УДК 681.325

Н.Н. Пономаренко

Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков

БЫСТРАЯ КЛАСТЕРИЗАЦИЯ В МНОГОМЕРНОМ ПРОСТРАНСТВЕ ДЛЯ ЗАДАЧ ПОИСКА ПОДОБИЯ

В работе предложен быстрый и эффективный метод кластеризации произвольного множества по расстоянию до одного из его граничных элементов. Метод является пригодным не только для традиционных задач векторного квантования, но и для задач поиска подобия в многомерном пространстве, таких, как поиск подобного блока изображения во фрактальном сжатии или же поиск подобных изображений по заданному образцу. Спецификой кластеризации для таких задач часто является наличие только некоторой нелинейной функции расстояния между заданными элементами множества, зависящей от значений различных признаков элемента множества, число которых может достигать нескольких тысяч. Высокая эффективность предложенного метода продемонстрирована в сравнении с традиционным методом медианного сечения на примере задачи поиска подобных блоков изображения.

Ключевые слова: *кластеризация, векторное квантование, поиск изображений по подобию, сжатие изображений с потерями.*

Введение

Кластеризация, или же векторное квантование [1], широко используется при решении различных задач, таких, как сжатие данных, понимание данных, обнаружение новизны, поиск подобия, сегментация изображений и др.

Исходными данными для кластеризации могут служить либо признаковое описание элементов множества (каждый из них описывается набором своих характеристик, называемых признаками), либо матрица расстояний между элементами множества. Последнее характерно для задачи поиска подобных элементов множества, которую необходимо уметь эффективно решать в таких приложениях, как, например, поиск изображений по подобию [2], фрактальное сжатие изображений [3], подавление шума на изображениях [4] и многих других. Число элементов множеств при этом может достигать от миллионов для задач обработки изображений до миллиардов для задач поиска подобных изображений в Интернете [2]. При таком большом количестве элементов множеств практически невозможным является полное построение матрицы расстояний, и основным требованием для метода кластеризации становится минимизация количества вычислений расстояний в процессе кластеризации.

Наиболее известные методы кластеризации, такие как обобщенный алгоритм Ллойда [5] (метод k-средних), являются слишком медленными (требуют слишком большого числа вычислений функции расстояния) [6], чтобы их можно было использовать в рассматриваемой задаче. К тому же они подразумевают вычисление центров кластеров усреднением значений признаков входящих в них элементов, что никак не учитывает возможный нелинейный вид заданной функции расстояния. Быстрые алгоритмы, такие, как алгоритм медианного сечения, эффективны только для относительно небольшого количества признаков [7] и так же не способны учитывать возможный нелинейный вид функции расстояния. Поэтому актуальной является задача разработки новых методов кластеризации, оперирующих в процессе построения кластеров только вычисленными значениями расстояний между элементами множества.

В данной работе предлагается новый быстрый метод кластеризации, обеспечивающий, на выбор, заданные размеры кластеров или же заданное максимальное расстояние между элементами кластера, и обладающий сложностью меньше $O(M*\log_2 M)$, где M – число элементов в исходном множестве. Предлагаемый метод является универсальным и может эффективно использоваться для задач кластеризации обоих типов (с известными наборами признаков элементов либо с заданной функцией расстояния).

1. Описание предлагаемого метода

Пусть $\mathbf{X^m} = \{x_1, \ldots, x_m\}$ – множество объектов, которое требуется кластеризовать. Предположим, что о характеристиках объектов x_i ничего не известно, либо они настолько сложны, что их практически невозможно учесть напрямую при кластеризации. Например, для задачи поиска подобных изображений элементами множества $\mathbf{X^m}$ могут быть комбинации данных разного типа, общим объемом до нескольких тысяч байт [2]. В то же время предположим, что задана функция расстояния [2] между элементами множества $\rho(x, x')$. Требуется разбить множество $\mathbf{X^m}$ на кластеры так, чтобы каждый кластер состоял из элементов, близких по метрике ρ, а

элементы, принадлежащие разным кластерам, существенно различались. Отметим, что обычно кластерами являются непересекающиеся множества, но для задач поиска подобия может быть целесообразным создание пересекающихся кластеров (один и тот же элемент принадлежит одновременно нескольких кластерам), если это может повысить качество поиска. Идея предлагаемого метода кластеризации иллюстрируется рис. 1.

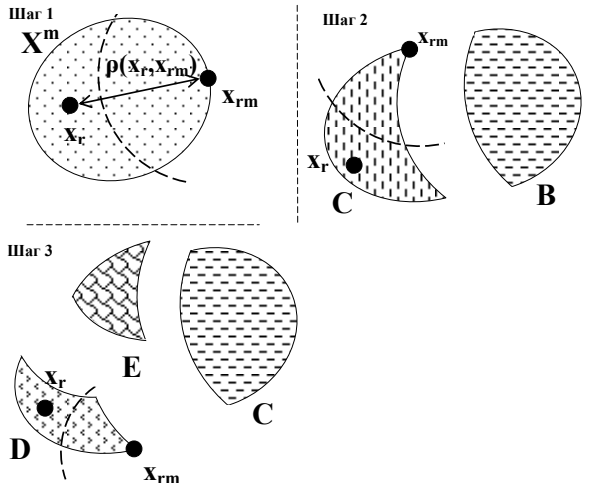

Рис. 1. Графическая иллюстрация предлагаемого метода кластеризации

На первом шаге кластеризации из кластеризуемого множества $\mathbf{X}^m$ случайно выбирается один элемент x_r. Далее перебираются все элементы множества $\mathbf{X}^m$ и среди них находится элемент x_{rm}, максимально удаленный от x_r. Для этого потребуется M вычислений метрики ρ, где M - число элементов в множестве $\mathbf{X}^m$. Очевидно, что выбранный подобным образом элемент x_{rm} будет лежать на какой-либо границе множества $\mathbf{X}^m$. Затем еще раз перебираются все элементы множества $\mathbf{X}^m$, и для каждого x_i вычисляется расстояние $\rho_i = \rho(x_{rm}, x_i)$. Для этого потребуется еще M вычислений метрики. В завершение шага кластеризации множество $\mathbf{X}^m$ разделяется на два кластера **B** и **C** в соответствии с одной из выбранных стратегий:

Стратегия 1: элементы x_i со значением ρ_i, меньшим заданного порога ρ_{tr}, помещаются в кластер **B**, а остальные - в кластер **C**. Данная стратегия хорошо подходит для работы с искаженными каким-либо шумом данными. Если кластер **B** в итоге содержит число элементов, меньшее некоторого заданного порога, то элемент x_{rm} можно считать шумоподобным и подвергнуть фильтрации.

Стратегия 2: множество $\mathbf{X}^m$ сортируется по возрастанию ρ_i, после чего первые M/2 элементов в отсортированной выборке помещаются в кластер **B**, а оставшиеся - в кластер **C**. Такая стратегия обеспечивает равные по количеству элементов кластеры. Рис. 1 иллюстрирует именно вторую стратегию кластеризации. Пунктирной линией обозначена черта разделения кластера на две равные по числу элементов (в данном случае - равные по площади) части.

На втором шаге кластеризации подвергается уже множество (кластер) **C**, который разделяется на два кластера **D** и **E** таким же образом, как это делалось на шаге 1. На третьем шаге уже кластер **D** разделяется на два кластера и т.д.

Кластеризация осуществляется рекурсивно для всех сформированных кластеров до тех пор, пока не выполнится условие завершения кластеризации. Для первой стратегии кластеризация прекращается, если для всех x_i выполняется условие $\rho_i \le \rho_{tr}$. Для второй стратегии кластеризация прекращается, если число элементов в кластере меньше заданного порога. В пределе, при разбиении множества $\mathbf{X}^m$ на кластеры, содержащие по 2 элемента, потребуется $2M*\log_2 M$ вычислений расстояния ρ, то есть сложность метода кластеризации не превышает $O(M*\log_2 M)$.

Отметим, что стратегии 1 и 2 можно применять непосредственно к x_r вместо x_{rm}, однако в этом случае с большей вероятностью могут возникать нежелательные для обеспечения высокого качества кластеризации ситуации, проиллюстрированные рис. 2.

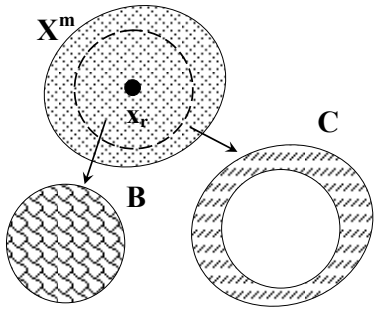

Рис. 2. Нежелательная ситуация, которая может возникнуть при кластеризации по расстоянию от случайно выбранного элемента x_r

В ситуации, показанной на рис. 2 кластер **C** плохо соответствует основному требованию, предъявляемому к кластерам - кластер должен состоять из элементов, близких по метрике ρ. При дальнейшем дроблении кластера **C** положение может исправиться, но использование граничных элементов x_{rm} позволяет избежать большинства подобных ситуаций.

Другая нежелательная ситуация иллюстрируется рис. 3. Граничный элемент x_{rm} в данном случае является аномальным (лежит за пределами основного множества).

В этом случае множество $\mathbf{X}^m$ может разделяться на кластеры **B** и **C** неоптимальным образом. Более того, кластер **B** по-прежнему содержит проблемный аномальный элемент, который и дальше может выбираться в качестве x_{rm}.

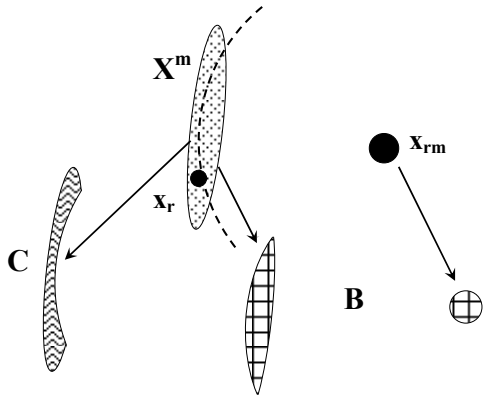

Рис. 3. Нежелательная ситуация, когда граничный элемент x_{rm} оказывается аномальным

Обеспечить робастность по отношению к аномальным элементам можно, модифицируя процедуру выбора x_{rm}. После вычисления расстояний от x_r до всех x_i множества $\mathbf{X^m}$ можно брать в качестве x_{rm} не максимально удаленный от x_r, а k-процентный квантиль в упорядоченной по возрастанию выборке, где k можно задавать в диапазоне от 95% до 98%. Чем больше k будет отличаться от 100%, тем большей робастностью будет обладать метод.

2. Анализ эффективности предложенного метода кластеризации

Оценим эффективность предложенного метода на примере задачи поиска подобных блоков изображений. В основу нелинейной функции расстояния между двумя блоками изображения R и Q положим коэффициент корреляции Пирсона между ними:

$$\rho(\mathbf{R},\mathbf{Q}) = 1 - \frac{\sum(R_i - \bar{R})(Q_i - \bar{Q})}{(n-1)\sigma_R\sigma_Q}, \qquad (1)$$

где σ_R и σ_Q – значения среднеквадратических отклонений соответственно для блоков **R** и **Q**.

При максимальном единичном значении коэффициента корреляции между блоками **R** и **Q** расстояние $\rho(\mathbf{R},\mathbf{Q})$ между ними, вычисленное в соответствии с (1), будет равно нулю.

Множество $\mathbf{X^m}$ сформируем следующим образом. Возьмем стандартные тестовые изображения Lena, Baboon, Barbara, Peppers, Goldhill, Bikes, Boat, House (все 512x512 пикселей, в оттенках серого цвета) и разобьем их на непересекающиеся блоки по 8x8 пикселей. Получим 32768 блоков, которые и составят множество $\mathbf{X^m}$.

Задача состоит в том, чтобы за ограниченное число сравнений для каждого x_i найти как можно более близкий к нему по расстоянию (1) блок.

Предложенный метод сравнивался с кластеризацией методом медианного сечения и со случайным перебором.

При кластеризации использовалась вторая стратегия, то есть все множество разбивалось на кластеры одинакового размера. Кластеризация осуществлялась на 32, 64, 128, 256, 512 и 1024 кластера. При этом для каждого блока осуществлялся поиск наиболее близкого блока внутри его кластера, и подсчитывалось общее количество вычисления расстояний в процессе кластеризации и поиска подобия. Ошибка поиска (расстояние до ближайшего найденного кластера) усреднялась для всех блоков.

На рис. 4 приведены полученные зависимости для трех сравниваемых методов.

Видно, что предложенный метод обеспечивает существенно более высокое качество поиска, чем метод медианного сечения, и, тем более, чем случайный перебор.

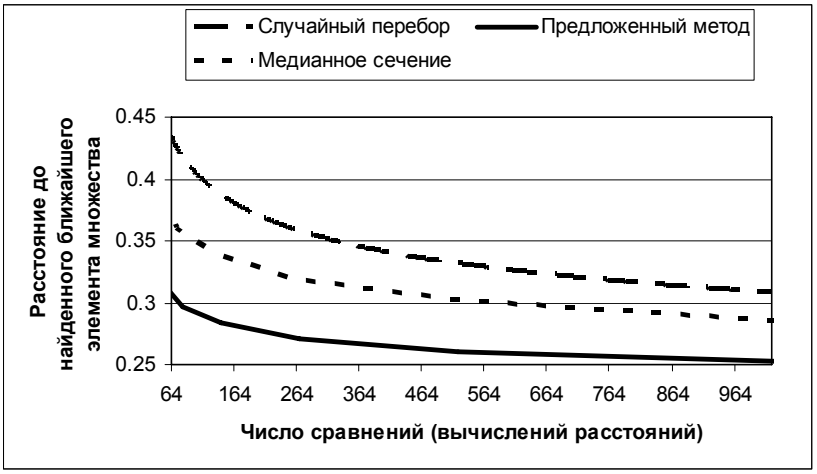

Рис. 4. Зависимость качества поиска подобия от числа сравнений для трех рассматриваемых методов

На рис. 5 приведено сравнение стандартного варианта предложенного метода, а также варианта без использования граничных элементов (ситуация на рис. 2) и робастного варианта (устойчивого к наличию аномальных элементов, k = 98%).

Видно, что использование граничных точек даже для такой достаточно простой задачи кластеризации оправдано и всегда обеспечивает лучший результат.

Робастный вариант незначительно уступает стандартному, что свидетельствует об отсутствии аномальных блоков в этом множестве.

В то же время этот проигрыш можно с читать несущественным, а использование робастного варианта вполне оправданным из-за его более высокой надежности.

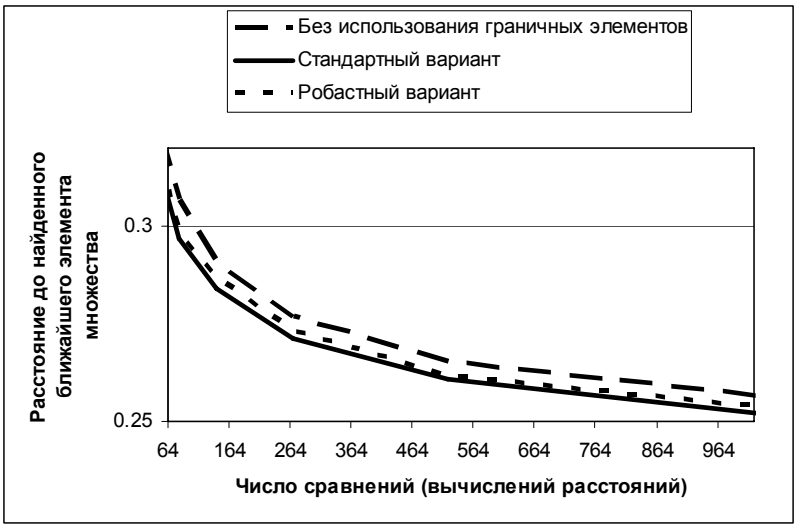

Рис. 5. Сравнение различных вариантов предложенного метода кластеризации

Заключение

В работе предложен новый эффективный метод кластеризации по расстоянию до граничного элемента множества.

Показана более высокая эффективность предложенного метода по сравнению с методом медианного сечения.

Список литературы

1. Gersho A. Vector Quantization and Signal Compression / A. Gersho, R.M. Gray. – Boston, MA: Kluwer, 1992. – 732 p.

2. Пономаренко Н.Н. Устойчивый поиск изображений по полному и тематическому подобию с использованием многопараметровой классификации / Н.Н. Пономаренко, В.В. Лукин, С.К. Абрамов // Интернет-математика (Яндекс, Россия). – Екатеринбург: Изд-во Уральского университета, 2007. – С. 171-180.

3. Fisher Y. Fractal Image Compression: Theory and Application / Y. Fisher. – Berlin, Germany: Springer-Verlag, 1995. – 341 p.

4. Image denoising by sparse 3-D transform-domain collaborative filtering / K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian // IEEE Trans. On Image Proc. – 2007. – Vol. 16, issue 8. – P. 2080-2095.

5. Lloyd S.P. Least squares quantization in PCM / S.P. Lloyd // Spec. issue on quantization, IEEE Trans. Inform. Theory. – 1982. – Vol. 28. – P. 129-137.

6. Arthur D. How Slow is the k-means Method? / D. Arthur, S. Vassilvitskii // Proceedings of the Symposium on Computational Geometry. – Arizona, USA. – June 5-7, 2006. – P. 144-153.

7. Speeding-up the Fractal Compression with Clustering / N.N. Ponomarenko, K. Egiazarian, V. Lukin, J. Astola // Proc. of All-Ukrainian Int. Conf. Signal/Image Processing and Pattern Recognition. – Kiev, Ukraine. – Nov. 27-Dec.1, 2000. – P. 55-58.

Поступила в редколлегию 19.01.2009

Рецензент: д-р техн. наук, проф. В.К. Волосюк, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.

ШВИДКА КЛАСТЕРИЗАЦІЯ У БАГАТОВИМІРНОМУ ПРОСТОРІ ДЛЯ ЗАДАЧ ПОШУКУ ПОДІБНОСТІ

М.М. Пономаренко

В роботі запропоновано швидкий та ефективний метод кластеризації довільної множини за відстанню до одного з її крайових елементів. Метод є придатним не тільки для традиційних задач векторного квантування, але й для задач пошуку подібності у багатовимірному просторі, таких, як пошук подібного блоку зображення у фрактальному стиску або пошук подібних зображень за заданим зразком. Специфікою кластеризації для таких задач часто є наявність лиш деякої нелінійної функції відстані між заданими елементами множини, яка залежить від значень різних ознак елемента множини, кількість яких може досягати кількох тисяч. Високу ефективність запропонованого методу продемонстровано у порівнянні з традиційним методом медіанного перерізу на прикладі задачі пошуку подібних блоків зображення.

Ключові слова: *кластеризація, векторне квантування, пошук зображень за подібністю, стиск зображень с втратами*

FAST CLUSTERING IN MULTIDIMENSIONAL SPACE FOR SIMILARITY SEARCH TASKS

N.N. Ponomarenko

A fast and effective method of clustering of a set by distance to an edge set member is proposed in this paper. The method is appropriate for tasks of similarity search in multidimensional space as well as for traditional tasks of vector quantization. An example of such similarity search tasks is searching of similar block of the image in fractal compression or searching of images similar to given sample. A peculiarity of clustering for such task is presence only a nonlinear distance function between set members. Often this function depends from thousands of set members features. Effectiveness of the proposed method in comparison to well known median splitting method is shown for task of searching similar blocks of images.

Keywords: *clustering, vector quantization, image similarity, image lossy compression.*