DOI: 10.26565/2304-6201-2020-47-01

УДК 004.932

Image collections clustering in large databases on the basis of recurrent optimization

S.I. Bogucharskyi

Bogucharskyi Sirhii *Candidate of Engineering Science, senior researcher*
V. N. Karazin Kharkiv National University, 4, Svobody Sq., Kharkiv, 61022
e-mail: sbogucharskiy@karazin.ua
https://orcid.org/0000-0003-4971-4314

The following paper considers methods for clustering large amounts of data and proposes a modification of the density-based approach to clustering multimedia objects with disturbance. The analysis of the existing DENCLUE method is carried out, and the matrix influence function is introduced, which makes it possible to effectively use this approach in the analysis of multidimensional objects, the collections of images, video and multimedia data in particular. The introduced matrix form makes it possible to increase the speed of clustering due to the absence of vectorization-devectorization of the initial data.

Keywords*: *clustering, image databases, DENCLUE, influence function.

Кластеризація колекцій зображень у великих базах даних на основі рекурентної оптимізації

С.І. Богучарський

Богучарський Сергій Іванович *Кандидат технічних наук, старший науковий співробітник*
Харківський національний університет імені В.Н.Каразіна
площа Свободи, 4, місто Харків, 61022, Україна

В даній роботі розглянуті методи кластерізації великих об’ємів даних та пропонується модифікація підходу кластеризації мультимедійних об’єктів з збуреннями, заснованого на щільності. Проведено аналіз існуючого метода DENCLUE, та запропонована матрична функція впливу, що дозволяє ефективно використовувати зданий підхід при аналізі багатовимірних об’єктів, в частині, колекцій зображень, відео та мультимедіа даних. Впроваджена матрична форма дозволяє підвищити швидкодію клатеризації за рахунок відсутності векторизації-девекторизації вихідних даних.

Дотепер обробка відео викликає цілу низку труднощів, що пов’язані насамперед із розмаїттям тематики, якості та умовами зйомки, для яких неможливо підібрати уніфіковану процедуру розпізнання. Запропонований авторами підхід до обробки відеоданих дозволив виконати скорочення відеороликів та вилучення значущих кадрів, які мають назву ключових кадрів, з урахуванням контенту. Пошук ключових кадрів реалізовано за допомогою математичного апарату діаграм Вороного, які раніше використовувались тільки у сфері геодезії, матеріалознавстві та комп’ютерній графіці для тривимірного моделювання. У статті вирішуються важливі питання щодо пошуку опорних точок (за якими будуються діаграми Вороного) та покращення місця їх розміщення у кольорових зображеннях, якими є відеокадри. Порівняння відеокадрів за допомогою відповідних до них діаграм Вороного надало можливість отримати машинне уявлення про переміщення об’єктів зйомки у просторі та часі.

В статті розглянуто існуючий метод кластеризації мультимедійних даних з підвищеним рівнем шумів. Запропоновано матричний аналог метода кластеризації DENCLUE, призначений для обробки колекцій зображень, збереження у великих базах неструктурованих даних. Алгоритм достатньо простий у чисельній реалізації та характеризується збільшеною швидкодією за рахунок відмови від реалізації допоміжних операцій векторизації-девекторизації вхідних зображень.

Ключові слова: *кластеризація, бази даних зображень, DENCLUE, функція впливу, матричний аналог, відеокадр.*

Кластеризация коллекций изображений в больших базах данных на основе рекуррентной оптимизации

С. И. Богучарский

Богучарский Сергей Иванович *Кандидат технічних наук, старший научний сотрудник*
Харковский национальний университет имени В.Н.Каразина
площадь Свободы, 4, город Харьков, 61022, Украина

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.

Ключевые слова: *кластеризация, базы данных изображений, DENCLUE, функция влияния.*

1. Introduction

The problem of clustering arrays of multidimensional observations is often encountered nowadays and a large number of methods, procedures and algorithms, ranging from purely empirical to strictly mathematical, have been developed for its solution [1-6].

In the most cases, it is assumed that there is a group of N objects described by n - dimensional feature vectors $x(k) \in R^n$, $k = 1, 2, ..., \mathrm{N}$ which must be divided into p clusters, while this number may be unknown in advance, i.e. $1 < p < N$.

Due to the fact that there is no universal algorithm suitable for all possible situations it becomes clear that there is a large number of possible approaches to solving this problem.

A special group of clustering methods is formed by the algorithms designed to process information stored in very-large databases (VLDB) [2, 5], where speed and simplicity of numerical implementation come to the forefront.

In this situation, clustering methods based on the density of data distribution have proven themselves to be quite effective, while the concept of density used here is close in meaning to the distribution density used in probability theory and mathematical statistics. It is the density-based methods that make it possible to form clusters of arbitrary shape when the processed data are distorted by perturbations, and the number of clusters p is not known in advance. Within the framework of the «density» approach, clusters are understood as areas in the n-dimensional space of features with a high level of data concentration. These areas are separated by areas with low density and it is here that the disturbances are located.

Thus, algorithms based on the concept of density, in the process of data processing, form areas of arbitrary shape, where the data is concentrated most densely.

The purpose of this work is to analyze density-based clustering methods and develop a modification of the clustering method.

2. Existing methods of clustering extremely large amounts of data

The most common method from this class is DBSCAN (Density-Based Spatial Clustering of Applications with Noise), which is computationally simple and resistant to disturbances [7]. The method is based on a number of concepts and definitions, the main of which are internal and boundary points, D-reachability (Density reachability) and D-connectivity (D-connectedness), threshold ($\varepsilon = Eps$) and the minimum number of observations in a cluster (MinPts). In this case, it is assumed that an arbitrary point is directly reachable from any point $x(q)$ if it is removed in the sense of the accepted metric (traditionally Euclidean) by a distance not exceeding the threshold ε which is set a priori. The threshold $\varepsilon = Eps$ is the main initial parameter of the algorithm set by the user, which is assumed to be a qualified specialist in a specific subject area, in our case, in the field of video processing and computer science.

Based on the selected threshold, the ε-neighborhood of the $x(q)$ point is formed, which consists of all the points that satisfy the inequation

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

As for the minimum number of observations in the MinPts cluster, this is a parameter chosen experimentally, usually $\mathrm{N>MinPts} \geq \mathrm{N+1}$, it is argued that if the ε-neighborhood of a $x(q)$ point contains at least MinPts points, then both $x(k)$ and $x(q)$ belong to the same cluster.

If we consider the concept of D-reachability, then a $x(k)$ point is considered as D-reachable from $x(q)$ if such «chain» of observations can be formed that each of its elements is directly reachable by its neighbors.

An important factor is that the concept of D-reach is not considered to be symmetric. If $x(k)$ lies on the cluster boundary, then the symmetry is broken, i.e. this point may contain fewer than MinPts of points in its neighborhood.

It is precisely by finding such boundary points that the formation of clusters is completed. It is clear that in this case it is a priori assumed that the clusters being formed do not intersect. All observations belonging to a specific cluster and having at least MinPts of observations in their neighborhood are called internal points of the cluster. The described asymmetry gives rise to the concept of D-connection, and points $x(k)$ and $x(q)$ are called D-connected if they are both reachable from $x(r)$, and it is obvious that the concept of D-connection is symmetric.

Based on the introduced concepts, it is possible to define a cluster as a set of D-connected points, and, what is important, this formulation can be extended to other approaches to the clustering problem, where the concept of a metric is used. The clustering process itself can be reduced to a sequence of elementary actions, which, starting from an arbitrary point, finds a set of D-connected data. After all such observations are found, the procedure starts again from an arbitrary previously unanalyzed point and finds all the D-connected data related to it. This happens until all observations of the analyzed group of image objects are exhausted. The set of all objects that are not included in any cluster and contain less than MinPts observations in their neighborhood are treated as noise in the framework of the standard approach, although it may turn out that these points contain unique information that should be carefully analyzed outside of the DBSCAN scope.

It should be noted that the DBSCAN method, due to its simplicity and clarity, has become widespread in many applied problems of data analysis, including segmentation of various kinds of images, where a multidimensional set of features specified in vector form is assigned to each pixel. It is clear that the number of such vectors in the sample can be very large. Of course, some additional characteristics of the analyzed image can be introduced into consideration, however, to successfully solve the problem the user's qualification must be high enough. It is this circumstance, as well as the low level of formalization of this method and the sensitivity to the choice of the algorithm parameters, that gave rise to a number of modifications, devoid of some of the disadvantages of the prototype.

Today, a number of modifications are known, and each new of them sought to minimize the influence of the subjective factor associated with each specific user and additionally formalize the basic procedure.

One of such modifications is DBCLASD (Distribution-Based Clustering of Large Spatial Databases) [8], which can also be used to form clusters of arbitrary shape from “noisy” data. The main advantage of DBCLASD is the ability to process data in a sequential (on-line) mode, while each newly received image can be assigned to one or another cluster based on the analysis of the distributions of distances from the analyzed image to each of the clusters based on the χ^2-test. This method has a reduced sensitivity to the choice of the Eps and MinPts parameters, however, it is based on the assumption that the data in each cluster are subject to a uniform distribution law, which is not always the case in real problems, especially those related to image processing.

The development of DBSCAN is also the OPTICS (Ordering Points to Identify the Clustering Structure) algorithm [9], which allows solving clustering problems in conditions when the clusters have not only different shapes, but also different data distribution densities in each class. OPTICS, in addition to the basic concepts and definitions used in DBSCAN, introduces additional characteristics for each observation such as core distance and reachability distance. OPTICS is structurally equivalent to DBSCAN, has advanced functionality, but from a computational point of view, it is much more complex and slower than the prototype, which complicates its use in tasks related to VLDB.

An interesting hybrid of DBSCAN and the popular averages method is Bridge [10], with the help of which the original data array is first processed using the standard averages method, and then DBSCAN is applied to each formed data group, which suppresses noise and restores the data density in each cluster. It is clear that Bridge from a computational point of view is more complex than DBSCAN, however, it is currently used to solve a number of problems related to VLDB [2].

3. Clustering based on density

The most formalized and mathematically sound density-based algorithm is DENCLUE (DENsity-based CLUstEring) [11], created for processing large arrays of multimedia data, by forming clusters of arbitrary shape at a high noise level. This method is based on a number of assumptions:

1) the influence of each vector-image on neighboring observations can be formally described by using some function, usually a nuclear one, called the influence function, which describes the relationship of all observations in some neighborhood of the given image;

2) the general density of data distribution in the n-dimensional space of attributes is formally described as the sum of the influence functions of each observation;

3) clusters are defined as neighborhoods of density attractors (D-attractors), which are, in fact, local maxima of the general data distribution density function.

For some arbitrary point in the feature space, its influence on the image can be described by using the influence function $f^{y}(x) = f(x, y)$, moreover, such functions are most often either a rectangular structure (1)

$$f(x, y) = \begin{cases} 0, \textit{if } D(x, y) > \sigma; \\ 1, \textit{if} - \textit{otherwise}, \end{cases} \tag{1}$$

or Gaussian (2)

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

where σ is the parameter of the width of the nuclear function, is the distance, usually Euclidean, between the points *x* and *y*.
Then, for a set of observations, the general density function can be represented in the form (3)

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

Function (3), is the sum of nuclear functions, characterized by the presence of a set of local extrema-maxima, called D-attractors, each of which represents a separate cluster and can be determined by using one or another optimization procedure. Here we note that the use of the influence function (1) turns DENCLUE into a standard DBSCAN, and if $f(x, y)$ is continuous and differentiable, such as (2), the standard gradient optimization can be used to find the local maxima. In this case, an arbitrary point is attracted to the D-attractor if the sequence of iterations (4)

$$x^{i} = x^{i-1} + \eta \frac{\nabla f^{x}(x^{i-1})}{\left\| \nabla f^{x}(x^{i-1}) \right\|}; i=1, 2,...; x^{0} = -x \tag{5}$$

converges to x^*.

If $f^{x}(x)$ are used as relations (2), (3), then $\nabla f^{x}(x) = \sum_{k=1}^{N} (x(k)-x) f(x, x(k))$, and the procedure (4) takes the form (5)

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

where η is a search step parameter.

Each of the D-attractors is characterized by its own density function (6)

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

where $\text{near } x^* = \{x(k): D(x^*, x(k)) \leq \sigma_{near}\}$ and its extremum determines the coordinates of the cluster centroid.

Of course, from a computational point of view, DENCLA is more complex than any of the algorithms described above, however, its advantages include a high level of formalization, as well as the fact that it generalizes the density-based clustering procedures discussed above.

4. DENCLUE in tasks of clustering image collections

When solving clustering problems, it is always assumed that each multidimensional observation-image is described by a n-dimensional vector $\mathrm{x(k)}$, and the entire solution process is associated precisely with vector operations.

In a situation where there is a large collection of images to be clustered, each two-dimensional image must first be vectorized, then the clustering problem is solved, and its result is devectorized, which transforms the vector description into a matrix form. It is possible to significantly simplify the process of clustering arrays without converting them into a vector form, but operating directly with matrices. Thus, the set of initial images is the set of matrices $\mathrm{x(k)=\{x_{i_1 i_2}(k)\}}$, $x_1 = 1, 2..., m$; $x_2 = 1, 2..., n$; $k = 1, 2..., N$, $x(k) \in R^{m\times n}$.

Further, instead of the standard vector Euclidean norm its spherical matrix analogue is introduced (7)

$$\mathrm{D_S^2(x,y)=Sp(x\text{-}y)(x\text{-}y)^T}, \tag{7}$$

and the matrix density function

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

In this case, an arbitrary $\mathrm{(m\times n)}$ matrix-image x is attracted to the matrix D-attractor x^* if the sequence of iterations of type (4)

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

$$\mathrm{i=1, 2,...;\ x^0=x}$$

converges to x^*. Here $\left\{\frac{\partial \mathrm{f_S^x(x)}}{\partial \mathrm{x_{i_1 i_2}}}\right\}$ $\mathrm{(m\times n)}$ matrix, formed by derivatives $\mathrm{f_S^x(x)}$ with respect to the components of the matrix x.

If the expression (7) is used instead of the matrix function, the optimization algorithm (8) can be rewritten in a simple form

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

where $\mathrm{r(k,i\text{-}1)=(x(k)\text{-}x^{i-1})f_S(x^{i-1}, x(k))}$.

Note that this is essentially an extension of (5) to the matrix case.

The use of its matrix analogue instead of a vector description makes it possible to significantly increase the speed of information processing and avoid a number of issues arising in the problem of clustering data described by high-dimensional vectors, which in turn allows processing not only image databases, but also solving problems of clustering video data.

5. Conclusion

The article discusses the existing method for clustering multimedia data with a high level of noise. A matrix analogue of the DENCLUE clustering method is introduced, intended for processing image collections stored in large unstructured databases. The numerical implementation of the algorithm is quite simple and its performance is increased due to rejecting auxiliary operations of vectorization-devectorization of the original images.

REFERENCES

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: Birkhäuser, 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.

Clustering of image collections in large databases based on a recurrent optimization model / S.I. Bogucharskyi // V. N. Karazin Kharkiv National University.

Approaches to clustering multimedia objects with density-based perturbations are considered. A modification of the existing DENCLUE method has been made, which is based on the introduction of a matrix influence function, which allows using this approach effectively in the analysis of multidimensional objects, in particular, collections of images, video and multimedia. The introduced matrix form allows accelerating the speed of clustering due to absence of vectorization-devectorization of initial data. Refs., 11 titles.

Надійшла – 25.08.2020