УДК 516.6

С. И. Богучарский[1], С. В. Машталир[2]
[1] ХНУРЭ, г. Харьков, Украина
[2] ХНУРЭ, г. Харьков, Украина, mashtalir_s@kture.kharkov.ua;

КЛАСТЕРИЗАЦИЯ КОЛЛЕКЦИЙ ИЗОБРАЖЕНИЙ В БОЛЬШИХ БАЗАХ ДАННЫХ НА ОСНОВЕ РЕКУРРЕНТНОЙ ОПТИМИЗАЦИИ

В данной работе рассмотрены методы кластеризации больших объемов данных и предлагается модификация подхода кластеризации мультимедийных объектов с возмущениями, основанного на плотности. Проведен анализ существующего метода DENCLUE, и введена матричная функция влияния, что позволяет эффективно использовать данный подход при анализе многомерных объектов, в частности, коллекций изображений, видео и мультимедиа данных. Введенная матричная форма позволяет повысить быстродействие кластеризации за счет отсутствия векторизации-девекторизации исходных данных.

КЛАСТЕРИЗАЦИЯ, БАЗЫ ДАННЫХ ИЗОБРАЖЕНИЙ, DENCLUE, ФУНКЦИЯ ВЛИЯНИЯ

Введение

Задача кластеризации массивов многомерных наблюдений часто встречается во множестве реальных приложений, а для ее решения на сегодня разработано большое количество методов, процедур, алгоритмов [1-6] от сугубо эмпирических до строго математических. В наиболее общей постановке предполагается, что имеется группа из N объектов, описываемых n-мерными векторами-признаками $x(k) \in R^n$, $k = 1, 2, ..., N$, которую необходимо разбить на p кластеров, при этом это число может быть заранее неизвестно, т.е. $1 < p < N$. Понятно, что столь большое количество возможных подходов к решению задачи связано с тем, что принципиально не существует универсального алгоритма, пригодного для эффективного использования во всех ситуациях, возникающих в реальных задачах.

Особую группу методов кластеризации образуют алгоритмы, предназначенные для обработки информации, хранящейся в сверхбольших базах данных (VLDB) [2, 5], где на первый план выходят быстродействие и простота численной реализации. В данной ситуации в качестве достаточно эффективных показали себя методы кластеризации, основанные на плотности распределения данных, при этом применяемое здесь понятие плотности по смыслу близко к плотности распределения, используемой в теории вероятностей и математической статистике. Именно методы, основанные на плотности, позволяют формировать кластеры произвольной формы в условиях, когда обрабатываемые данные искажены возмущениями, а само число кластеров p заранее неизвестно. В рамках «плотностного» подхода под кластерами понимают области в n-мерном пространстве признаков с высоким уровнем концентрации данных. Эти области разделены участками с низкой плотностью и именно здесь располагаются возмущения. Таким образом, алгоритмы, основанные на понятии плотности, в процессе обработки данных формируют области произвольной формы, где данные наиболее густо сконцентрированы.

Таким образом, целью данной работы является анализ методов кластеризации на основе плотностей и разработка модификации метода кластеризации.

1. Существующие методы кластеризация сверхбольших объемов данных

Наиболее распространенным методом из этого класса является DBSCAN (Density-Based Spatial Clustering of Applications with Noise), отличающийся вычислительной простотой и устойчивостью к возмущениям [7]. В основе метода лежит ряд понятий и определений, основными из которых являются внутренние и граничные точки, D-досягаемость (Density reachability) и D-связность (D-connectedness), порог ($\varepsilon = Eps$) и минимальное количество наблюдений в кластере (MinPts). При этом полагается, что произвольная точка $x(t)$ непосредственно досягаема из любой точки $x(q)$, если она удалена в смысле принятой метрики (традиционно евклидовой) на расстояние, не превышающее заданного априори порога ε. Именно $\varepsilon = Eps$ является основным исходным параметром алгоритма, задаваемым пользователем, при этом подразумевается, что этот пользователь является квалифицированным специалистом в конкретной предметной области, в рассматриваемом нами случае – специалистом в области видеообработки и компьютерных наук. На основании выбранного порога формируется ε-окрестность точки $x(q)$, которую образуют все точки x, удовлетворяющие неравенству

$$\|x - x(q)\| < \varepsilon .$$

Что же касается минимального количества наблюдений в кластере MinPts, то это также параметр, выбираемый экспериментально, обычно $N > MinPts \geq N+1$, при этом утверждается, что если в ε-окрестности точки $x(q)$ содержится не менее MinPts точек, то $x(k)$ и $x(q)$ относятся к одному кластеру.

Если рассмотреть понятие D-досягаемости, то точка $x(k)$ является D-досягаемой из $x(q)$, если

может быть сформирована «цепочка» наблюдений $x(q),...,x(r),...x(k)$ такая, что каждый ее элемент $x(r)$ непосредственно досягаем своими соседями.

Важным фактом является то, что понятие D-досягаемости не является симметричным. Если $x(k)$ лежит на границе кластера, то симметрия нарушается, т.е. эта точка может содержать меньше, чем MinPts точек в своей окрестности. Именно нахождением таких граничных точек и завершается формирование кластеров. Понятно, что в этом случае априорно предполагается, что формируемые кластеры не пересекаются. Все наблюдения, принадлежащие конкретному кластеру и имеющие не менее MinPts наблюдений в своей ε-окрестности, называются внутренними точками кластера. Описанная асимметрия порождает понятие D-связности, при этом точки $x(k)$ и $x(q)$ называются D-связными, если они обе досягаемы из $x(r)$, при этом очевидно, что понятие D-связности – симметрично.

Опираясь на введенные понятия, можно определить кластер как множество D-связных точек, причем, что важно, такую формулировку можно распространить и на другие подходы к задаче кластеризации, где используется понятие метрики. Сам же процесс кластеризации может быть сведен к последовательности элементарных действий, которая, стартуя из произвольной точки, находит множество D-связных с ней данных. После того как все такие наблюдения найдены, процедура запускается вновь из произвольной прежде непроанализированной точки и находит все относящиеся к ней D-связные данные. Так происходит до исчерпания всех N наблюдений анализируемой группы объектов-образов. Множество всех объектов, не включенных ни в один кластер и содержащих менее, чем MinPts наблюдений в своей ε-окрестности, в рамках стандартного подхода рассматриваются как шумы, хотя может оказаться, что именно в этих точках содержится уникальная информация, которая должна быть тщательно проанализирована вне рамок DBSCAN.

Следует отметить, что метод DBSCAN в силу своей простоты и наглядности получил широкое распространение во множестве прикладных задач анализа данных, в том числе и для сегментации различного рода изображений, где в соответствие каждому пикселю ставится многомерный набор признаков, задаваемый в векторной форме. Понятно, что количество таких векторов в выборке может быть очень велико. Конечно, в рассмотрение могут быть введены некоторые дополнительные характеристики анализируемого изображения, однако, во всех случаях для успешного решения задачи квалификация пользователя должна быть достаточно высока. Именно это обстоятельство, а также низкий уровень формализации этого метода и чувствительность к выбору параметров алгоритма породили целый ряд модификаций, лишенных некоторых недостатков прототипа.

На сегодня известен целый ряд модификаций, причем каждая новая из них стремилась минимизировать влияние субъективного фактора, связанного с каждым конкретным пользователем, и дополнительно формализовать базовую процедуру.

Одной из таких модификаций является DBCLASD (Distribution-Based Clustering of LArge Spatial Databases) [8], с помощью которой также можно формировать кластеры произвольной формы из «зашумленных» данных. Основным достоинством DBCLASD является возможность обработки данных в последовательном (on line) режиме, при этом каждый вновь поступающий на обработку образ может быть отнесен к тому или иному кластеру на основе анализа распределений расстояний от анализируемого образа до каждого из кластеров, основанных на χ^2-тесте. Данный метод обладает пониженной чувствительностью к выбору параметров Eps и MinPts, однако, в его основе лежит предположение, что данные в каждом кластере подчинены равномерному закону распределения, что далеко не всегда бывает в реальных задачах, особенно связанных с обработкой изображений.

Развитием DBSCAN также является алгоритм OPTICS (Ordering Points to Identify the Clustering Structure) [9], позволяющий решать задачи кластеризации в условиях, когда кластеры имеют не только различную форму, но и разную плотность распределения данных в каждом классе. OPTICS кроме основных понятий и определений, используемых в DBSCAN, вводит дополнительные характеристики для каждого наблюдения такие, как внутреннее расстояние (core distance) и расстояние досягаемости (reachability distance). OPTICS структурно эквивалентен DBSCAN, обладает расширенными функциональными возможностями, однако с вычислительной точки зрения значительно сложнее и медленнее прототипа, что усложняет его использование в задачах, связанных с VLDB.

Интересным гибридом DBSCAN и популярного метода k-средних является Bridge [10], с помощью которого исходный массив данных сначала обрабатывается с помощью стандартного метода k-средних, а затем к каждой сформированной группе данных применяется DBSCAN, подавляющий шумы и восстанавливающий плотность распределения данных в каждом кластере. Понятно, что Bridge с вычислительной точки зрения сложнее, чем DBSCAN, однако в настоящее время он используется для решения ряда задач, связанных с VLDB [2].

2. Кластеризация на основе плотности

Наиболее формализованным и математически обоснованным алгоритмом, основанным на плотности, является DENCLUE (DENsity-based CLUstEring) [11], созданный для обработки

больших массивов мультимедийных данных, формируя кластеры произвольной формы при высоком уровне шумов. Данный метод основан на ряде предположений:

– влияние каждого вектора-образа на соседние наблюдения формально может быть описано с помощью некоторой функции, обычно ядерной, называемой функцией влияния, которая описывает взаимосвязь всех наблюдений в некоторой окрестности данного образа;

– общая плотность распределения данных в n-мерном пространстве признаков формально описывается как сумма функций влияния каждого наблюдения;

– кластеры определяются как окрестности аттракторов плотности (D-аттракторов), являющихся, по сути, локальными максимумами общей функции плотности распределения данных.

Для некоторой произвольной точки в пространстве признаков y ее влияние на образ x может быть описано с помощью функции влияния

$$f^y(x) = f(x, y),$$

при этом наиболее часто в качестве таких функций используется либо прямоугольная конструкция

$$f(x, y) = \begin{cases} 0, \text{если } D(x, y) > \sigma; \\ 1, \text{в противном случае,} \end{cases} \quad (1)$$

либо гауссиан

$$f(x, y) = \exp\left(-\frac{D^2(x, y)}{2\sigma^2}\right), \quad (2)$$

где σ – параметр ширины ядерной функции, $D(x, y)$ – расстояние, обычно евклидово, между точками x и y.

Тогда для множества наблюдений $X = \{x(1), x(2), x(k), ..., x(N)\}$ общая функция плотности может быть представлена в виде

$$f^x(x) = \sum_{k=1}^{N} f(x, x(k)). \quad (3)$$

Функция (3) является суммой ядерных функций, характеризуется наличием множества локальных экстремумов-максимумов, именуемых D-аттракторами, каждый из которых представляет отдельный кластер и может быть определен с помощью тех или иных оптимизационных процедур. Здесь же заметим, что использование функции влияния (1) превращает DENCLUE в стандартный DBSCAN, а если $f(x, y)$ непрерывна и дифференцируема, как например (2), для нахождения локальных максимумов может быть использована стандартная градиентная оптимизация. При этом произвольная точка x притягивается к D-аттрактору x^*, если последовательность итераций

$$x^i = x^{i-1} + \eta \frac{\nabla f^x(x^{i-1})}{\left\| \nabla f^x(x^{i-1}) \right\|}; i = 1, 2, ...; x^0 = -x \quad (4)$$

сходится к x^*.

Если в качестве $f^x(x)$ используются соотношения (2), (3), то $\nabla f^x(x) = \sum_{k=1}^{N} (x(k) - x) f(x, x(k))$, а процедура (4) приобретает вид

$$x^i = x^{i-1} + \eta \frac{\sum_{k=1}^{N} (x(k) - x^{i-1}) f^{x^i}, x(k))}{\left\| \sum_{k=1}^{N} (x(k) - x^{i-1}) f^{x^i}, x(k)) \right\|}, \quad (5)$$

где η – параметр шага поиска.

Каждый из D-аттракторов характеризуется собственной функцией плотности

$$\hat{f}^{x^*}(x) = \sum_{x(k) \in near\, x^*} f(x, x(k)), \quad (6)$$

где $near\, x^* = \{x(k) : D(x^*, x(k)) \le \sigma_{near}\}$, а ее экстремум определяет координаты центроида кластера.

Конечно, с вычислительной точки зрения DENCLA сложнее любого из описанных выше алгоритмов, однако к его преимуществам следует отнести высокий уровень формализации, а также то, что он обобщает рассмотренные выше процедуры кластеризации, основанные на плотности.

3. DENCLUE в задачах кластеризации коллекций изображений

При решении задач кластеризации всегда предполагается, что каждое многомерное наблюдение-образ описывается n-мерным вектором $x(k)$, а весь процесс решения связан именно с векторными операциями. В ситуации, когда имеется большая коллекция изображений, подлежащих кластеризации, каждое двумерное изображение сначала должно быть подвергнуто векторизации, далее решается задача кластеризации, а ее результат подвергается девекторизации, переводящей векторное описание в матричную форму. Существенно упростить процесс кластеризации массивов можно, не переводя их в векторную форму, а непосредственно, оперируя с матрицами. Таким образом, набором исходных образов является набор матриц $x(k) = \{x_{i_1 i_2}(k)\}$, $x_1 = 1, 2..., m$; $x_2 = 1, 2..., n$; $k = 1, 2..., N$, $x(k) \in R^{m \times n}$. Далее, вводя вместо стандартной векторной евклидовой нормы ее сферический матричный аналог

$$D_S^2(x, y) = Sp(x - y)(x - y)^T,$$

можно ввести матричную функцию влияния

$$f_S(x, y) = \exp\left(-\frac{D_S^2(x, y)}{2\sigma^2}\right) = \exp\left(-\frac{Sp(x - y)(x - y)^T}{2\sigma^2}\right) \quad (7)$$

и матричную функцию плотности

$$f_S^x(x) = \sum_{k=1}^{N} f_S(x, x(k)).$$

При этом произвольная $(m \times n)$ матрица-образ x притягивается к матричному D-аттрактору x^*, если последовательность итераций типа (4)

$$x^i = x^{i+1} + \eta \frac{\left\{\frac{\partial f_S^x(x^{i-1})}{\partial x_{i_1 i_2}}\right\}}{\left(Sp\left\{\frac{\partial f_S^x(x^{i-1})}{\partial x_{i_1 i_2}}\right\}\left\{\frac{\partial f_S^x(x^{i-1})}{\partial x_{i_1 i_2}}\right\}^T\right)^{\frac{1}{2}}},$$

$$i = 1, 2, \ldots ; \; x^0 = x \qquad (8)$$

сходится к x^*. Здесь $\left\{\frac{\partial f_S^x(x)}{\partial x_{i_1 i_2}}\right\}$ – $(m \times n)$ матрица, образованная производными $f_S^x(x)$ по компонентам матрицы x.

Если в качестве матричной функции используется выражение (7), алгоритм оптимизации (8) может быть переписан в простой форме

$$x^i = x^{i-1} + \eta \frac{\sum_{k=1}^{N} r(k, i-1)}{\left(Sp(\sum_{k=1}^{N} r(k, i-1)(\sum_{k=1}^{N} r(k, i-1))^T\right)^{\frac{1}{2}}},$$

где $r(k, i-1) = (x(k) - x^{i-1}) f_s(x^{i-1}, x(k))$. Отметим, что по сути это расширение (5) на матричный случай.

Использование вместо векторного описания его матричного аналога позволяет существенного повысить быстродействие процесса обработки информации и избежать ряда проблем, возникающих в задаче кластеризации данных, описываемых векторами высокой размерности, что в свою очередь позволяет проводить обработку не только баз данных изображений, но и решать задачи кластеризации видеоданных.

Выводы

В статье рассмотрен существующий метод кластеризации мультимедийных данных с высоким уровнем шумов. Введен матричный аналог метода кластеризации DENCLUE, предназначенный для обработки коллекций изображений, хранящихся в больших базах неструктурированных данных. Алгоритм достаточно прост в численной реализации и характеризуется повышенным быстродействием за счет отказа от реализации вспомогательных операций векторизации-девекторизации исходных изображений.

Список литературы: 1. *Han J., Kamber M.* Data Mining: Concepts and Techniques. – 2-nd ed. – San Francisco: Morgan Kaufmann, 2006. – 800 p. **2.** *Gan G., Ma C., Wu J.* Data Clustering: Theory, Algorithms, and Applications. – Philadelphia: SIAM, 2007. – 466 p. **3.** *Abonyi J., Feil B.* Cluster Analysis for Data Mining and System Identification. – Basel: Birkhduser, 2007. – 303 p. **4.** *Olson D.L., Dursun D.* Advanced Data Mining Techniques. – Berlin: Springer, 2008. – 180 p. **5.** *Xu R., Wunsch D.C.* Clustering. – Hoboken: John Wiley&Sons, 2008. – 358 p. **6.** *Kohonen T.* Self-Organizing Maps. – 1-st ed. – Berlin: Springer, 1995. – 501 p. **7.** *Ester M., Kriegel H.-P., Sander J., Xu X.* A density-based algorithm for discovering clusters in large spatial database with noise // Proc. Int. Conf. on Knowledge Discovery in Databases and Data Mining. – Portlend, Oregon: AAAIO Press, 1996. – P. 226-331. **8.** *Xu X., Ester M., Kriegel H., Sander J.* A distribution-based clustering algorithm for mining in large spatial databases // Proc. 14-th Int. Conf. in Data Clustering "ICDE'98" – Orlando FLA: IEEE Computer Society, 1998 – P. 324-331. **9.** *Ankerst M., Breunig M., Krilgel H., Sander J.* OPTICS: Ordering points to identify the clustering structure // Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data. – Philadelphia, PA, 1999 – P. 49-60. **10.** *Dash M.* "1+1>2": Merging distance and density based clustering // Proc. Int. Conf. on Database systems for Advanced Applications. – Hong Kong. AEEE Computer Society, 2001. – P. 30-33. **11.** *Hu H., Ester M., Sander A.* Distribution-based clustering algorithm for mining in large spatial databases // Proc. 14-th Int. Conf. on Data Clustering "ICDE'98". – Orlando: FLA AEEE Computer Society, 1998. – P. 324-331.

Поступила в редколлегию 30.01.2014

УДК 516.6

Кластеризація колекцій зображень в великих базах даних на базі рекурентної моделі оптимизації / С.І. Богучарський, С.В. Машталір // Біоніка інтелекту: наук.-техн. журнал. – 2014. – № 1 (82). – С. 43–46.

Розглянуто підходи до кластеризації мультимедійних об'єктів зі збуреннями, які базуються на щільностях. Проведено модифікацію існуючого методу DENCLUE, яка базується на введенні матричної функції впливу, що дозволяє ефективно використовувати даний підхід при аналізі багатовимірних об'єктів, зокрема, колекцій зображень, відео та мультимедійних. Введена матрична форма дозволяє додати в швидкодії кластеризації за рахунок відсутності векторизації-девекторизації вихідних даних.

Бібліогр.: 11 найм.

UDC 516.6

Image collections clustering in large databases based on recursive optimization / S.I. Bogucharskiy, S.V. Mashtalir // Bionics of Intelligence: Sci. Mag. – 2014. – № 1 (82). – P. 43–46.

Approaches to multimedia objects with noises clustering based on density are describes. DENCLUE modification, based on the introduce the matrix form influence function, which allows efficient use of this approach in the analysis of multi-dimensional objects , in particular, image collections, video and multimedia data. Introduced matrix form can improve performance of clustering due to the lack of source data vectorization-devectorization.

Refs.: 11 titles.