	4	0	0
10	0	0	0	0	0	0	0	0	4	1	6	0
11	0	0	2	0	0	0	2	0	0	6	1	6
12	0	4	0	0	0	0	0	0	0	0	6	1

Примечание: элементы таблицы n_{ij} — суммарное число общих граней (смежность) i-го и j-го элементов разбиения при отображениях их выборочных точек; элементы диагонали — номера кластеров.

достаточно далекие их образы. Действительно, точки отрезка имеют левого и правого соседей, а точки пространства R^n имеют соседей по 2^n направлениям. Поэтому для выявления кластерной структуры многомерной выборки следует выявить в этом пространстве смежные элементы разбиения с выборочными точками (под смежностью понимается наличие между элементами разбиения общей грани).

Для выявления смежных элементов разбиения в в работе [9] выполняются отображения гиперкубов, которые получены из исходного гиперкуба некоторым сдвигом позиций выборочных точек. Недостаток такого подхода состоит в трудности формализации направления и численного значения такого сдвига, а также количества требуемых отображений. Здесь решение этой проблемы следует ортогональным вращениям исходного гиперкуба вокруг начала координат. Суммарное число поворотов развертки при отображении n-мерного гиперкуба на одномерный отрезок равно 2^n [18]. Матрица вращения определяется выражением

$$A = \begin{pmatrix} \cos\theta & \mp\sin\theta \\ \pm\sin\theta & \cos\theta \end{pmatrix},$$

где θ — угол поворота. При двумерном признаковом пространстве число таких вращений равно четырем (см. рис. 3). Поэтому здесь $\theta = 0, \pi/2, \pi, 3\pi/2$.

Распределения маркеров объектов на рис. 4 (см. третью сторону обложки) показывает, что они заполняют 12 элементов разбиения. При этом конфигурация образов выборочных точек относительно друг друга не зависит от вращения (тут уместна аналогия с распределениями точек абсолютно твердого тела при его движении [19]). Анализ случаев подтверждает достоверность наличия в исследуемой выборке двух кластеров. Один из них составляют ирисы *Setosa*, а второй — ирисы совокупности *Versicolor* и *Virginica*.

Здесь для автоматической классификации объектов предлагается такой алгоритм. При его построении учитывается тот факт, что при различных отображениях узор конфигурации и распределение выборочных точек не изменяются. Для каждого результата вращения выполняется отображение выборочных точек на отрезок вещественной оси.

Далее образы элементов разбиения упорядочиваются согласно номерам элементов разбиения первого отображения. Например, если при первом отображении интервал отрезка соответствует i-му номеру исходного гиперкуба, то образы выборочных точек этого гиперкуба при других отображениях также получат этот номер. Сведенные в единую таблицу (табл. 3) эти номера позволяют определить смежность гиперкубов. К одному кластеру приписывают те выборочные точки, которые принадлежат только смежным гиперкубам. На диагонали таблицы указываются номера кластеров.

Заключение

В работе разработан алгоритм автоматической классификации выборочного материала на основании рекурсивного метода построений кривой Пеано. Работоспособность алгоритма иллюстрируется результатом решения этой проблемы на примере общедоступной выборки Р. Фишера 150 четырехмерных наблюдений.

Список литературы

1. **Springel V.** The cosmological simulation code GADGET-2 // Mon. Not. R. Astron. Soc. 2005. 364, pp. 1105—1134.
2. **Ефремов Ю. Н., Чернин А. Д.** Крупномасштабное звездообразование в галактиках // УФН. 2003. Т. 173, № 1. С. 3—25.
3. **Фукунага К.** Введение в статистическую теорию распознавания образов. М.: Наука, 1979. — 368 с.
4. **Лузин Н. Н.** Теория функций действительного переменного. 2-е изд. М.: Учпедгиз, 1948. 321 с.
5. **Александров В. В., Горский Н. Д., Поляков А. О.** Рекурсивные алгоритмы обработки и представления данных. — Препр. Ленинград: НИВЦ АН СССР, 1979. 53 с.
6. **Peano G.** Sur une courbe, qui remplit toute une aire plane // Math. Ann., 1890. N. 36 (1). P. 57—160.
7. **Hilbert D.** Uber die stetige Abbildung einer Linie auf ein Flachenstuck // Math. Annln., 1891. 38. P. 459—460.
8. **Стронгин Р. Г.** Численные методы решения многоэкстремальных задач (информационно-статистические алгоритмы). М.: Наука, 1978. 240 с.
9. **Александров В. В., Горский Н. Д.** Алгоритмы и программы структурного метода обработки данных. Л.: Наука, 1983. 208 с.
10. **Четырбоцкий А. Н.** Методы и алгоритмы решения задач снижения размерности пространства описания. Владивосток: ДВО АН СССР, 1991. 95 с.
11. **Стронгин Р. Г., Гергель В. П., Баркалов К. А.** Параллельные методы решения задач глобальной оптимизации // Известия высших учебных заведений. Приборостроение. 2009. Т. 52, № 1. С. 25—32.
12. **Jin G., Mellor-Crummey J.** Using Space-filling Curves for Computation Reordering // LACSI, 2005.
13. **Бауман К. Е.** О квадратно-линейном отношении правильных кривых Пеано. Автореф. ... дис. ... канд. физ.-мат. наук. М.: МГУ, 2012. http://mech.math.msu.su/~snarkyfiles/vak/arzh2.pdf
14. **Fisher R. A.** The use of multiple measurements in taxonomic problems // Ann. Eugen. 1936-7, 179. P. 466—475.
15. **Кендалл М. Дж., Стьюарт А. Т.** Многомерный статистический анализ и временные ряды. М.: Наука. 1976. 736 с.
16. http://matlab.exponenta.rU/fuzzylogic/book2/4/irisfcm.php
17. **Болч Б., Хуань Дж.** Многомерные статистические методы для экономики. М.: Статистика, 1979. 317 с.
18. **Баркалов К. А., Рябов В. В., Сидоров С. В.** О некоторых способах балансировки локального и глобального поиска в параллельных алгоритмах глобальной оптимизации // Выч. мет. и программирование, 2010. Т. 11, вып. 4. С. 382—387.
19. **Жилин П. А.** Рациональная механика сплошных сред: учеб. пособие. СПб.: Изд-во политех. ун-та, 2012. 584 с.

A. N. Chetyrbotsky, D. Sc., Leading Researcher, Far East Geological Institute,
The Far Eastern Federal University e-mail: Chetyrbotskv@vandex.ru

Automatic Classification of Objects of Multidimensional Sampling Recursive Method of Constructing Curves of Peano (for Example, Fisher's iris data)

Based on the recursive method of constructing curves of Peano developed an algorithm of automatic classification of objects of multidimensional sampling. The essence of development is to agree among themselves the respective distributions of images of objects on the segment of the real axis. To evaluate the performance of the algorithm used publicly available for use by the sample of 150 observations of four-three species of iris flowers (Fisher's iris data). The results of the algorithm and performed their analysis.

Keywords: *multidimensional sampling, cluster analysis, space-filling curves*

References

1. **Springel V.** The cosmological simulation code GADGET-2, *Mon. Not. R. Astron. Soc*, 2005, 364, pp. 1105—1134.
2. **Efremov Ju. N., Chernin A. D.** Krupnomasshtabnoe zvezdoobrazovanie v galaktikah, *UFN*, 2003, vol. 173, no. 1, pp. 3—25.
3. **Fukunaga K.** *Vvedgniev statisticheskuju teoriju raspoznavanija obrazov,* Moscow, Nauka, 1979, 368 p.
4. **Luzin N. N.** *Teorija funkcij dejstvitel'nogo peremennogo*, 2-e izd, Moscow, Uchpedgiz, 1948, 321 p.
5. **Aleksandrov V. V., Gorskij N. D., Poljakov A. O.** Rekursivnye algoritmy obrabotki i predstavlenija dannyh, Prepr. Leningrad, NIVC AN SSSR, 1979, 53 p.
6. **Peano G.** Sur une courbe, qui remplit toute une aire plane, *Math. Ann.*, 1890, 36 (1), pp. 57—160.
7. **Hilbert D.** Uber die stetige Abbildung einer Linie auf ein Flachenstuck, *Math. Annln.*, 1891, 38, pp. 459—460.
8. **Strongin R. G.** Chislennye metody resheniia mnogojekstremarnyh zadach (informacionno-statisticheskie algoritmy), Moscow, Nauka, 1978, 240 p.
9. **Aleksandrov V. V., Gorskij N. D.** Algoritmy i programmy strukturnogo metoda obrabotki dannvh. Leningrad, Nauka, 1983, 208 p.
10. **Chetyrbockij A. N.** *Metody i algoritmy reshenija zadach snizhenija razmernosti prostranstva opisanija*, Vladivostok, DVO AN SSSR, 1991, 95 p.
11. **Strongin R. G., Gergel' V. P., Barkalov K. A.** Parallel'nye metody reshenija zadach global'noj optimizacii, *Izvestija vysshih uchebnyh zavedenij, Priborostroenie*, 2009, vol. 52, no. 10, pp. 25—32.
12. **Jin G., Mellor-Crummey J.** Using Space-filling Curves for Computation Reordering, *LACSI*, 2005.
13. **Bauman K. E.** O kvadratno-linejnom otnoshenii pravil'nyh krivyh Peano. — avtoref. ... dis. kand. fiz-mat nauk, Moscow: MGU, 2012. http://mech.math.msu.su/~snark/files/vak/arzh2.pdf
14. **Fisher R. A.** The use of multiple measurements in taxonomic problems, *Ann. Eugen.*, 1936-7, 179, pp. 466—475.
15. **Kendall M. Dzh., St'juart A. T.** *Mnogomernyj statisticheskij analiz i vremennye rjady*, Moscow, Nauka, 1976, 736 p.
16. http://matlab.exponenta.ru/fuzzylogic/book2/4/irisfcm.php
17. **Bolch B., Huan' Dzh.** *Mnogomernye statisticheskie metody dlja jekonomiki*, Moscow, Statistika, 1979, 317 p.
18. **Barkalov K. A., Rjabov V. V., Sidorov S. V.** O nekotoryh sposobah balansirovki v lokal'nogo i global'nogo poiska parallel'nyh algoritmah global'noj optimizacii, *Vych. Met. Programmirovanie*, 2010. — t. 11, vyr. 4. — s. 382—381.
19. **Zhilin P. A.** Racional'naja mehanika sploshnyh sred: ucheb. posobie, Sankt Peterburge, Izd-vo politeh. un-ta, 2012, 584 p.