УДК 004.658

С.И. Богучарский[1], А.Г. Каграманян[2], С.В. Машталир[1]

[1] *Харьковский национальный университет радиоэлектроники, Харьков*
[2] *Харьковский национальный университет имени В.Н. Каразина, Харьков*

ИЕРАРХИЧЕСКАЯ АГЛОМЕРАТИВНАЯ КЛАСТЕРИЗАЦИЯ ИЗОБРАЖЕНИЙ В БОЛЬШИХ БАЗАХ ДАННЫХ

Работа посвящена задаче анализа больших баз данных изображений с точки зрения решения задачи интеллектуального поиска. Для решения этой задачи удобно представить все множество изображений в виде некоторых классов по мере их сходства. Для этого возможно применить аппарат кластерного анализа, методы которого и рассмотрены в данной работе. При этом предложены две матричные модификации известных подходов, позволяющие упростить анализ изображений за счет исключения операций векторизации-девекторизации исходных данных.

Ключевые слова: *изображение, матричная иерархическая кластеризация, выборка, кластер.*

Введение

Задача кластеризации (автоматической классификации) массивов многомерных наблюдений, основной целью которой является нахождение в обрабатываемых выборках данных однородных в принятом смысле групп (классов, сегментов, кластеров), является важной частью активно развивающегося в настоящее время научного направления, известного как интеллектуальный анализ данных [1 – 5].

Несмотря на то, что на сегодня известны десятки, если не сотни подходов, методов и алгоритмов кластеризации, говорить об «универсальном решении» всех возможных задач кластеризации не приходится, поскольку специфика каждой конкретной задачи определяется в значительной мере спецификой объекта или явления порождающего исходную выборку.

Весьма специфической задачей в широком смысле является различного типа обработка видеоинформации, включая и ее поиск в сверхбольших базах данных (VLDB). Основными факторами, определяющими сложность этой задачи, являются огромные объемы анализируемой информации, описание изображений, как правило, в матричной форме, наличие сегментов изображений сложной формы, их искаженность различного рода возмущениями и шумами. В таких задачах классические традиционные методы кластеризации оказываются неэффективными либо вообще неработоспособными.

Следуя классификации, введенной в [6], методы кластеризации подразделяются на два больших класса: основанные на разбиении и иерархические.

На сегодня наибольшее распространение получении процедуры, основанные на разбиении, которые разделяют массив информации, содержащий N многомерных наблюдений, описываемых n-мерными векторами признаков $x(k) \in R^n$, $k = 1, 2, \ldots, N$, на p классов (сегментов), где p – основной параметр, задаваемый, как правило, априорно из эмпирических соображений. Такие алгоритмы начинают свою работу с некоторого произвольного разбиения, которое в процессе оптимизации некоторой также априори заданной целевой функции, основанной на той или иной метрике (обычно евклидовой или манхэттенской), непрерывно корректируется с помощью той или иной итерационной процедуры. При этом каждое наблюдение, содержащееся в исходном массиве, неоднократно просматривается. Основной характеристикой каждого формируемого кластера является его центроид, вокруг которого группируются наблюдения конкретного класса. Наиболее характерными представителями этого подхода являются алгоритмы k-средних, k-медоидов и т.п. Несмотря на популярность и достаточно строгую формализацию алгоритмов разбиения, им присущи и существенные недостатки. Во-первых, они формируют выпуклые сегменты, которые в реальных изображениях присутствуют далеко не всегда. Конечно, всякую невыпуклую фигуру можно покрыть множеством кругов достаточно малого радиуса. Однако при этом возрастает вычислительная сложность алгоритма. А необходимость неоднократного просмотра каждого вектора-образа делает использование подобных алгоритмов в VLDB крайне проблематичным.

В отличие от этого подхода иерархические алгоритмы, которые, в свою очередь, делятся на агломеративные и дивизимные, автоматически определяют число классов путем слияния отдельных наблюдений в кластеры либо дробления исходного массива данных на подвыборки – кластеры. Понятно, что в этом случае число формируемых кластеров может лежать в интервале $2 \le p \le N-1$. Иерархический подход может формировать кластеры произвольной формы, крайне прост с алгоритмической точки зрения, однако в силу необходимости многократного просмотра всех наблюдений, хранящихся в базе данных, крайне

неудобен для обработки информации, содержащейся в VLDB. Кроме того, получаемые с помощью этого подхода результаты весьма чувствительны к различного рода возмущениям и шумам, всегда присутствующим в реальных данных.

Нельзя не отметить также интенсивно развивающиеся последние годы методы кластеризации, основанные на плотности распределению данных [1, 2]. Эти методы позволяют формировать кластеры произвольной формы в условиях, когда данные «зашумлены» возмущениями различной природы. При этом в рамках данного подхода под кластерами понимаются области в пространстве признаков с наиболее высокой плотностью распределения данных. Эти области чередуются с областями с низкой плотностью, где и концентрируются возмущения и помехи. Таким образом, алгоритмы, основанные на плотности, в процессе вычислений «выращивают» области с высокой плотностью распределения и формируют кластеры произвольной формы, отделяя при этом возмущения и шумы. Также нельзя не отметить, что данный подход может быть адаптирован для работы с VLDB. Вместе с тем алгоритмы, основанные на плотности, требуют предварительного задания ряда параметров, определяющих в конечном итоге качество получаемого результата. Таким образом, пользователь должен быть специалистом в конкретной предметной области, а математическая сложность соответствующих алгоритмов затрудняет интерпретацию этих результатов.

Интуитивно понятно, что наиболее простым, понятным, хорошо интерпретируемым и наглядным является иерархический подход, и если бы удалось существенно сократить объем «просматриваемых» данных для формирования устойчивых кластеров, его можно было бы рекомендовать и для работы с VLDB. Понятно также, что эта идея не могла не привлечь внимание исследователей [7 – 9].

1. Матричный итеративный иерархический балансовый метод кластеризации

Исторически первым иерархическим агломеративным алгоритмом кластеризации, ориентированным на работу с VLDB, является BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [7], вычислительная сложность которого линейно зависит от объема обрабатываемой выборки, а возможность обработки данных в последовательном режиме делает его особенно привлекательным.

В рамках BIRCH вводится два основных понятия этого метода: «признак кластеризации» (Clustering Feature – CF) и «дерево признаков кластеризации» (CF – Tree, CF – дендрограмма), при этом здесь под дендрограммой понимается вложенная группировка объектов, образов, векторов признаков и т.п., изменяющаяся по определенным правилам на различных уровнях иерархии. Использование введенных понятий позволяет упростить вычисления, повысить скорость обработки и организовать динамическую кластеризацию вновь поступающих данных.

Чтобы ввести эти понятия, допустим, что l-й кластер Cl_l образован N_l n-мерными объектами-образами $x(k) \in Cl_l$. Для этого кластера вводится его центроид

$$C(l) = \frac{1}{N_l} \cdot \sum_{k=1,\, x(k) \in Cl_l}^{N_l} x(k),$$

радиус

$$R(l) = \left(\frac{1}{N_l} \cdot \sum_{k=1,\, x(k) \in Cl_l}^{N_l} \| x(k) - C(l) \|^2 \right)^{1/2}$$

и диаметр

$$Dm(l) =$$

$$= \left(\frac{1}{N_l (N_l - 1)} \cdot \sum_{k=1,\, x(k) \in Cl_l}^{N_l} \sum_{q=1,\, x(q) \in Cl_l}^{N_l} \| x(k) - x(q) \|^2 \right)^{1/2}.$$

Как несложно заметить, эти признаки порождены евклидовой метрикой, хотя несложно заменить ее метрикой Минковского.

Собственно же кластеризация на верхних уровнях иерархии реализуется на основе CF-векторов, определяемых тройкой

$$CF(l) = (N_l, LS^T(l), SS(l))^T,$$

где $LS(l) = \sum_{k=1,\, x(k) \in Cl_l}^{N_l} x(k)$, $SS(l) = \sum_{k=1,\, x(k) \in Cl_l}^{V_l} \| x(k) \|^2$.

Таким образом, размерность вектора $CF(l)$ есть $(n+2) \times 1$. Важно, что CF-векторы обладают свойством аддитивности, т.е.

$$CF(l,\, l+1) = CF(l) + CF(l+1) =$$

$$= (N_l + N_{l+1},\, LS^T(l) + LS^T(l+1),\, SS(l) + SS(l+1))^T$$

и могут уточняться по мере поступления новых данных.

В дальнейшем BIRCH оперирует только с CF, которые накапливаются и анализируются с помощью CF-дендрограммы. Сама же CF-дендрограмма характеризуется двумя признаками: фактором ветвления B, который определяет максимальное число элементов в каждом кластере (субкластере) на нижнем уровне иерархии, и максимальным диаметром субкластера T в каждом узле дерева (уровня иерархии). Понятно, что чем больше значение B, тем меньшее количество кластеров будет сформировано, а чем больше T – тем меньше уровней иерархии будет иметь дендрограмма.

Кластеризация на основе BIRCH происходит в два этапа: на первом этапе в результате однократного просмотра базы данных формируется исходная

дендрограмма, при этом на нижнем уровне обрабатываются исходные данные $x(k)$, а на верхних – векторы признаков $CF(l)$. На втором этапе производится кластеризация сформированных на первом этапе субкластеров, при этом субкластеры, содержащие малое число объектов, удаляются как шумы и выбросы, а субкластеры, чьи центроиды расположены достаточно близко, т.е.

$$D(Cl_l, Cl_r) = \| C(l) - C(r) \| < \Delta,$$

сливаются в один.

Таким образом, достигается робастность процесса кластеризация, а последовательная обработка сводится к тому, что вновь поступивший образ $x(k)$ просто «вставляется» в субкластер с наиболее близко расположенным центроидом.

Реализация BIRCH предполагает, что каждый обрабатываемый объект описывается n-мерным вектором $x(k)$, а сам процесс обработки информации связан с векторными операциями. В ситуации, когда необходимо обрабатывать двумерные изображения, они должны быть подвергнуты векторизации, что ведет к резкому возрастанию размерности CF-векторов, после чего решается собственно задача кластеризации, результат решения которой далее должен быть девекторизован. Процесс кластеризации массивов изображений можно упростить, используя вместо векторных операций соответствующие матричные операции, при этом исходная информация задается не в форме n-мерных векторов, а в виде матриц $x(k) = \{x_{i_1, i_2}(k)\}$, $i_1 = 1, 2, ..., m$; $i_2 = 1, 2, ..., n$; $k = 1, 2, ..., N$; $x(k) \in R^{m \times n}$.

Далее можно ввести в рассмотрение центроид l-го кластера

$$C(l) = \frac{1}{N_l} \cdot \sum_{k=1,\, x(k) \in Cl_l}^{N_l} x(k),$$

матричный радиус

$$R_l = \left(\frac{1}{N_l} \sum_{k=1,\, x(k) \in Cl_l}^{N_l} Sp(x(k) - C(l))(x(k) - C(l))^T \right)^{1/2},$$

матричный диаметр

$$Dm(l) = \left(\frac{1}{N_l (N_{l-1} - 1)} \times \right.$$

$$\left. \times \sum_{k=1,\, x(k) \in Cl_l}^{N_l} \sum_{q=1,\, x(q) \in Cl_l}^{N_l} Sp(x(k) - x(q))(x(k) - x(q))^T \right)^{1/2}$$

и CF-матрицу $CF(l) = \begin{pmatrix} N_l & \vdots \\ SS(l) & \vdots \\ \cdots\cdots & LS(l) \\ \vec{0} & \vdots \end{pmatrix}$ размерности $(m \times (n+1))$.

Здесь

$$LS(l) = \sum_{k=1,\, x(k) \in Cl_l}^{N_l} x(k), \quad SS(l) = \sum_{k=1,\, x(k) \in Cl_l}^{N_l} Spx(k) x^T(k),$$

$\vec{0}$ – вектор, образованный нулями.

При этом в качестве меры расстояния используется сферическая норма

$$D_S(x(k), x(r)) = (Sp(x(k) - x(r))((x(k) - x(r))^T)^{1/2}.$$

Далее может быть реализована BIRCH-процедура, где вместо CF-векторов используются введенные нами CF-матрицы. Данная процедура может быть реализована в виде следующей схемы, представленной на рис. 1.

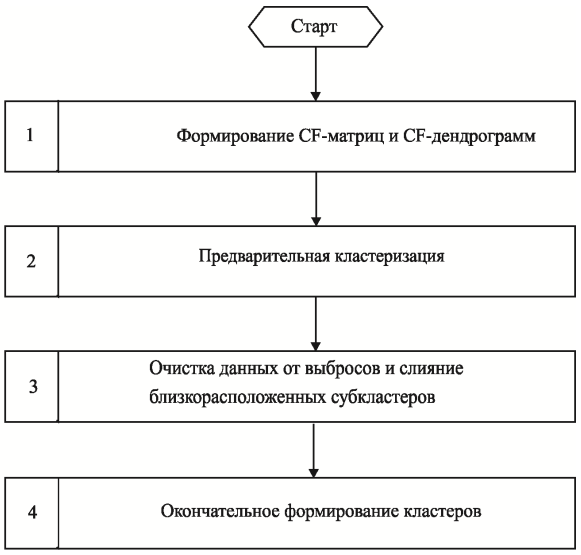

Рис. 1. Матричный метод BIRCH

Таким образом, можно говорить о том, что матричная модификация метода BIRCH проста в численной реализации, достаточно быстра, робастна к различного рода выбросам, допускает последовательную обработку данных.

Главный недостаток процедуры состоит в том, что в результате ее применения формируются только выпуклые кластеры, что, естественно, ограничивает ее применимость и заставляет искать альтернативные подходы.

2. Матричный иерархический метод кластеризации на основе сокращенной репрезентативной выборки

Формировать кластеры не сферической формы различных размеров и плотности позволяет метод CURE (Clustering Using Representatives) [8], сохраняя при этом работоспособность и возможность работы с VLDB. CURE является своеобразным гибридом методов кластеризации, основанных на разбиении, и иерархических агломеративных алгоритмов [9]. В рамках этого подхода вначале формируется фиксированное множество репрезентативных точек, характеризующих каждый кластер, при этом в пределе это число может быть равным единице, что

роднит CURE с популярными алгоритмами k-средних и k-медоидов. При этом использование не одной точки-центроида, а именно множества позволяет формировать кластеры произвольной формы. Далее эти репрезентативные точки, а не вся выборка из VLDB подвергается агломеративной кластеризации. Таким образом, резко сокращается число обрабатываемых наблюдений, что и позволяет успешно использовать этот метод при работе с очень большими массивами данных.

Работа CURE реализуется за шесть последовательных этапов. На первом этапе случайным образом формируется так называемая репрезентативная выборка, такая, чтобы она сохраняла информацию о геометрии кластеров. Размер такой выборки определяется так называемыми границами Чернова [10] и задается выражением

$$S \ge \mu N + \frac{N}{N_l}\log\frac{1}{\varepsilon} + \frac{N}{N_l}\sqrt{\left(\log\frac{1}{\varepsilon}\right)^2 + 2\mu N_l \log\frac{1}{\varepsilon}},$$

где $0 \le \mu \le 1$.

Данное выражение дает оценку, что вероятность того, то репрезентативная выборка содержит меньше чем μN_l наблюдений в кластере Cl_l, меньше чем $0 \le \varepsilon \le 1$. Здесь можно заметить, что популярный метод кластеризации CLARANS [11] также оперирует со случайной подвыборкой, однако в CURE объем этой подвыборки задается на основе формальных соображений.

На втором этапе репрезентативная выборка разбивается на h частей, т.е. формируется h выборок, каждая из которых содержит sh^{-1} наблюдений.

На третьем этапе с помощью любого из известных алгоритмов кластеризации производится независимая предкластеризация каждой из подвыборок-сегментов репрезентативной выборки, в результате чего на каждом из сегментов может быть сформировано различное число подкластеров.

На четвертом этапе подкластеры, содержащие малое число наблюдений, рассматриваются как шумы и выбросы и исключаются из дальнейшего анализа и обработки. Именно четвертый этап обеспечивает робастные свойства методу CURE.

На пятом этапе производится стандартная агломеративная иерархическая кластеризация всех сформированных подкластеров, однако поскольку на этом этапе объединяются не отдельные наблюдения, а подкластеры, содержащие множество наблюдений, вместо расстояния между отдельными образами (в нашем случае матрицами) используются межкластерные расстояния такие, как:

$$D_{mcan}(Cl_l, Cl_r) = \|C_l - C_r\| = \left(Sp(C_l - C_r)(C_l - C_r)^T\right)^{1/2},$$

$$D_{ave}(Cl_l, Cl_r) = \frac{1}{N_l N_r} \times$$

$$\times \sum_{x(k)\in Cl_l} \sum_{x(r)\in Cl_r} \left(Sp(x(k)-x(r))(x(k)-x(r))^T\right)^{1/2},$$

$$D_{max}(Cl_l, Cl_r) =$$

$$= \max_{x(k)\in Cl_l,\, x(r)\in Cl_r} \left(Sp(x(k)-x(r))(x(k)-x(r))^T\right)^{1/2},$$

$$D_{min}(Cl_l, Cl_r) =$$

$$= \min_{x(k)\in Cl_l,\, x(r)\in Cl_r} \left(Sp(x(k)-x(r))(x(k)-x(r))^T\right)^{1/2}.$$

И, наконец, на последнем (шестом) этапе каждое из наблюдений, содержащихся в VLDB, но не принадлежащее репрезентативной выборке, «приписывается» к одному из сформированных кластеров по признаку минимального расстояния от этого наблюдения до любого из образов репрезентативной выборки.

Процедура кластеризации большого массива изображений-матриц может быть реализована в виде схемы, приведенной на рис. 2.

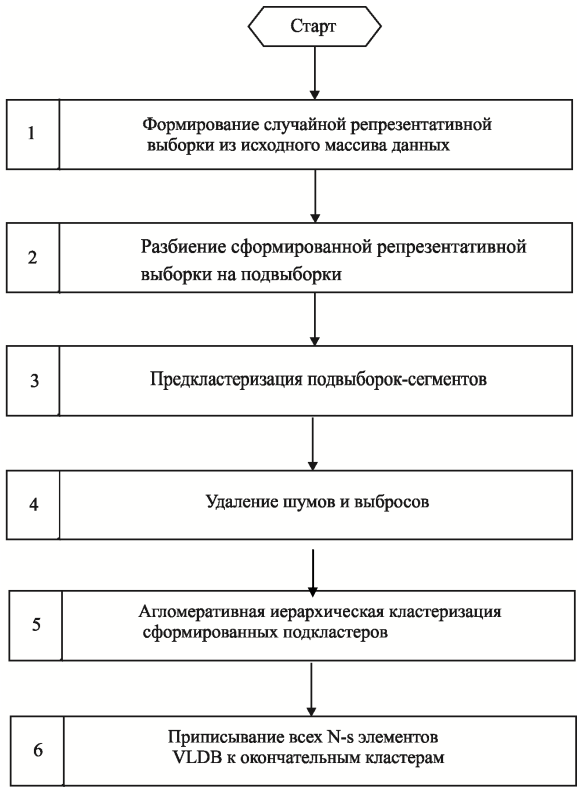

Рис. 2. Матричный метод CURE

Заключение

Рассмотрена задача кластеризации больших массивов изображений на основе иерархического агломеративного подхода. Введены матричные модификации соответствующих методов, позволяющие обрабатывать изображения без использования операций векторизации-девекторизации.

Предлагаемые процедуры кластеризации просты в численной реализации, не требуют многократного просмотра обрабатываемого массива, обеспечивают последовательную обработку поступающей информации, формируя кластеры произ-

вольной формы в условиях воздействия интенсивных возмущений.

Список литературы

1. Han J. Data Mining: Concepts and Techniques. – 2-nd ed. / J. Han, M. Kamber. – San Francisco: Morgan Kaufmann, 2006. – 800 p.

2. Gan G. Data Clustering: Theory, Algorithms, and Applications / G. Gan, C. Ma, J. Wu. – Philadelphia: SIAM, 2007. – 466 p.

3. Abonyi J. Cluster Analysis for Data Mining and System Identification / J. Abonyi, B. Feil. – Basel: Birkhäuser, 2007. – 303 p.

4. Olson D.L. Advanced Data Mining Techniques / D.L. Olson, D. Dursun. – Berlin: Springer, 2008. – 180 p.

5. Xu R. Clustering / R. Xu, D.C. Wunsch. – Hoboken: John Wiley&Sons, 2008. – 358 p.

6. Kaufman L. Finding Groups in Data: An Introduction to Cluster Analysis / L. Kaufman, P.J. Rousseeuw. – N.Y.: John Wiley&Sons, 1990. – 342 p.

7. Zhang T. BIRCH: An efficient data clustering method for very large databases / T. Zhang, R. Ramakrishnan, M. Livny // Proc. of the ACM SIGMOD Conf. on Management of Data. – Montreal: ACM Press, 1996. – P. 103-114/

8. Yuha S. CURE: am efficient clustering algorithm for large databases / S. Yuha, R. Rastogi, K. Shim // Information Systems. – 2001. – 26. – № 1. – P. 35-58/

9. Jain A.K. Algorithms for Clustering Data / A.K. Jain, R.C. Dubes. – Emglewood Cliffs, N. J.: Prentice Hall, 1988. – 318 p.

10. Motwani R. Randomized algorithms / R. Motwani, P. Raghavan. – Cambridge: CUP, 1995. – 420 p.

11. Ng R.T. Efficient and effective clustering methods for spatial data mining / R.T. Ng, J. Han // Proc. 20th Int. Conf. on Very Large Data Bases. – Santiago, Chile, 1994. – P. 144-145.

Поступила в редколлегию 10.09.2014

Рецензент: д-р техн. наук, проф. Е.В. Бодянский, Харьковский национальный университет радиоэлектроники, Харьков.

ІЕРАРХІЧНА АГЛОМЕРАТИВНА КЛАСТЕРИЗАЦІЯ ЗОБРАЖЕНЬ В ВЕЛИКИХ БАЗАХ ДАНИХ

С.І. Богучарський, О.Г. Каграманян, С.В. Машталір

Робота присвячена задачі аналізу великих баз даних зображень з точки зору розв'язання задачі інтелектуального пошуку. Для розв'язання цієї задачі зручно представити всі множину зображень у вигляді деяких класів у міру їх подібності. Для цього можливо застосувати апарат кластерного аналізу, методи якого і розглянуті в даній роботі. При цьому запропоновані дві матричні модифікації відомих підходів, що дозволяють спростити аналіз зображень за рахунок виключення операцій векторизації-девекторізації вихідних даних.

Ключові слова: *зображення, матрична ієрархічна кластеризація, вибірка, кластер.*

HIERARCHICAL AGGLOMERATIVE CLUSTERING IMAGES IN LARGE DATABASES

S.I. Bogucharskii, A.G. Kagramanyan, S.V. Mashtalir

Work is devoted to the problem of images large databases analysis in terms of solving the problem of intelligent search. To solve this problem it is convenient to introduce the set of all images in the form of some classes as their similarities. For this unit may apply cluster analysis methods and which are discussed in this paper. In this matrix proposed two modifications of known approaches to simplify the image analysis by eliminating operations vectoring-devektoring of source data.

Keywords: *images, matrix hierarchical clustering, sample, cluster.*