Information Sciences 591 (2022) 103–127

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

On exploring data lakes by finding compact, isolated clusters

Patricia Jiménez [a,*], Juan C. Roldán [a], Rafael Corchuelo [a]

[a] Universidad de Sevilla, ETSI Informática, Avda. de la Reina Mercedes, s/n, Sevilla E-41012, Spain

ARTICLE INFO

Article history:
Received 17 February 2021
Received in revised form 31 October 2021
Accepted 11 December 2021
Available online 15 January 2022

Keywords:
Data lakes
Clustering
Meta-heuristics
Genetic algorithms

ABSTRACT

Data engineers are very interested in data lake technologies due to the incredible abundance of datasets. They typically use clustering to understand the structure of the datasets before applying other methods to infer knowledge from them. This article presents the first proposal that explores how to use a meta-heuristic to address the problem of multi-way single-subspace automatic clustering, which is very appropriate in the context of data lakes. It was confronted with five strong competitors that combine the state-of-the-art attribute selection proposal with three classical single-way clustering proposals, a recent quantum-inspired one, and a recent deep-learning one. The evaluation focused on exploring their ability to find compact and isolated clusterings as well as the extent to which such clusterings can be considered good classifications. The statistical analyses conducted on the experimental results prove that it ranks the first regarding effectiveness using six standard coefficients and it is very efficient in terms of CPU time, not to mention that it did not result in any degraded clusterings or timeouts. Summing up: this proposal contributes to the array of techniques that data engineers can use to explore their data lakes.

1. Introduction

The Web is currently the most important data source since it provides a plethora of datasets on virtually any topics. A data lake is a repository to which data engineers dump as many datasets as possible in an attempt not to miss any chances to infer new valuable knowledge [41,25,33,15]. It is then not surprising that many IT providers are competing to devise technologies that help data engineers work with their data lakes [22,4,34].

The key here is that data engineers are not expected to have any clues on the structure of the data; particularly, the data are not expected to be pre-classified. This precludes using supervised machine-learning methods directly on them and argues for using unsupervised methods on the hope to discover clusters of data that are compact, i.e., the data are similar within the clusters, and isolated, i.e., they are dissimilar amongst the clusters [18]. Every cluster is then a sub-group of data that may help data engineers understand the structure of the datasets in their data lakes. Once they are studied, they constitute the starting point to apply other machine-learning methods whose ultimate goal is to infer new knowledge [20]. That knowledge is expressed using models that capture the relationships amongst the attributes that flag a datum as belonging to one or another cluster. Much effort is currently being put on explaining the models so that data engineers can interpret them [17].

Ideally, a clustering technique must meet the following requirements in the context of data lakes: R1) it must be able to deal with business data, since typical data lakes provide datasets in which data are observations or aggregations of other data

* Corresponding author.
E-mail addresses: patriciajimenez@us.es (P. Jiménez), jcroldan@us.es (J.C. Roldán), corchu@us.es (R. Corchuelo).

https://doi.org/10.1016/j.ins.2021.12.045
0020-0255/

that cannot be assumed to be normally distributed or to have any spatial and/or temporal relationships; R2) it must be multi-way, which means that it must select a single subspace of informative attributes and cluster the original dataset in that subspace simultaneously, since single-way proposals perform these tasks independently and they are known to produce worse clusterings and multiple subspaces are often confusing for data engineers; R3) it must be able to deal with small- and high-dimensional data, since typical data lakes have datasets whose dimensionalities range from a few to thousands of attributes; R4) it must be automatic, that is, it must not require the user to provide the number of clusters manually, since the data engineer may not be assumed to know anything about the structure of the datasets; R5) it must not require per-dataset configuration, since typical data lakes provide far too many.

In the literature, there are many proposals that attempt to address the previous challenges [48,21,26,23,42,7,16,46,36,2,47,14,9]. Many of them use algorithmic approaches whose goal is to find a good enough solution, but they usually have trouble to deal with large or high-dimensional datasets [11]. Many other proposals use meta-heuristic approaches that map the problem onto nature-inspired processes. They are interesting insofar they can explore complex search spaces in parallel, which definitely contributes to both effectiveness and efficiency; it is then not surprising that they have found their way into many scientific and engineering problems [37,11]. Recently, some deep-learning approaches have been published and they have proven to be very effective when dealing with images, text, or sounds [35,24]. Unfortunately, meta-heuristics and deep learning are insufficiently explored; particularly, there are not any proposals to address multi-way single-subspace automatic clustering using meta-heuristics or the single- and multiple-subspace problems using deep learning.

This article presents RóMULO, which explores the research niche regarding using meta-heuristics. It was confronted with five strong competitors that integrate the state-of-the-art GSPPCA method [8] to find subspaces of informative attributes and several methods to perform the clustering, namely: Affinity-Propagation, Mean-Shift, and OPTICS-Xi, which are classical proposals, as well as PQC, which is a quantum-inspired proposal [12], and DCC, which is a deep-learning proposal [43]. The evaluation focused on two key points, namely: first, exploring their ability to find compact and isolated clusterings, which was evaluated on 46 real-world data lakes with a total of 2 561 datasets that provide 15 435 171 unclassified data; second, exploring the extent to which such clusterings can be considered good classifications, which was evaluated on five additional data lakes with a total of one hundred datasets that provide 1 200 552 pre-classified data. The comparison was performed in terms of both effectiveness and efficiency. Regarding effectiveness, six standard coefficients were computed, namely: Silhouette, Davies-Bouldin, Caliński-Harabasz, Adjusted Rand, Fowlkes-Mallows, and Accuracy. Regarding efficiency, the CPU time was used as the main measure. Additionally, the ratio of degraded clusterings and the ratio of timeouts were computed. The experimental results were studied using a statistically sound method, which confirmed that RóMULO ranks at the first position regarding the six effectiveness coefficients and it is very efficient for practical purposes; furthermore, it was the only proposal that did not result in any degraded clusterings or timeouts.

The rest of the article is organised as follows: Section 2 analyses the related work; Section 3 describes the details of the proposal; Section 4 presents the experimental analysis; finally, Section 5 concludes the article.

2. Related work

This section reports on the most closely-related work. First, it introduces a conceptual framework that homogenises the vocabulary in this area; then, the literature is reviewed along two axes: single-way versus multi-way proposals; finally, there is a discussion that makes it clear the motivation behind RóMULO.

2.1. Conceptual framework

A dataset is a set of the form $\{x_1, x_2, \ldots, x_n\}$, where each x_i is a d-dimensional vector of attributes $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$ $(n \geqslant 0, d \geqslant 1, 1 \leqslant i \leqslant n)$.

Clustering dataset X seeks to find k clusters $(1 \leqslant k \leqslant n)$ such that they are as compact (high intra-similarity) and isolated (low inter-similarity) as possible [10]. The problem can be formalised as the following optimisation problem:

$$\begin{array}{lll} \max & f(X, M, C) & \\ \text{st} & \sum_{j=1}^{k} M[i,j] > 0 & 1 \leqslant i \leqslant n \\ & 0 \leqslant M[i,j] \leqslant 1 & 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant k \\ & C[j] \in \mathbb{P}_1 X \times \mathbb{P}_1(A \times \mathbb{R}) & 1 \leqslant j \leqslant k \end{array}$$

where f denotes a fitness function, X denotes the input dataset, M denotes a membership matrix, C denotes a clustering, and A denotes a set of attributes. As usual, $\mathbb{P}_1$ denotes the non-empty powerset and $\mathbb{R}$ denotes the real numbers.

The fitness function helps assess how compact and isolated the clusters are. In the literature, there are many choices to implement it. They range from simple coefficients to multi-objective functions [47,5,19,30,27,31]. Some coefficients can be computed on the clusterings themselves because they focus on compactness and isolation only, but other coefficients require the clustering plus a ground truth because they also measure the extent to which the clustering can be used to classify the

data into a number of pre-defined classes. They all can be used to assess the same clustering from different, complementary perspectives, which argues for a means to combine them using multi-objective functions. The Lexicase method is such a multi-objective function and it was specifically designed to work in the context of genetic search strategies [19,27].

The membership matrix M has n rows and k columns; intuitively, $M[i,j]$ indicates if the i-th datum in X belongs to cluster $C[j]$ or not ($n = |X|, 1 \leqslant k \leqslant n, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant k$). A clustering proposal is hard if it assigns every datum to a single cluster ($\forall 1 \leqslant i \leqslant n : \exists_1 1 \leqslant j \leqslant k : M[i,j] = 1$) and overlapping otherwise ($\forall 1 \leqslant i \leqslant n : \exists 1 \leqslant j \leqslant k : M[i,j] > 0$). The latter can be crisp if the membership matrix is Boolean ($\forall 1 \leqslant i \leqslant n : \forall 1 \leqslant j \leqslant k : M[i,j] \in \{0,1\}$) or fuzzy if it is a likelihood matrix ($\forall 1 \leqslant i \leqslant n : \forall 1 \leqslant j \leqslant k : 0 \leqslant M[i,j] \leqslant 1$).

The components of clustering $C[j]$ are pairs of the form (X', A'), where X' denotes a non-empty subset of the input dataset X and A' is a non-empty subset of $A \times \mathbb{R}$ ($1 \leqslant j \leqslant k, 1 \leqslant k \leqslant n, n = |X|$). Intuitively, X' denotes the subset of data in a particular cluster and A' denotes a set of pairs (a_i, w_i) in which each a_i refers to an attribute and each w_i to a weight that represents how informative it is ($1 \leqslant i \leqslant d, d \geqslant 1$). A clustering proposal is single-way [23,47] if it clusters the data using all of the input attributes, i.e., $A' = \{(a_i, 1)\}_{i=1}^{d}$ ($d \geqslant 1$); it is multi-way [23,46,14,9] if it finds the subspace of most informative attributes and clusters the datasets simultaneously, i.e., $A' = \{(a_i, w_i)\}_{i=1}^{d}$ ($d \geqslant 1, 0 \leqslant w_i \leqslant 1$). The latter can be further subclassified as crisp or fuzzy depending on whether the weights are Boolean or real. The single-way proposals must necessarily rely on a pre-processor that selects the informative attributes [28,8], but Jain [23] found out that this typically results in worse clusterings. A proposal is single-subspace if all of the clusters refer to the same subspace of attributes; it is multiple-subspace if each cluster may refer to a different subset of attributes. A proposal is manual if the user must provide the number of clusters to find beforehand; it is automatic if it can find the number of clusters automatically.

Clustering is inherently complex from a computational point of view. A basic hard single-way manual proposal has to explore a search space whose size is the Stirling partition number [3,47], namely:

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^n.$$

If the number of clusters k is not known beforehand, then the search must be performed in a larger space whose size is the Bell number, namely:

$$B(n) = \sum_{k=0}^{n} S(n,k).$$

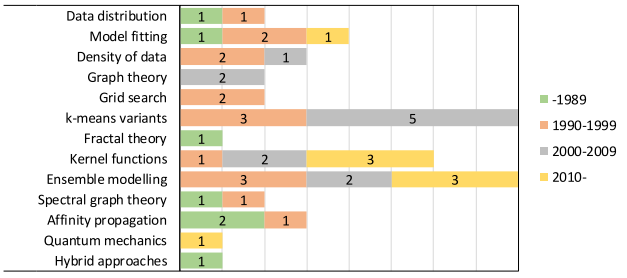
The search space grows by $O(2^d)$ in multiple-subspace clustering problems with regard to single-subspace problems because there are $2^d - 1$ non-empty subsets of attributes in a dataset with d-dimensional data ($d \geqslant 1$). Thus, in a data engineering context, it does not generally make sense to try to find the optimal solution to a clustering problem, but an approximate solution that is good enough for practical purposes.

2.2. Single-way clustering

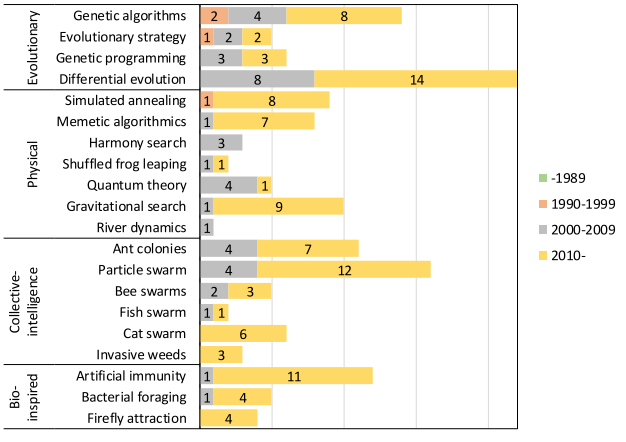
Originally, most approaches to single-way clustering were algorithmic, but meta-heuristic approaches have found their way into this field because their nature-inspired approaches help explore complex search spaces using multi-core and/or multi-threaded CPUs. Recently, some proposals that are based on deep learning have also found their way into clustering, with a focus on unstructured data like images, text, or sounds. Fig. 1 summarises the existing approaches; the labels on the vertical axis describe the approach and the bars show the total number of proposals grouped by several temporal periods.

The surveys by Jain [23] and Xu and Tian [47] focus on algorithmic approaches. The most common ones are the following: a) data distribution, which assumes that the clusters can be identified by finding sub-distributions of data; b) model fitting, which generalises the previous idea to arbitrary statistical models whose parameters are fit to the data; c) density of data, which assumes that clusters are high-density groups of data; d) graph theory, which assumes that the data are the nodes of a graph in which the edges represent distance-based relationships amongst them; e) grid search, which assumes that the data may be arranged in a grid in which the clusters are squared groups of data; f) k-means variants like bisecting k-means, kd-means, single-pass k-means, k-medoids, kernel k-means, sort-means, k-harmonic means, or x-means; g) fractal theory, which assumes that the clusters are subgroups that share some geometric properties with the whole dataset; h) kernel functions, which project the input data onto higher-dimensionality attribute spaces in which clusters can be made apart easily; i) ensemble modelling, which first generates candidate clusters using other techniques and then merges the results using consensus functions; j) spectral graph theory, which addresses the problem as a graph partitioning problem; k) affinity propagation, which computes the centroids of the clusters as the data with the highest affinity to the other data; l) quantum mechanics, which maps the clustering problem onto some phenomena that are formalised using Schrödinger's equation; and m) hybrid approaches that have managed to successfully combine two or more of the previous ones.

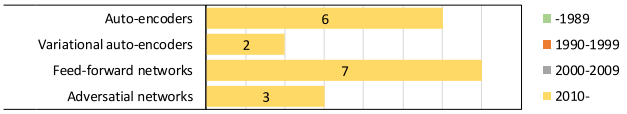
The surveys by Hruschka et al. [21], Bong and Rajeswari [7], Rana et al. [42], Alam et al. [2], Nanda and Panda [36], García and Gómez-Flores [14], and Figueiredo et al. [13] focus on meta-heuristic approaches. The most common ones are the following: a) evolutionary approaches, which are based on genetic algorithms, evolutionary strategies, genetic programming, or



a) Algorithmic approaches.



b) Meta-heuristic approaches.



c) Deep-learning approaches.

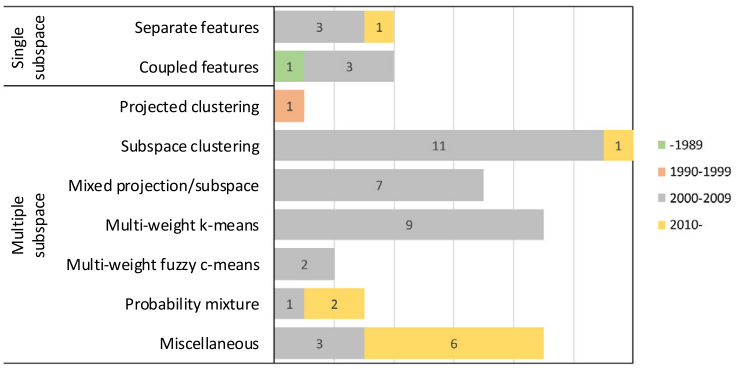
Fig. 1. Single-way clustering proposals.

differential evolution; b) physical approaches, which are based on simulated annealing, memetic algorithms, harmony search, shuffled frog leaping, quantum theory, gravitational search, or river dynamics; c) collective-intelligence approaches, which are based on ant colonies, particle swarms, bee swarms, fish swarms, cat swarms, or invasive weeds; and d) bio-inspired approaches, which are based on artificial immunity, bacterial foraging, or fireflies.

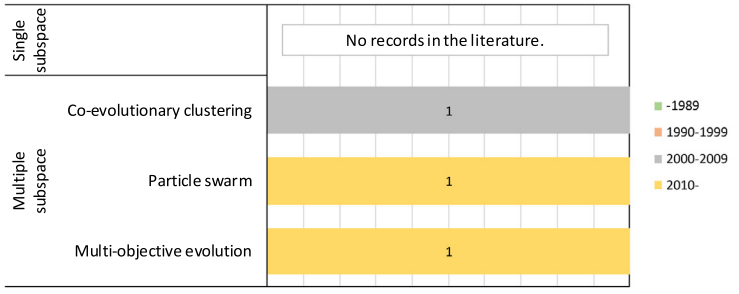
The surveys by Min et al. [35] and Karim et al. [24] focus on deep-learning approaches, which can be grouped as follows: a) proposals that use an auto-encoder to compute a latent attribute space on which clustering techniques like *k*-means have proven to work better than on the original datasets; b) proposals that use variational auto-encoders, which learn the distribution model of the input data so that it can be used to generate new synthetic data for learning purposes; c) proposals that rely on fully connected networks, convolutional networks, and deep-belief networks that must be trained using two complementary loss functions; and d) proposals that rely on min–max adversarial games between two neural networks, namely: a generative network that generates both similar and dissimilar synthetic data regarding a particular cluster and a discriminative network that learns to classify a new datum as belonging to that cluster or not.

2.3. Multi-way clustering

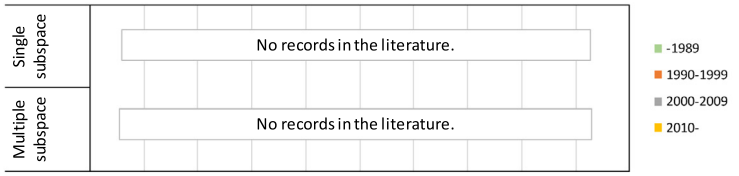
Jain [23], Sim et al. [46], García and Gómez-Flores [14], and Deng et al. [9] surveyed the existing multi-way clustering proposals, which can be subclassified according to whether they attempt to find the clusters in a single or multiple subspaces. Most proposals use algorithmic approaches, only a few use meta-heuristic approaches, and none uses a deep-learning approach, cf. Fig. 2.



a) Algorithmic approaches.



b) Meta-heuristic approaches.



c) Deep-learning approaches.

Fig. 2. Multi-way clustering proposals.

Regarding the algorithmic approaches to single-subspace clustering, there are only two: a) the separate-feature approach, which learns the weights of the attributes before finding the clusters, and b) the coupled-feature approach, which learns the weights of the attributes during the clustering process.

Regarding the multiple-subspace problem, there are a few more algorithmic approaches, namely: a) projected clustering, which relies on distance functions and/or data patterns; b) subspace clustering, which aims at finding all possible clusters in all possible subspaces; c) projection/subspace approaches, which somewhat combine the previous approaches; d) multi-weight *k*-means, e) multi-weight fuzzy *c*-means, and g) probability mixture, which are somewhat inspired by the classical single-way counterparts; finally, h) there are some miscellaneous approaches that address the three most important problems with the previous ones, namely [26,9]: they focus on cluster compactness and neglect cluster isolation, they are very sensitive to the configuration parameters, and they do not deal well with datasets with clusters of diverging sizes.

Regarding the meta-heuristic approaches to single-subspace clustering, no proposal was found in the literature. Regarding multiple-subspace clustering, there are three approaches, namely: a) co-evolutionary clustering, which leverages attribute weighting methods to deal with complex data and noisy and correlated attributes; b) particle swarm, which seeks for near-optimal variable weights for a given objective function; and c) multi-objective evolution, which benefits from both the merits of crisp subspace clustering and the good properties of the multi-objective optimisation-based approach for fuzzy clustering. Unfortunately, the previous approaches have stability problems that hinder applying them to many real-world datasets [26,9].

Realise that there are not any deep-learning approaches to multi-way clustering since such techniques are not intended to select any informative attributes, but to transform the original attributes into so-called latent attributes. They somewhat encode the original attributes by aggregating the results output by several neurons on the hope that other techniques, including clustering, become more effective and/or efficient when working on them. The latent attribute space is then not generally expected to be understood by a data engineer; contrarily, a subspace of the original attributes helps data engineers focus on the informative attributes that he or she needs to understand in order to grasp the structure of a dataset.

2.4. Discussion

The ideal clustering technique in the context of data lakes must meet the following requirements:

R1: Must be able to deal with business data. Typical data lakes provide business data that represent real-world entities or events by means of vectors with many attributes. They can be observations or aggregations of other data that cannot be assumed to be normally distributed or to have any spatial and/or temporal relationships. They may also include unstructured data like images, text, or sounds, which can be dealt with using specific-purpose proposals [39].
R2: Must perform multi-way clustering. Typical data lakes provide many attributes that are uninformative regarding the structure of the data and make it difficult for data engineers to find it. Single-way proposals require a pre-processor to find the most informative attributes and then cluster the dataset that results from projecting the original dataset onto that subspace; multi-way proposals find the subspace of attributes and cluster the projected datasets simultaneously, which has been proven to produce better results [23].
R3: Must deal with high-dimensional data. High data dimensionality is problematic because of two reasons [48,3,47,9]: on the one hand, many attributes might be uninformative and introduce noise that does not help the data engineer understand the structure of the data or the clustering algorithm to find good clusters; on the other hand, more attributes means more inefficiency and more chances to miss the important attributes and the relationships amongst them that help find good clusters [29].
R4: Must not require the user to provide the number of clusters. This means that the proposal must be able to guess the number of clusters in the input datasets automatically and not require the user to set it beforehand. This is the case of data lakes, since data engineers cannot be assumed to have any prior knowledge of the datasets in their data lakes, but their topic. Unfortunately, not knowing the number of clusters beforehand is known to increase the size of the search space exponentially [38,1], which is even more problematic in the context of data lakes as the size of the datasets increases.
R5: Must not require per-dataset configuration. Finding the appropriate values of the configuration parameters typically requires to perform grid search, which is an inherently costly and difficult procedure [29]. Furthermore, fine-tuning the configuration parameters may induce a particular clustering instead of finding the actual clusters in the dataset [46]. In the context of data lakes, there are typically too many datasets, which implies that fine-tuning the configuration parameters for each particular dataset is not generally a good idea.

The analysis of the literature in the previous sections reveals that there are many clustering proposals that might well meet the previous requirements. What shines in this analysis is that there are some clear research niches, cf. Figs. 1 and 2, namely: using meta-heuristics to address the multi-way single-subspace problem or deep-learning to address the multi-way single- or multiple-subspace problem. This article presents a proposal that explores the first research niche and meets the previous requirements.

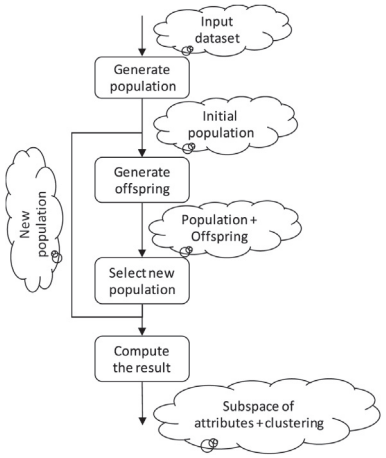
A quick reader might think that single-subspace clustering is not an actual problem since one might use a multiple-subspace technique and then select the best subspace. That is not possible due to a subtle, but very important difference between single- and multiple-subspace clustering. Single-subspace clustering refers to finding one subspace of attributes in which the best possible clustering can be computed; note that a proposal like RóMULO can find multiple subspaces and clusterings, but returns the best one, which is typically more than enough for practical purposes. What matters here is that the best subspace (or the suboptimal ones) results in complete clusterings of the original dataset. Contrarily, multiple-subspace clustering refers to finding multiple clusters in likely different subspaces of attributes; there is not generally a best subspace because all of the subspaces complement each other to produce a set of clusterings of the original dataset. Selecting just one of the subspaces returned leads to a partial clustering of the original dataset. The idea of returning multiple complementary clusterings in the context of data lakes is not that appealing since the ultimate goal is to help a data engineer understand his or her datasets and to infer new knowledge from them. The more compact and isolated the data returned and the less dimensions is generally the better because this helps focus on groups of similar data.

Finally, it is worth mentioning that a reviewer highlighted a possible connection between clustering in the context of data lakes and so-called collaborative fuzzy clustering [40,44]. This technique was devised to deal with multiple datasets that provide complementary data, but cannot be merged, e.g., because of legal or performance issues. Collaborative fuzzy clustering is a very good clustering technique in that context, but it is not generally applicable in the context of data lakes because of the following reasons: it assumes that the datasets have the same attributes, it does not compute a subspace of informative attributes, and it requires the number of clusters to be set beforehand.

3. The algorithm

Algorithm 1: The main method of RóMULO.

```
method RóMULO (X) returns (A, C)
    – Step 1: generate the initial population
    P := generatePopulation(X)
    – Step 2: evolve the population
    repeat NGEN times
        – Step 2.1: generate the offspring
        S := generateOffspring(P)
        – Step 2.2: select the new population
        P := lexicase(P ∪ S, ⌈MU PSIZE⌉)
    end
    – Step 3: compute the result
    {(k, A)} := lexicase(P, 1)
    X′ := project(X, A)
    C := cluster(X′, k)
end
```



This section introduces the algorithm behind RóMULO. Algorithm 1 sketches its main method and illustrates it using a graphical abstract. It works on an input dataset *X* and outputs a subspace of attributes *A* and a clustering *C*. The method performs three steps in sequence: computing the initial population, evolving it, and computing the results. For the sake of readability, the configuration parameters of RóMULO are introduced as constants that are denoted using uppercase, multi-letter identifiers.

Algorithm 2: Method to generate a population.

```
method generatePopulation(X) returns P
  n := |X|
  d := dim X
  P := ∅
  repeat PSIZE times
    k := uniform(2, ⌈n/2⌉)
    A := (bernoulli(0.50), ..., bernoulli(0.50))
    P := P ∪ {(k, A)}
  end
end
```

The first step consists in generating a population *P* with *PSIZE* individuals, cf. Algorithm 2. Generally speaking, the individuals are expected to encode the information that the meta-heuristic requires to make decisions regarding how to explore the search space. Since RóMULO outputs a number of clusters and a subspace of attributes, then the search space consists of tuples of the form (k, A), where k denotes the number of clusters and A is a Boolean vector that encodes the attributes that must be used to compute the clusters. The method first initialises n and d to the number of data and attributes in the input dataset X, respectively, and P to an empty set. It then iterates to generate a population with *PSIZE* individuals in set P. The individuals are generated with random numbers of clusters (which are drawn from a Uniform distribution in interval $[2, n/2]$, where n denotes the number of data in the input dataset), and random subspaces of attributes (which are drawn from a Bernoulli distribution with mean 0.50, i.e., every attribute has the same chances to be selected).

The second step consists in evolving the initial population *NGEN* times using a $(MU + LAMBDA)$ genetic strategy [6]. The evolution is performed by generating $\lceil LAMBDA\,PSIZE \rceil$ offsprings and then selecting the best $\lceil MU\,PSIZE \rceil$ individuals to create the new generation. Both *LAMBDA* and *MU* are introduced as percentages of the initial population size.

Algorithm 3: Method to generate the offspring.

```
method generateOffspring(P) returns
    S := ∅
  repeat ⌈LAMBDA PSIZE⌉ times
    if bernoulli(CXPB) then
      x, y := crossover(P)
      S := S ∪ {x, y}
    else if bernoulli(MUTPB) then
      x := mutate(P)
      S := S ∪ {x}
    else
      {x} := pick(P, 1)
      S := S ∪ {x}
    end
  end
end
```

Algorithm 3 shows the method that generates the offspring. It iterates a total of $\lceil LAMBDA\,PSIZE \rceil$ times and produces the offspring as follows: it first determines if two new offsprings must be generated by crossing two existing individuals, which happens according to a Bernoulli random variable with mean probability *CXPB*; if crossing is not selected, then the method determines if a new offspring must be generated by mutating an existing individual, which happens according to a Bernoulli random variable with mean probability *MUTPB*; if neither crossing nor mutation are selected, then one random individual is picked from the population and cloned. (In the pseudo-code, pick(k, P) denotes a random subset of k different individuals from population *P.)*.

Algorithm 4: Method to crossover two individuals.

```
method crossover(P) returns (x′, y′)
  {x, y} := pick(P, 2)
  (k1, A1) := x
  (k2, A2) := y
  d := dim P
  p := uniform(1, d − 1)
  x′ := (k2, A2[1 : p] · A1[p + 1 : d])
  y′ := (k1, A1[1 : p] · A2[p + 1 : d])
end
```

Algorithm 4 presents the method to perform crossover. First, it picks any two individuals x and y from population P and extracts their components (k_1, A_1) and (k_2, A_2). Next, it computes the number of attributes d in population P and generates a random natural p in interval $[1, d-1]$. Then it generates the following offsprings: $x' = (k_2, A_2[1:p] \cdot A_1[p+1:d])$ and $y' = (k_1, A_1[1:p] \cdot A_2[p+1:d])$, where $v[i:j]$ denotes the slice of vector v from its i-th position up to its j-th position and $v \cdot u$ denotes the catenation of vectors v and u. Summing up, the offsprings exchange the number of clusters in their parents and a part of their subspaces of attributes.

Algorithm 5: Method to mutate an individual.

```
method mutate(P) returns x′
  {x} := pick(P, 1)
  (k, A) := x
  k′ := uniform(k − 1, k + 1)
  p := uniform(1, |A|)
  d := dim P
  A′ := A[1 : p − 1] · (not A[p]) · A[p + 1 : d]
  x′ := (k′, A′)
end
```

Algorithm 5 presents the method to perform mutation. First, it picks an individual x from population P and extracts its components (k, A). Next, it computes a random natural in interval $[k-1, k+1]$, which will be used as the number of clusters in the offspring; it also computes a random natural in interval $[1, |A|]$ that represents the attribute to be mutated in the offspring. Finally, it assembles and returns the new offspring.

Once the new offspring has been generated, the $(MU + LAMBDA)$ genetic strategy selects the best $\lceil MU\,PSIZE \rceil$ individuals from both the current population and the offspring. It is implemented using the version of the Lexicase method that was described by Cava et al. [27]. Helmuth et al. [19] published an in-depth analysis of Lexicase; their conclusion was that this method is very appropriate in cases in which it is not easy to aggregate multiple quality indicators into a single value. This is the case of RóMULO, whose search strategy seeks to minimise the number of clusters and to maximise the Silhouette coefficient, to minimise the Davis-Bouldin coefficient, and to maximise the Caliński-Harabasz coefficient. Intuitively, it seeks to minimise the number of clusters since the smaller number of clusters the easier for a data engineer to understand them; it seeks to optimise the other coefficients because they are known to achieve their best values when the clusters are compact and isolated. Aggregating the four indicators into a meaningful single value is not easy because they range in different intervals whose lower and upper bounds do not have homogeneous interpretations. In such cases, the Lexicase method provides a good solution to implement a multi-objective fitness function. Note that selecting the best individuals requires to evaluate them using the three previous coefficients. Simply put: the original dataset must be projected onto the subspace of attributes selected by each individual and then the projection must be clustered using any applicable proposal in the literature. Note that the decision on which exact proposal must be used requires some experimentation, which is the reason why it is deferred to the section on the experimental analysis.

The main loop of the algorithm evolves the initial population *NGEN* times. When it finishes, the best individual from the last population is selected. Let that individual be (k, A). The main method then projects the input dataset onto the subspace of attributes denoted by A and then clusters the resulting projection using the same method used to compute the coefficients that guide the search process.

4. Experimental analysis

This section presents the details behind the experimental analysis, namely: first, the experimental setting is described; then, the configuration procedure is explained; next, the experimental methodology is introduced; after that, the experimental results regarding clustering power are presented; finally, the experimental results regarding classification power are also presented.

4.1. Experimental setting

RóMULO[1] was implemented using Python 3.7.6 and several components: DEAP 1.3.1 to implement the meta-heuristic, Pandas 1.0.3 to implement the datasets, NumPy 1.18.2 to implement vector and matrix operations, and Scikit-Learn 0.24.2 to leverage the implementation of some classical clusterers and to compute performance measures; the implementations of GSPPCA [8], PQC [12], and DCC [43] were provided by their authors.

The experiments regarding the ability to find compact and isolated clusterings were run on a collection of 46 data lakes that provide 15 435 171 unclassified data that are grouped into 2 561 datasets. They were sampled from several open-data governmental repositories (Brazil, Canada, France, Spain, USA, and UK) and several reputed organisations (Kaggle, UCI, World Bank, and World Health Organisation). Table 1 provides a description in terms of name, number of datasets, number of data, number of attributes, and ratio of informative attributes according to GSPPCA. The data lakes have an average of 55.67 datasets, with a global minimum of four datasets and a global maximum of 531 datasets. They range in average size from 556.28 to 7 632.70 data per dataset, with an average of 3 755.78 data, a global minimum of three data, and a global maximum of 43 828 data; the average number of attributes ranges from 38.76 to 2 995.02 attributes per dataset, with an average of 477.64 attributes, a global minimum of three attributes, and a global maximum of 69 166 attributes. The average ratio of informative attributes ranges from 0.19 to 0.87 per dataset, with an average of 0.55, a global minimum smaller than 0.01, and a global maximum that equals 1.00. As of the time of writing this article, this is the largest experimentation repository in the context of data lakes. These data were not pre-classified because they are business data without any particular purposes.

The experiments regarding the extent to which the clusterings returned by RóMULO or the competitors can be considered good classifications were run on five additional data lakes that provide 1 200 552 pre-classified data that are grouped into 100 datasets. They were randomly sampled from the open-data repositories by Kaggle, OpenML, and the UCI, plus some standard evaluation datasets from Scikit-Learn and some common synthetic datasets. Table 2 summarises them: the data lakes have an average of 20.00 datasets, with a global minimum of four datasets and a global maximum of 53 datasets. They range in average size from 154.60 to 35 461.20 data per dataset, with an average of 7 674.36 data, a global minimum of seven data, and a global maximum of 52 619 data; the average number of attributes ranges from 15.60 to 3 465.20 attributes per dataset, with an average of 520.96 attributes, a global minimum of two attributes, and a global maximum of 10 000 attributes. The

[1] RóMULO and the experimental data lakes are available at https://doi.org/10.17632/y5v2zy356t.1.

Table 1
Description of the clustering data lakes.

		Number of data			Number of attributes			Ratio of informative attributes		
Data lake	Number of datasets	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum
Adolescents NIH	22	4679	9110.00	22671	23	55.32	138	0.04	0.80	1.00
Agrifood Exports	11	6	27.91	53	10	33.64	63	0.27	0.85	1.00
Agro climatic	47	444	8970.28	10000	15	15.00	15	0.20	0.77	1.00
Argentina Climate	23	27	2722.35	8000	7	126.87	535	0.01	0.29	1.00
Art Funding	7	12	306.71	1000	34	359.00	798	0.05	0.61	1.00
Beaches Euskadi	16	40	229.88	1159	12	84.88	231	0.03	0.10	0.25
Brasilian Taxes	31	8	4704.16	8788	6	15.61	30	0.15	0.39	1.00
Citizen Workforce	23	6	2233.87	8000	10	172.78	925	0.13	0.76	1.00
COVID	113	26	4898.96	9092	5	258.94	9291	0.00	0.27	1.00
Crimes	149	35	8512.27	10000	24	81.41	449	0.02	0.59	1.00
Deputies Canary	29	3	72.34	667	6	20.00	46	0.11	0.33	0.67
Disasters	19	109	1641.11	10000	53	760.16	4570	0.01	0.55	1.00
Divorces	17	24	354.76	1176	34	207.24	353	0.01	0.29	1.00
Doctoral Graduates	178	33	7304.76	12201	15	58.73	3932	0.00	0.63	1.00
EU Fruit Data	30	87	2674.77	5000	21	231.57	1218	0.01	0.75	1.00
Farm Operators	29	792	7013.00	10000	36	184.14	257	0.02	0.80	1.00
Gas Emissions	17	4	2148.76	5000	4	257.71	3530	0.01	0.85	1.00
Hiking Tours	16	44	2495.13	5000	12	2346.38	3508	0.00	0.20	1.00
Homicide Victims	11	100	1503.09	3396	18	56.55	141	0.04	0.51	1.00
Husa Data	350	3402	10704.45	12345	190	190.00	190	1.00	1.00	1.00
Marriage	4	17	17.00	17	22	74.50	100	0.71	0.76	0.82
Microsoft Kinetic	20	806	2458.90	5775	82	82.00	82	1.00	1.00	1.00
National Households	70	14	8686.50	10000	4	1922.59	2185	0.00	0.80	1.00
NBA Raptors	8	179	5774.00	7000	31	9092.50	69166	0.00	0.17	1.00
Puerto Rico Media	7	37	50.00	84	7	8.71	13	0.23	0.41	0.50
Regular Force Outflow	10	20	20.90	22	3	5.60	17	0.29	0.53	1.00
Rental Property NY	15	493	3303.13	5000	208	965.87	3643	0.00	0.74	1.00
Riddler Castles	5	886	1067.20	1403	11	52.40	159	0.04	0.30	0.55
Rockfall Risk	20	362	1951.75	2627	27	33.90	35	0.93	0.98	1.00
Second Language	20	9	3522.55	10000	11	656.80	3249	0.00	0.29	1.00
Steller Sea Lions	41	24	9280.07	43828	35	247.17	1232	0.01	0.11	1.00
Tax Filers	531	1430	4616.57	9284	65	148.95	1318	0.00	0.79	1.00
Teen Pregnancy	18	117	6787.83	10000	35	72.89	120	0.03	0.73	1.00
Tourist Accommodation	21	26	1206.57	6011	46	1661.00	17570	0.16	0.92	1.00
Trump world trust	7	15	31.00	37	40	42.14	43	0.12	0.30	0.83
Tuition Fees	25	2494	7934.44	10000	40	226.36	4598	1.00	1.00	1.00
TV Access Services	5	324	3022.60	8688	125	318.00	507	0.02	0.61	1.00
TV Commercials	13	1546	9975.77	33117	201	203.31	205	0.54	0.59	0.82
UBER Trips	11	59	3079.73	7154	7	254.27	1196	0.00	0.30	1.00
University Spending	30	792	6491.93	7500	47	83.03	794	1.00	1.00	1.00
US Weather History	10	365	365.00	365	16	16.00	16	0.19	0.48	0.88
Vehicle Pedestrian Inv	326	121	8563.55	8737	34	35.93	37	0.08	0.20	0.31
Voter Turnout	4	5512	6312.75	7428	49	53.00	58	0.09	0.12	0.16
WiFi Hotspots	30	4	553.50	13395	4	91.27	1091	0.03	0.64	1.00
Women World Cup	88	24	32.06	52	46	55.18	65	0.05	0.08	0.11
World Cup	84	32	32.00	32	52	52.00	52	0.06	0.08	0.10
Grand average	55.67	556.28	3755.78	7632.70	38.76	477.64	2995.02	0.19	0.55	0.87

average ratio of informative attributes ranges from 0.23 to 0.98 per dataset, with an average of 0.63, a global minimum smaller than 0.01, and a global maximum that equals 1.00. The data in these data lakes were pre-classified by their corresponding authors, but the classes were not used during the clustering processes, only to compute the effectiveness measures.

There are not any direct competitors with which RóMULO can be compared experimentally because there are not any multi-way single-subspace automatic clustering techniques in the literature. The idea was then to implement some indirect competitors by assembling a pipeline in which the first stage computes a subspace of informative attributes and the second stage performs single-way automatic clustering. The choice regarding the first stage was to use GSPPCA [8], which is the most recent unsupervised proposal that can deal with business data [28]. The choice regarding the second stage was to use some classical proposals, including Affinity-Propagation, Mean-Shift, and OPTICS-Xi, as well as PQC [12] and DCC [43], which are the most recent quantum-inspired and deep-learning proposals, respectively.

The machinery used to run the experiments consisted in a computer that was equipped an Intel Core i7-9700F processor with eight single-threaded cores at 3.70 GHz and 16 GiB of DDR4 RAM memory at 2.67 GHz. The operating system used was Windows 10 Pro.

Table 2
Description of the classification data lakes.

		Number of data			Number of attributes			Ratio of informative attributes		
Data lake	Number of datasets	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum
Kaggle	21	165	8139.81	51047	5	206.00	2219	0.01	0.49	1.00
OpenML	53	7	6992.00	52619	3	373.00	10000	0.03	0.56	1.00
Sckit-Learn	4	400	13745.00	39550	64	1799.00	4096	0.70	0.93	1.00
Synthetic	9	60	6687.00	10000	2	165.00	450	0.33	0.85	1.00
UCI	13	141	2808.00	24090	4	61.00	561	0.06	0.32	0.91
Grand average	20.00	154.60	7674.36	35461.20	15.60	520.80	3465.20	0.23	0.63	0.98

4.2. Configuring RóMULO and the competitors

A subset of 100 datasets was randomly selected from the experimental data lakes. They were used to perform grid search over the space of configuration parameters of RóMULO and the competitors. Table 3 describes the configuration parameters. The first column refers to the proposal, the second column describes the configuration parameters and the values that were examined; the values that were selected are highlighted in boldface. The proposals were executed using all of the combinations of values for their configuration parameters and the ones that resulted in better performance according to the Lexicase method were selected; see the performance measures used in the following subsection.

RóMULO requires to use a base clusterer to compute the fitness of the individuals and to compute its final result. The decision was to use the well-known *k*-Means algorithm using a custom initialisation procedure [32]. The alternative was to use Birch, but, unfortunately, it failed to handle many high-dimensional datasets in the experimental data lakes.

Note, too, that Mean-Shift and PQC do not require any configuration parameters to be adjusted using grid search because the authors devised estimation methods that are applied to the input datasets automatically.

4.3. Experimental methodology

The datasets were cleaned as follows: text fields were removed, dates, times, co-ordinates, and other such structured fields were split into their constituent parts, and enumerated fields were one-hot encoded. Each competitor was run on each dataset using its best configuration parameters and several performance measures were computed and collected.

Table 3
Configuration parameters.

Proposal	Configuration parameter/ Values									
RóMULO	*CXPB*: crossover probability									
	0.10	0.20	**0.30**	0.40	0.50	0.60	0.70	0.80	0.90	1.00
	MUTPB: mutation probability									
	0.01	0.05	0.10	**0.15**	0.20	0.25	0.30	0.35	0.40	0.45
	PSIZE: size of the initial population									
	10.00	15.00	20.00	25.00	30.00	**35.00**	40.00	45.00	50.00	55.00
	LAMBDA: percentage of offsprings to generate (relative to the SIZE parameter)									
	0.10	0.20	0.30	0.40	0.50	0.60	**0.70**	0.80	0.90	1.00
	MU: percentage of individuals to select (relative to the SIZE parameter)									
	0.10	0.20	0.30	0.40	0.50	0.60	**0.70**	0.80	0.90	1.00
	NGEN: number of generations									
	10.00	15.00	20.00	25.00	30.00	**35.00**	40.00	45.00	50.00	55.00
	epsi: convergence criterion									
	1.00-10	1.00-09	1.00-08	1.00-07	1.00-06	**1.00-05**	1.00-04	1.00-03	1.00-02	0.00 + 00
	nit: dimension of the latent space									
GSPPCA	25.00	50.00	75.00	**100.00**	125.00	150.00	200.00	250.00	300.00	350.00
	dampling: extent to which the current value is maintained relative to incoming values									
Aff.-Prop.	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	**0.90**	1.00
	bandwidth: bandwidth used in the RBF kernel									
Mean-Shift	Adjusted using the estimation procedure provided by Scikit Learn.									
	xi: minimum steepness on the reachability plot that constitutes a cluster boundary									
OPTICS-Xi	**0.01**	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80
	k: proportion of neighbours used to estimate the local covariance matrix for each datum									
PQC	Adjusted using the estimation procedure provided by the authors.									
	d: embedding dimensionality (ratio of the input dimensionality)									
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
	k: number of neighbours (ratio of the number of input data)									
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
	M: period of graduated non-convexity									
DCC	10.00	**20.00**	30.00	40.00	50.00	60.00	70.00	80.00	90.00	100.00

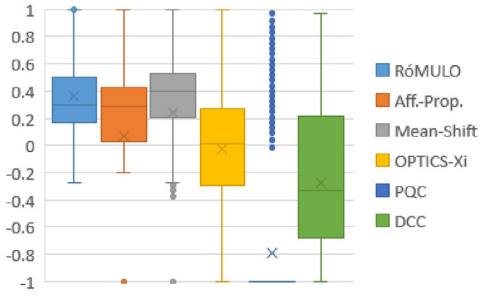
Table 4
Clustering power: Silhouette.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC	Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Adolescents NIH	0.33	0.48	0.25	0.06	-0.70	-0.70	NBA Raptors	0.12	0.21	-0.30	-0.24	-0.78	-0.26
Agrifood Exports	0.45	0.20	0.15	-0.51	0.09	-0.33	Puerto Rico Media	0.44	0.69	0.37	0.56	0.35	-0.14
Agro climatic	0.47	-0.01	0.31	-0.30	-0.97	-0.91	Regular Force Outflow	0.52	0.47	0.49	0.11	0.55	0.11
Argentina Climate	0.29	0.10	-0.02	-0.02	0.09	-0.46	Rental Property NY	0.46	0.12	0.33	-0.10	-1.00	-0.10
Art Funding	0.34	0.19	0.26	-0.17	-0.36	-0.56	Riddler Castles	0.32	0.11	0.34	-0.24	-1.00	-0.30
Beaches Euskadi	0.35	0.09	0.28	0.33	0.37	-0.32	Rockfall Risk	0.95	0.27	0.41	-0.63	-0.43	-0.26
Brasilian Taxes	0.27	0.39	0.28	-0.10	-0.63	-0.50	Second Language	0.40	0.32	0.30	0.12	-0.12	0.12
Citizen Workforce	0.37	0.42	0.19	-0.12	-0.51	-0.38	Steller Sea Lions	0.67	0.19	0.00	-0.23	-0.96	-0.23
COVID	0.62	0.32	0.61	-0.16	-0.14	-0.63	Tax Filers	0.30	-0.31	0.27	0.21	-1.00	0.21
Crimes	0.62	0.35	0.19	0.06	-0.90	-0.61	Teen Pregnancy	0.42	0.33	0.41	0.24	-0.51	0.24
Deputies Canary	0.45	0.12	0.00	-0.29	0.12	-0.43	Tourist Accomodation	0.36	0.17	0.12	-0.08	-0.59	-0.32
Disasters	0.49	0.28	0.17	-0.36	-0.31	-0.52	Trump world trust	0.45	-1.00	0.35	-0.12	-0.67	-0.19
Divorces	0.32	-0.10	0.57	0.15	0.24	-0.54	Tuition Fees	0.36	0.43	0.37	0.00	-1.00	0.00
Doctoral Graduates	0.20	0.37	0.36	0.01	-0.90	-0.80	TV Access Services	0.30	-0.60	0.43	0.26	-0.38	-0.12
EU Fruit Data	0.61	-0.13	0.48	0.09	-0.38	-0.64	TV Commercials	0.14	-0.21	0.07	-0.39	-1.00	-0.30
Farm Operators	0.26	0.49	0.34	0.16	-0.75	-0.38	UBER Trips	0.54	0.39	-0.40	-0.12	0.13	-0.12
Gas Emissions	0.44	0.12	0.44	0.00	-0.14	-0.50	University Spending	0.35	0.59	0.18	0.43	-1.00	0.43
Hiking Tours	0.45	0.31	0.32	-0.06	-0.60	-0.13	US Weather History	0.45	0.28	0.29	-0.14	0.28	-0.01
Homicide Victims	0.59	0.21	0.62	0.39	0.15	-0.31	Vehicle Pedestrian Inv	0.26	0.29	0.64	0.05	-0.97	0.05
HUSA data	0.34	-0.36	-0.20	-0.43	-1.00	-0.61	Voter Turnout	0.89	0.30	-1.00	-0.26	-1.00	-0.26
Marriage	0.80	-0.15	0.40	0.52	-0.09	0.16	WiFi Hotspots	0.38	0.13	-0.04	-0.25	-0.22	-0.33
Microsoft Kinetic	0.86	0.17	0.58	-0.10	-0.26	-0.69	Women World Cup	0.25	0.35	0.39	0.35	-0.79	-0.28
National Households	0.08	0.15	-0.84	-0.81	-0.80	-0.80	World Cup	0.35	0.49	0.44	0.30	-1.00	-0.08

a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	-0.28	-1.00	-1.00	-1.00	-1.00	-1.00
25th percentile	0.17	0.03	0.21	-0.29	-1.00	-0.68
Median	0.30	0.29	0.40	0.01	-1.00	-0.33
75th percentile	0.50	0.42	0.53	0.27	-1.00	0.21
Maximum	1.00	1.00	1.00	1.00	1.00	0.97
Average	0.36	0.07	0.24	-0.03	-0.79	-0.27
Standard deviation	0.25	0.56	0.55	0.42	0.53	0.50
Variance	0.06	0.31	0.31	0.17	0.28	0.25
Conf. Int. at 95%	0.49	1.10	1.08	0.82	1.03	0.98

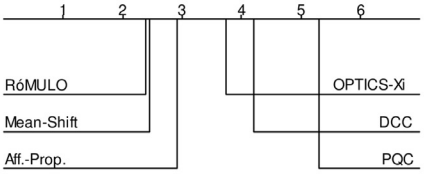
b) Global statistics.



c) Boxplot chart.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	2.39		N/A
Aff.-Prop.	2.92		0.00
Mean-Shift	2.44	0.00	< 0.01
OPTICS-Xi	3.74		0.00
PQC	5.31		0.00
DCC	4.20		0.00

d) Statistical analysis.



e) Critical diference diagram.

The effectiveness measures regarding the ability to find compact and isolated clusterings were the Silhouette, the Davies-Bouldin, and the Caliński-Harabasz coefficients. The Silhouette coefficient ranges in interval $[-1.00, +1.00]$, where the higher its value, the better; the Davies-Bouldin coefficient ranges in interval $[0.00, +\infty]$, where the lower its value, the better; and the Caliński-Harabasz coefficient ranges in interval $[0.00, +\infty]$, where the higher its value, the better. We also computed the degradation ratio to assess the proportion of useless clusterings that consists of a single cluster or a cluster per datum. It ranges in interval $[0.00, +1.00]$, where the smaller its value, the better.

The effectiveness measures regarding the extent to which the clusterings can be considered good classifications were the Adjusted Rand, the Fowlkes-Mallows, and the Accuracy coefficients. They all address the problem of mapping the classes that are inferred from the clustering, which are machine-learnt and have no meaning for the user, onto the classes in the ground truth, which are user-provided and expected to have a meaning for the user; they differ in their approach and the perspective from which they address the comparison. The Adjusted Rand coefficient takes into account the effects of randomness in clustering; it ranges in interval $[-1.00, +1.00]$, where the greater its value, the better. The Fowlkes-Mallows coefficient takes into

account the effects of noise and the Accuracy coefficient puts an emphasis on both true positives and true negatives. They both range in interval $[0.00, +1.00]$, where the higher their values, the better. We also computed the degradation ratio.

The efficiency measures were the time to execute each experiment and the timeout ratio. The time was measured in CPU minutes, which helps reduce noise as much as possible since the times measured at the CPU level are generally more stable and consistent throughout repeated experiments than the wall time. A timeout of 240 min (four hours) was set to compute the subspace of informative attributes using GSPPCA and an additional 240-min timeout was set to run the clusterers. Very likely, few data engineers would be willing to wait for eight hours to compute a clustering, but it puts an upper, sensible limit to the total time required to run the experiments.

The experimental results were analysed using the following statistical method at the standard significance level ($\alpha = 0.05$) [45]: first, the empirical ranks regarding each measure were computed; second, Iman-Davenport's omnibus test was used to determine if there were any significant differences in the empirical ranks; if there were, then Hommel's post hoc

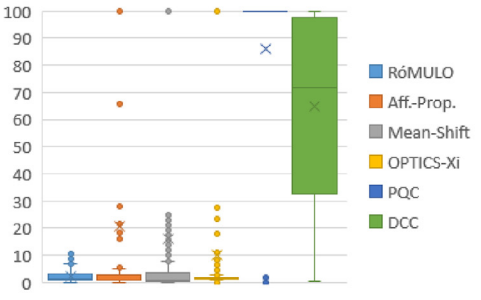
Table 5
Clustering power: Davies-Bouldin.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC	Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Adolescents NIH	1.78	5.19	14.21	14.82	81.95	77.56	NBA Raptors	1.78	16.51	50.35	14.89	87.52	62.44
Agrifood Exports	0.78	1.41	18.69	64.31	18.59	64.04	Puerto Rico Media	0.45	0.40	0.74	0.70	1.20	55.44
Agro climatic	1.78	16.07	2.93	1.40	97.90	93.10	Regular Force Outflow	0.95	0.67	0.58	21.21	0.60	55.89
Argentina Climate	0.81	18.25	30.88	5.97	14.13	55.49	Rental Property NY	1.78	14.04	0.53	1.86	100.00	53.83
Art Funding	0.73	1.12	0.59	17.09	57.37	76.73	Riddler Castles	0.35	20.73	0.69	1.95	100.00	34.39
Beaches Euskadi	0.49	25.64	0.82	7.48	1.06	48.94	Rockfall Risk	0.17	0.43	5.17	2.25	70.06	31.00
Brasilian Taxes	0.51	7.05	7.10	9.18	71.38	82.35	Second Language	1.29	15.61	15.46	11.38	40.51	56.56
Citizen Workforce	1.41	0.76	17.85	23.23	65.57	42.69	Steller Sea Lions	1.68	3.69	29.84	6.21	97.59	59.27
COVID	0.94	4.37	2.24	4.85	48.09	77.96	Tax Filers	1.73	43.83	19.44	6.19	100.00	61.30
Crimes	0.86	3.52	18.62	1.34	92.72	86.74	Teen Pregnancy	1.46	11.80	0.67	1.29	67.00	58.63
Deputies Canary	0.57	24.54	28.11	45.71	18.06	56.83	Tourist Accomodation	1.24	3.76	14.92	9.78	67.28	32.07
Disasters	1.08	11.89	16.22	33.03	58.26	59.43	Trump world trust	0.60	100.00	0.99	30.00	72.01	31.81
Divorces	1.18	41.63	6.24	13.21	24.09	53.49	Tuition Fees	1.31	3.35	4.60	5.19	100.00	52.57
Doctoral Graduates	0.55	6.33	4.20	5.32	93.91	81.88	TV Access Services	0.31	80.00	0.50	1.71	60.53	35.55
EU Fruit Data	1.43	34.18	0.55	2.10	60.31	69.73	TV Commercials	2.05	24.94	10.66	9.19	100.00	25.32
Farm Operators	1.82	14.23	10.74	7.99	82.95	59.81	UBER Trips	1.07	0.76	63.82	2.54	19.12	62.34
Gas Emissions	1.56	18.30	0.44	13.78	47.38	54.75	University Spending	1.63	1.22	14.03	4.43	100.00	51.79
Hiking Tours	0.54	1.11	6.90	2.00	75.25	28.24	US Weather History	0.75	1.12	11.03	1.73	1.83	30.60
Homicide Victims	0.80	27.69	0.51	0.92	27.75	64.37	Vehicle Pedestrian Inv	4.05	6.30	1.00	1.49	98.16	55.95
HUSA data	4.21	38.99	32.40	15.34	100.00	79.82	Voter Turnout	0.68	0.95	100.00	1.26	100.00	73.24
Marriage	0.37	50.11	0.60	0.55	50.92	42.93	WiFi Hotspots	0.95	20.68	30.58	41.32	47.09	43.49
Microsoft Kinetic	0.46	11.00	0.55	1.33	60.10	86.63	Women World Cup	0.52	12.97	0.87	3.84	81.97	39.92
National Households	1.09	5.23	88.64	83.59	85.87	79.24	World Cup	0.61	0.67	0.97	4.46	100.00	40.43

a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	0.34
25th percentile	0.62	0.72	0.45	1.28	100.00	32.53
Median	1.33	1.00	0.75	1.50	100.00	71.84
75th percentile	3.02	2.50	3.29	1.74	100.00	97.43
Maximum	12.25	100.00	100.00	100.00	100.00	100.00
Average	1.95	20.38	16.02	9.95	85.88	64.99
Standard deviation	1.76	39.16	33.91	27.42	34.65	32.76
Variance	3.10	> 1E3	> 1E3	752.06	> 1E3	> 1E3
Conf. Int. at 95%	0.07	1.52	1.31	1.06	1.34	1.27

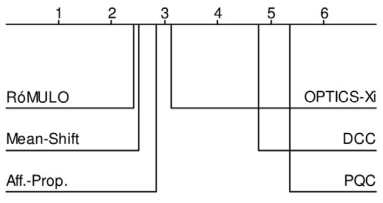
b) Global statistics.



c) Boxplot chart.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	2.42	0.00	N/A
Aff.-Prop.	2.84		< 0.01
Mean-Shift	2.50		0.00
OPTICS-Xi	3.13		0.01
PQC	5.35		0.00
DCC	4.76		0.00

d) Statistical analysis.



e) Critical diference diagram.

test was used to find the exact significant differences. Note that non-parametrical tests were used because the analysis of the experimental results using Shapiro–Wilk's and Levene's tests concluded that they were distributed neither normally nor homocedastically.

4.4. Analysis of clustering power

Tables 4–9 summarise the results regarding the ability of RóMULO and the competitors to find compact and isolated clusters. Their structure is very homogeneous, namely: part a) shows the averages per data lake, which helps have an overall picture of which ones resulted easier to cluster and which ones resulted more difficult; part b) shows some statistics that

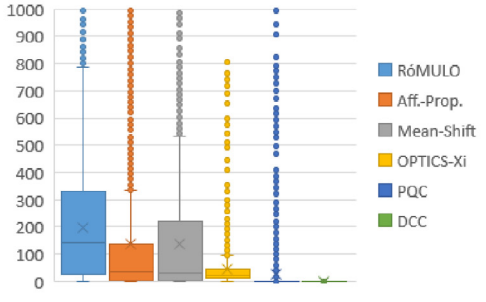
Table 6
Clustering power: Caliński-Harabasz.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC	Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Adolescents NIH	288.90	220.07	268.86	127.57	90.00	0.81	NBA Raptors	31.27	65.32	115.42	13.69	3.11	0.00
Agrifood Exports	42.83	78.17	29.93	10.00	49.37	0.32	Puerto Rico Media	83.31	134.46	145.26	92.53	23.69	0.00
Agro climatic	516.69	121.27	165.20	13.43	0.78	0.44	Regular Force Outflow	20.94	41.85	35.42	17.00	35.25	0.00
Argentina Climate	177.61	179.70	16.53	14.89	17.69	0.02	Rental Property NY	34.09	132.94	122.54	6.48	0.00	0.00
Art Funding	2.91	9.64	10.31	2.11	6.38	0.09	Riddler Castles	212.48	223.92	111.48	11.07	0.00	0.00
Beaches Euskadi	31.39	43.36	108.61	25.50	50.87	0.00	Rockfall Risk	64.15	117.20	377.57	9.13	7.85	0.00
Brasilian Taxes	188.43	106.89	155.15	23.63	29.13	0.16	Second Language	72.61	400.30	211.83	35.45	144.36	0.00
Citizen Workforce	117.00	319.89	82.85	26.30	18.58	0.66	Steller Sea Lions	81.35	210.03	90.81	16.61	0.35	0.00
COVID	258.69	348.87	356.86	44.92	244.63	1.75	Tax Filers	160.45	51.01	83.27	115.44	0.00	0.00
Crimes	371.68	310.20	192.34	25.96	3.40	0.13	Teen Pregnancy	201.29	327.22	249.10	135.85	37.15	0.00
Deputies Canary	80.72	150.33	120.92	36.48	11.89	0.01	Tourist Accomodation	21.75	13.42	11.99	2.36	15.08	0.00
Disasters	13.45	242.31	91.26	7.10	96.12	0.00	Trump world trust	2.58	0.00	23.53	8.72	0.70	0.00
Divorces	20.89	80.99	136.51	26.40	88.95	0.14	Tuition Fees	155.62	196.43	207.98	67.72	0.00	0.00
Doctoral Graduates	63.78	42.27	59.46	28.59	26.94	0.06	TV Access Services	47.70	74.21	68.27	92.60	60.76	0.00
EU Fruit Data	126.57	127.30	134.53	19.34	130.99	0.01	TV Commercials	334.02	25.54	84.77	8.96	0.00	0.00
Farm Operators	91.63	248.58	46.70	23.61	21.87	0.01	UBER Trips	303.74	311.19	96.07	28.35	314.82	0.00
Gas Emissions	147.01	112.58	165.29	11.98	117.55	0.08	University Spending	288.44	378.42	377.16	184.93	0.00	0.00
Hiking Tours	5.60	276.45	231.31	15.94	14.86	0.01	US Weather History	318.74	224.86	328.84	13.98	88.41	0.00
Homicide Victims	174.96	228.60	190.68	35.40	144.88	0.07	Vehicle Pedestrian Inv	338.59	278.80	364.89	30.50	2.15	0.00
HUSA data	315.41	19.96	7.11	15.28	0.00	0.02	Voter Turnout	132.95	182.88	0.00	15.51	0.00	0.00
Marriage	61.47	51.72	85.46	34.90	337.37	0.01	WiFi Hotspots	40.00	55.47	62.13	36.09	9.33	0.00
Microsoft Kinetic	65.57	348.84	305.82	24.46	104.91	0.20	Women World Cup	21.66	42.50	32.39	25.07	0.33	0.00
National Households	9.43	37.98	14.09	1.64	6.52	0.10	World Cup	27.68	55.23	42.35	23.53	0.00	0.00

a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	0.74	0.00	0.00	0.00	0.00	0.00
25th percentile	26.56	3.40	4.95	13.04	0.00	< 0.01
Median	141.34	34.74	32.00	21.81	0.00	< 0.01
75th percentile	332.32	136.67	217.74	45.46	0.00	< 0.01
Maximum	993.36	997.05	996.17	823.19	996.17	13.44
Average	197.08	136.16	137.77	46.69	25.12	0.12
Standard deviation	189.57	217.52	201.78	89.56	111.56	0.74
Variance	> 1E3	> 1E3	> 1E3	> 1E3	> 1E3	0.55
Conf. Int. at 95%	7.34	8.42	7.81	3.47	4.32	0.03

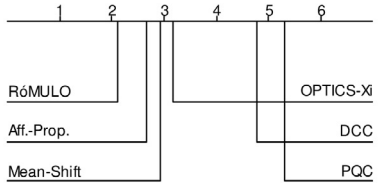
b) Global statistics.



c) Boxplot chart.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	2.12	0.00	N/A
Aff.-Prop.	2.67		0.00
Mean-Shift	2.94		0.00
OPTICS-Xi	3.17		0.00
PQC	5.32		0.00
DCC	4.78		0.00

d) Statistical analysis.



e) Critical diference diagram.

help understand how the results are distributed on the datasets, namely: the quartiles, the average, the standard deviation, the variance, and the length of the 95% confidence interval; part c) shows a boxplot that provides a graphical insight into the distribution of the performance measures; part d) shows the results of the statistical analysis, namely: the proposal, its corresponding empirical rank, the p-value computed by Iman-Davenport's omnibus test, and the p-value computed by Hommel's post-host test; and part e) provides a graphical representation of the statistical analysis using critical difference diagrams.

Regarding the Silhouette coefficient, cf. Table 4, RóMULO ranks at the first position and it is followed by the competitors based on Mean-Shift, Affinity-Propagation, OPTICS-Xi, DCC, and PQC. RóMULO attains an average Silhouette coefficient of 0.36, with a minimum of −0.28 and a maximum of 1.00. The minimum occurs in a dataset from the "Disasters" data lake

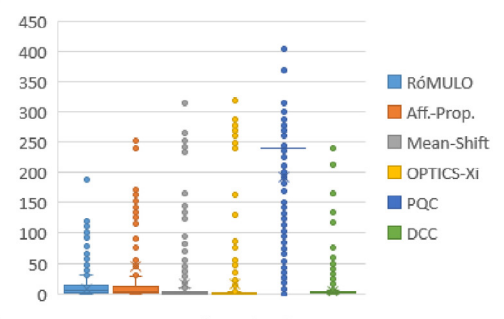
Table 7
Clustering power: Time.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC	Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Adolescents NIH	18.25	15.99	10.87	33.74	203.09	3.58	NBA Raptors	12.95	22.88	52.65	51.82	231.95	54.01
Agrifood Exports	0.11	0.03	0.03	0.03	0.70	0.15	Puerto Rico Media	0.02	0.02	0.02	0.02	0.21	1.76
Agro climatic	16.24	43.00	0.99	1.04	236.48	8.28	Regular Force Outflow	0.11	0.11	0.11	0.11	0.30	0.19
Argentina Climate	0.46	0.77	0.32	0.15	1.85	0.80	Rental Property NY	11.09	36.34	5.67	5.90	243.29	4.30
Art Funding	0.93	0.25	0.25	0.28	104.85	1.28	Riddler Castles	0.12	0.06	0.10	0.05	240.04	1.87
Beaches Euskadi	0.04	0.02	0.02	0.02	2.70	0.25	Rockfall Risk	0.16	0.14	0.08	0.08	168.48	1.35
Brasilian Taxes	0.92	0.71	0.33	0.11	172.25	1.65	Second Language	9.58	39.45	3.40	3.33	112.26	5.39
Citizen Workforce	5.23	1.01	6.34	2.10	135.96	0.95	Steller Sea Lions	27.95	13.22	7.88	9.31	237.07	5.38
COVID	5.71	1.85	2.10	1.66	151.48	5.70	Tax Filers	8.00	103.26	2.13	12.33	240.35	2.42
Crimes	15.96	13.25	5.31	1.54	223.66	1.89	Teen Pregnancy	14.18	28.93	2.43	1.55	171.78	1.77
Deputies Canary	0.02	0.01	0.01	0.01	0.61	0.12	Tourist Accomodation	1.17	18.02	34.06	29.26	153.29	19.42
Disasters	13.28	37.71	37.08	49.54	156.02	16.35	Trump world trust	0.02	0.01	0.01	0.01	0.40	1.30
Divorces	0.68	84.78	0.07	0.07	61.85	0.65	Tuition Fees	16.50	5.79	7.60	13.29	242.39	4.07
Doctoral Graduates	1.31	1.67	0.66	0.50	238.39	3.97	TV Access Services	0.65	2.25	3.25	2.27	233.25	2.20
EU Fruit Data	4.73	80.84	1.60	1.92	139.70	1.15	TV Commercials	2.85	40.86	33.21	24.14	240.72	1.91
Farm Operators	15.93	35.53	15.98	5.19	209.82	2.07	UBER Trips	5.96	0.70	0.67	0.58	50.66	2.94
Gas Emissions	2.23	43.14	0.49	14.65	114.17	2.41	University Spending	9.01	2.45	1.64	1.24	240.07	1.92
Hiking Tours	16.95	14.72	14.71	14.70	194.97	32.14	US Weather History	0.05	0.04	0.05	0.04	2.07	1.71
Homicide Victims	1.31	43.74	0.17	0.12	73.78	0.36	Vehicle Pedestrian Inv	14.71	14.41	1.01	0.55	235.85	1.91
HUSA data	9.39	98.58	64.41	44.42	240.25	3.17	Voter Turnout	9.09	1.12	0.92	0.27	240.13	2.00
Marriage	0.02	0.03	0.03	0.03	221.10	0.15	WiFi Hotspots	0.14	0.07	0.08	0.03	49.14	1.67
Microsoft Kinetic	4.03	24.24	0.75	0.39	186.68	0.82	Women World Cup	0.02	0.02	0.02	0.02	0.33	1.59
National Households	4.77	108.83	135.49	199.65	212.22	36.43	World Cup	0.02	0.02	0.02	0.02	0.31	1.50

a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	0.01	< 0.01	< 0.01	< 0.01	0.06	0.04
25th percentile	1.31	0.43	0.22	0.16	240.01	1.11
Median	5.88	3.08	0.95	0.59	240.10	2.08
75th percentile	12.96	12.32	4.34	1.88	240.25	3.14
Maximum	187.20	251.75	318.55	318.55	406.54	240.20
Average	8.30	46.02	15.20	16.23	193.28	4.16
Standard deviation	10.68	89.10	46.10	56.69	94.61	12.14
Variance	114.09	> 1E3	> 1E3	> 1E3	> 1E3	147.38
Conf. Int. at 95%	0.41	3.45	1.78	2.20	3.66	0.47

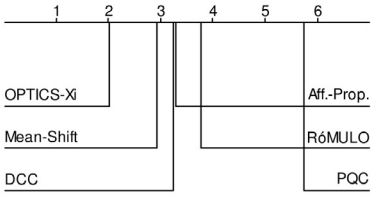
b) Global statistics.



c) Boxplot chart.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	3.77	0.00	0.00
Aff.-Prop.	3.29		0.00
Mean-Shift	2.92		0.00
OPTICS-Xi	2.03		N/A
PQC	5.75		0.00
DCC	3.25		0.00

d) Statistical analysis.



e) Critical diference diagram.

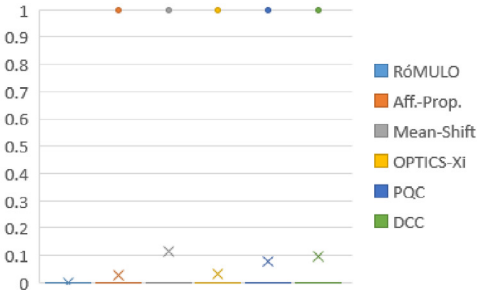
Table 8
Clustering power: Degradation.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC	Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Adolescents NIH	0.00	0.00	0.14	0.00	0.00	0.00	NBA Raptors	0.00	0.13	0.38	0.00	0.00	0.00
Agrifood Exports	0.00	0.00	0.18	0.64	0.18	0.36	Puerto Rico Media	0.00	0.00	0.00	0.00	0.00	0.14
Agro climatic	0.00	0.00	0.02	0.00	0.00	0.89	Regular Force Outflow	0.00	0.00	0.00	0.20	0.00	0.20
Argentina Climate	0.00	0.17	0.30	0.04	0.13	0.04	Rental Property NY	0.00	0.00	0.00	0.00	0.00	0.00
Art Funding	0.00	0.00	0.00	0.14	0.14	0.29	Riddler Castles	0.00	0.20	0.00	0.00	0.00	0.20
Beaches Euskadi	0.00	0.25	0.00	0.06	0.00	0.13	Rockfall Risk	0.00	0.00	0.05	0.00	0.00	0.25
Brasilian Taxes	0.00	0.06	0.06	0.06	0.00	0.23	Second Language	0.00	0.00	0.15	0.10	0.05	0.10
Citizen Workforce	0.00	0.00	0.17	0.22	0.09	0.00	Steller Sea Lions	0.00	0.00	0.29	0.02	0.00	0.05
COVID	0.00	0.04	0.02	0.04	0.00	0.04	Tax Filers	0.00	0.00	0.19	0.00	0.00	0.05
Crimes	0.00	0.00	0.18	0.00	0.00	0.00	Teen Pregnancy	0.00	0.00	0.00	0.00	0.00	0.00
Deputies Canary	0.00	0.24	0.28	0.45	0.17	0.31	Tourist Accomodation	0.00	0.00	0.10	0.05	0.14	0.19
Disasters	0.00	0.00	0.11	0.16	0.05	0.05	Trump world trust	0.00	1.00	0.00	0.29	0.71	0.00
Divorces	0.00	0.06	0.06	0.12	0.00	0.00	Tuition Fees	0.00	0.00	0.04	0.00	0.00	0.04
Doctoral Graduates	0.00	0.06	0.03	0.04	0.00	0.53	TV Access Services	0.00	0.80	0.00	0.00	0.00	0.00
EU Fruit Data	0.00	0.00	0.00	0.00	0.03	0.00	TV Commercials	0.00	0.08	0.00	0.00	0.00	0.15
Farm Operators	0.00	0.00	0.10	0.07	0.00	0.00	UBER Trips	0.00	0.00	0.64	0.00	0.00	0.00
Gas Emissions	0.00	0.00	0.00	0.06	0.00	0.18	University Spending	0.00	0.00	0.13	0.03	0.00	0.03
Hiking Tours	0.00	0.00	0.06	0.00	0.00	0.00	US Weather History	0.00	0.00	0.10	0.00	0.00	0.10
Homicide Victims	0.00	0.09	0.00	0.00	0.00	0.00	Vehicle Pedestrian Inv	0.00	0.01	0.01	0.00	0.00	0.00
HUSA data	0.00	0.00	0.02	0.00	0.00	0.00	Voter Turnout	0.00	0.00	1.00	0.00	0.00	0.00
Marriage	0.00	0.50	0.00	0.00	0.00	0.25	WiFi Hotspots	0.00	0.20	0.30	0.40	0.27	0.30
Microsoft Kinetic	0.00	0.00	0.00	0.00	0.00	0.55	Women World Cup	0.00	0.13	0.00	0.02	0.82	0.08
National Households	0.00	0.00	0.86	0.10	0.09	0.01	World Cup	0.00	0.00	0.00	0.02	1.00	0.04

a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	0.00	0.00	0.00	0.00	0.00	0.00
25th percentile	0.00	0.00	0.00	0.00	0.00	0.00
Median	0.00	0.00	0.00	0.00	0.00	0.00
75th percentile	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	0.00	1.00	1.00	1.00	1.00	1.00
Average	0.00	0.03	0.11	0.03	0.08	0.10
Standard deviation	0.00	0.16	0.32	0.18	0.26	0.30
Variance	0.00	0.03	0.10	0.03	0.07	0.09
Confidence (95%)	0.00	0.01	0.01	0.01	0.01	0.01

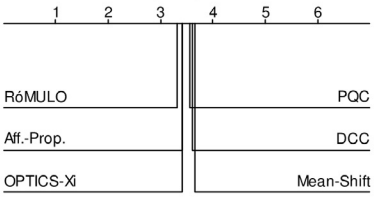
b) Global statistics.



c) Boxplot chart.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	3.33		N/A
Aff.-Prop.	3.41		< 0.01
Mean-Shift	3.66	< 0.01	0.00
OPTICS-Xi	3.42		< 0.01
PQC	3.56		0.00
DCC	3.62		0.00

d) Statistical analysis.



e) Critical diference diagram.

with roughly ten thousand 4 570-dimensional data. The problem with this dataset is that the data are too deeply interwoven[2] and RóMULO could not find a single subspace to make them apart. In this particular case, the competitors based on OPTICS-Xi and PQC failed to produce any results and the competitor based on DCC attained a coefficient of −0.79, but the competitors

[2] The datasets in which RóMULO performed the worst were explored as follows: they were projected onto 100 random subspaces of attributes varying the ratio of selected attributes from 10% to 100% in equally-sized increments; the projections were then re-projected onto a three-dimensional space using the well-known *t*-SNE method; the resulting data were analysed visually and clustered using the three classical methods in the experimental study. If the experimenter could not spot any compact and isolated clusters and the clusterings computed were degraded or resulted in very poor effectiveness coefficients, the conclusion was then that the data in those datasets are too deeply interwoven. That makes it difficult to set them apart without the help of a data engineer who can really understand the meaning of the data and performs some attribute engineering.

based on Affinity-Propagation and Mean-shift attained positive results; the latter were designed to find the clusters by selecting a number of data that is grown/shifted to form the clusters using affinity/density criteria, which generally helps them identify the clusters in the subspace of attributes returned by GSPPCA more easily than RóMULO can do on the original attributes. The statistical analysis confirms that RóMULO ranks at position 2.39 and it is followed by the competitors based on Mean-Shift at position 2.44, Affinity-Propagation at position 2.92, OPTICS-Xi at position 3.74, DCC at position 4.20, and PQC at position 5.31. The p-value returned by the statistical tests is zero or nearly zero in each comparison, which is a strong indication that the experimental results support the hypothesis that RóMULO ranks the first regarding the Silhouette coefficient.

Regarding the Davies-Bouldin coefficient, cf. Table 5, RóMULO ranks at the first position and it is followed by the competitors based on Mean-Shift, Affinity-Propagation, OPTICS-Xi, DCC, and PQC. RóMULO attains an average Davies-Bouldin

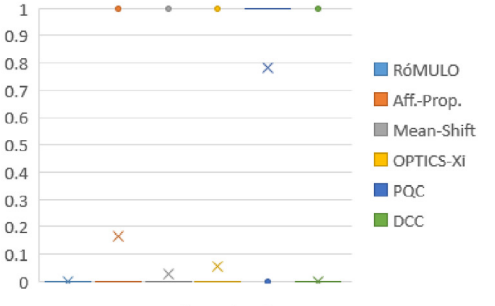
Table 9
Clustering power: Timeouts.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC	Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Adolescents NIH	0.00	0.05	0.00	0.14	0.82	0.00	NBA Raptors	0.00	0.00	0.13	0.13	0.88	0.13
Agrifood Exports	0.00	0.00	0.00	0.00	0.00	0.00	Puerto Rico Media	0.00	0.00	0.00	0.00	0.00	0.00
Agro climatic	0.00	0.15	0.00	0.00	0.98	0.00	Regular Force Outflow	0.00	0.00	0.00	0.00	0.00	0.00
Argentina Climate	0.00	0.00	0.00	0.00	0.00	0.00	Rental Property NY	0.00	0.13	0.00	0.00	1.00	0.00
Art Funding	0.00	0.00	0.00	0.00	0.43	0.00	Riddler Castles	0.00	0.00	0.00	0.00	1.00	0.00
Beaches Euskadi	0.00	0.00	0.00	0.00	0.00	0.00	Rockfall Risk	0.00	0.00	0.00	0.00	0.70	0.00
Brasilian Taxes	0.00	0.00	0.00	0.00	0.71	0.00	Second Language	0.00	0.15	0.00	0.00	0.35	0.00
Citizen Workforce	0.00	0.00	0.00	0.00	0.57	0.00	Steller Sea Lions	0.00	0.02	0.00	0.02	0.98	0.00
COVID	0.00	0.00	0.00	0.00	0.48	0.00	Tax Filers	0.00	0.43	0.00	0.05	1.00	0.00
Crimes	0.00	0.03	0.00	0.00	0.93	0.00	Teen Pregnancy	0.00	0.11	0.00	0.00	0.67	0.00
Deputies Canary	0.00	0.00	0.00	0.00	0.00	0.00	Tourist Accomodation	0.00	0.00	0.05	0.00	0.52	0.00
Disasters	0.00	0.11	0.05	0.16	0.53	0.00	Trump world trust	0.00	0.00	0.00	0.00	0.00	0.00
Divorces	0.00	0.35	0.00	0.00	0.24	0.00	Tuition Fees	0.00	0.00	0.00	0.04	1.00	0.00
Doctoral Graduates	0.00	0.00	0.00	0.00	0.94	0.00	TV Access Services	0.00	0.00	0.00	0.00	0.60	0.00
EU Fruit Data	0.00	0.33	0.00	0.00	0.57	0.00	TV Commercials	0.00	0.15	0.08	0.08	1.00	0.00
Farm Operators	0.00	0.14	0.00	0.00	0.83	0.00	UBER Trips	0.00	0.00	0.00	0.00	0.18	0.00
Gas Emissions	0.00	0.18	0.00	0.06	0.47	0.00	University Spending	0.00	0.00	0.00	0.00	1.00	0.00
Hiking Tours	0.00	0.00	0.00	0.00	0.75	0.00	US Weather History	0.00	0.00	0.00	0.00	0.00	0.00
Homicide Victims	0.00	0.18	0.00	0.00	0.27	0.00	Vehicle Pedestrian Inv	0.00	0.04	0.00	0.00	0.98	0.00
HUSA data	0.00	0.38	0.17	0.14	1.00	0.00	Voter Turnout	0.00	0.00	0.00	0.00	1.00	0.00
Marriage	0.00	0.00	0.00	0.00	0.50	0.00	WiFi Hotspots	0.00	0.00	0.00	0.00	0.20	0.00
Microsoft Kinetic	0.00	0.10	0.00	0.00	0.60	0.00	Women World Cup	0.00	0.00	0.00	0.00	0.00	0.00
National Households	0.00	0.04	0.03	0.73	0.77	0.00	World Cup	0.00	0.00	0.00	0.00	0.00	0.00

a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	0.00	0.00	0.00	0.00	0.00	0.00
25th percentile	0.00	0.00	0.00	0.00	1.00	0.00
Median	0.00	0.00	0.00	0.00	1.00	0.00
75th percentile	0.00	0.00	0.00	0.00	1.00	0.00
Maximum	0.00	1.00	1.00	1.00	1.00	1.00
Average	0.00	0.17	0.02	0.05	0.78	0.00
Standard deviation	0.00	0.37	0.16	0.22	0.41	0.02
Variance	0.00	0.14	0.02	0.05	0.17	0.00
Confidence (95%)	0.00	0.01	0.01	0.01	0.02	0.00

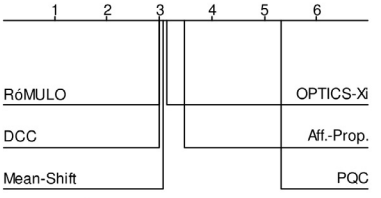
b) Global statistics.



c) Boxplot chart.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	2.99	0.00	N/A
Aff.-Prop.	3.49		0.00
Mean-Shift	3.06		0.03
OPTICS-Xi	3.15		< 0.01
PQC	5.33		0.00
DCC	2.99		0.49

d) Statistical analysis.



e) Critical diference diagram.

coefficient of 1.95, with a minimum that is nearly zero and a maximum of 12.25. The maximum occurs in a dataset from the "WiFi Hotspots" data lake with roughly thirteen thousand 18-dimensional data. Note that the dataset is relatively large in data, but small in dimensionality. The problem with this dataset is that the data are too deeply interwoven in all possible subspaces, which resulted in very bad results using all of the competitors. In this particular case, the competitors based on Mean-Shift and PQC failed to produce any results; the competitors based on OPTICS-Xi and DCC produced clusterings with the highest Davies-Bouldin coefficients; the exception was the competitor based on Affinity-Propagation which attained a coefficient as small as 0.80 but output a large number of small clusters due to its approach to select a subset of data and grow/shift them to form the clusters. The statistical analysis confirms that RóMULO ranks at position 2.42 and it is followed by the competitors based on Mean-Shift at position 2.50, Affinity-Propagation at position 2.84, OPTICS-Xi at position 3.13, DCC at position 4.76, and PQC at position 5.35. Note that the p-value returned by the statistical tests is zero or nearly zero in each comparison, which is a strong indication that the experimental results support the hypothesis that RóMULO ranks the first regarding the Davies-Bouldin coefficient.

Regarding the Caliński-Harabasz coefficient, cf. Table 6, RóMULO ranks at the first position and it is followed by the competitors based on Affinity-Propagation, Mean-Shift, OPTICS-Xi, DCC, and PQC. RóMULO attains an average Caliński-Harabasz coefficient of 197.08, with a minimum of 0.74 and a maximum of 993.36. The minimum occurs in a dataset from the "WiFi Hotspots" data lake with six 14-dimensional data. This is a very small dataset both in terms of number of data and attributes, but there is not a single subspace in which good clusters can be found because the data are deeply interwoven regarding all of the attributes. In this particular case, the competitors returned degenerated clusterings. The statistical analysis confirms that RóMULO ranks at position 2.12 and it is followed by the competitors based on Affinity-Propagation at position 2.67, Mean-Shift at position 2.94, OPTICS-Xi at position 3.17, DCC at position 4.78, and PQC at position 5.32. Note that the p-value returned by the statistical tests is zero in each comparison, which is a strong indication that the experimental results support the hypothesis that RóMULO ranks the first regarding the Caliński-Harabasz coefficient.

Regarding efficiency, cf. Table 7, RóMULO ranks at the fifth position. The statistical tests return a zero p-value for every comparison, which is a strong indication that the differences in rank are significant. Clearly, the search strategy used in RóMULO has a negative impact on its efficiency, which is clearly compensated because of its ability to find better clusterings.

Table 10
Classification power: Adjusted Rand.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Kaggle	0.07	-0.17	-0.14	-0.14	-0.66	-0.32
OpenML	0.00	-0.22	-0.17	-0.11	-0.60	-0.50
Sckit-Learn	0.15	-0.09	-0.50	-0.48	-0.73	-0.73
Synthetic	0.03	0.04	0.36	0.02	-0.73	-0.53
UCI	0.16	0.10	0.11	0.02	-0.21	-0.57

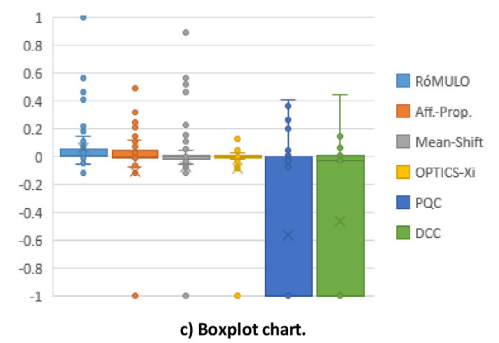
a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	-0.12	-1.00	-1.00	-1.00	-1.00	-1.00
25th percentile	-0.00	-0.00	-0.01	-0.01	-1.00	-1.00
Median	0.06	0.04	0.02	< 0.01	< 0.01	0.01
75th percentile	1.00	0.51	0.89	0.14	0.41	0.45
Maximum	1.00	0.51	0.89	0.14	0.41	0.45
Average	0.06	-0.14	-0.09	-0.10	-0.58	-0.48
Standard deviation	0.15	0.42	0.40	0.32	0.52	0.52
Variance	0.02	0.17	0.16	0.10	0.27	0.27
Conf. Int. at 95%	0.03	0.08	0.08	0.06	0.10	0.10

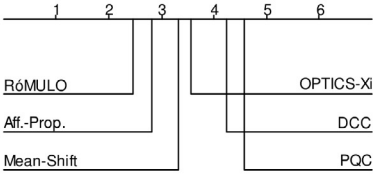
b) Global statistics.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	2.46		N/A
Aff.-Prop.	2.83		< 0.01
Mean-Shift	3.33	0.00	< 0.01
OPTICS-Xi	3.56		< 0.01
PQC	4.59		0.00
DCC	4.25		0.00

d) Statistical analysis.



c) Boxplot chart.



e) Critical diference diagram.

Note that the experiments run with RóMULO took from a minimum of 0.01 minutes to a maximum of 187.20 minutes. The minimum was achieved on the smallest datasets; the maximum occurred on a dataset in the "Steller Sea Lions" data lake that has 37270 data with 181 attributes each. Note that this is not the largest dataset in the experimental repositories, but one in which RóMULO could not easily find the appropriate subspace of informative attributes. The other proposals were also inefficient when working on this dataset and the reason was, again, that it provides many uninformative attributes that result in too deeply interwoven data. Note, too, that the timings vary significantly from data lake to data lake; recall that they consists of real-world datasets whose size ranges from an average minimum of 556.28 data to an average maximum of 7632.70 data per data lake and whose dimensionalities range from an average minimum of 38.76 attributes to an average maximum of 2995.02 attributes; inevitably, that has an important impact on the variance of the clustering time.

Tables 8 and 9 report on the ratio of degraded clusterings and timeouts, respectively. Note that RóMULO is the only proposal that attains 0.00 degradation ratio and 0.00 timeout ratio. The reason is that its search strategy quickly discards individuals that result in degraded clusterings and focus on promising candidates without timing out. Regarding the ratio of degraded clusterings, it is closely followed by the competitors based on Affinity-Propagation and OPTICS-Xi with an average ratio of 0.03, and then come the competitors based on PQC with an average ratio of 0.08, DCC with an average ratio of 0.10, and Mean-Shift with an average ratio of 0.11. Regarding the ratio of timeouts, it is identical to the competitor based on DCC with a ratio of 0.00, it is closely followed by the competitors based on Mean-Shift with an average ratio of 0.02, OPTICS-Xi with an average ratio of 0.05, and then come the competitors based on Affinity-Propagation with an average ratio of 0.17 and PQC with an average ratio of 0.78.

The conclusion is that RóMULO ranks the first regarding the Silhouette, the Davies-Bouldin, and the Caliński-Harabasz coefficients. It is closely followed by the competitors based on Affinity-Propagation and Mean-Shift and then come the competitors based on OPTICS-Xi, DCC, and PQC. It performs quite well regarding efficiency and does not result in any degraded clusterings or timeouts. The statistical analysis confirmed that the differences in rank are significant.

4.5. Analysis of classification power

Tables 10–15 summarise the results regarding the extent to which the clusterings returned by RóMULO and the competitors can be considered good classifications. The structure of these tables is the same as before.

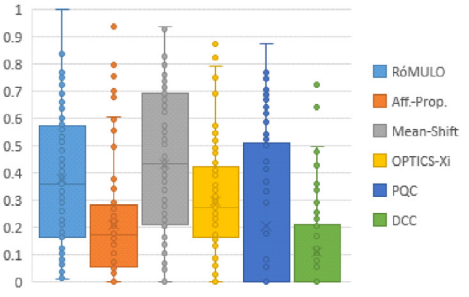
Table 11
Classification power: Fowlkes-Mallows.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Kaggle	0.46	0.21	0.49	0.35	0.21	0.18
OpenML	0.36	0.16	0.37	0.29	0.17	0.09
Sckit-Learn	0.19	0.20	0.12	0.11	0.08	0.03
Synthetic	0.26	0.32	0.61	0.25	0.18	0.11
UCI	0.46	0.36	0.55	0.32	0.39	0.13

a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	0.01	0.00	0.00	0.00	0.00	0.00
25th percentile	0.17	0.06	0.22	0.16	0.00	0.00
Median	0.57	0.28	0.68	0.42	0.51	0.21
75th percentile	1.00	0.93	0.94	0.87	0.87	0.72
Maximum	1.00	0.93	0.94	0.87	0.87	0.72
Average	0.38	0.21	0.43	0.30	0.20	0.11
Standard deviation	0.24	0.21	0.28	0.20	0.28	0.16
Variance	0.06	0.04	0.08	0.04	0.08	0.03
Conf. Int. at 95%	0.05	0.04	0.05	0.04	0.05	0.03

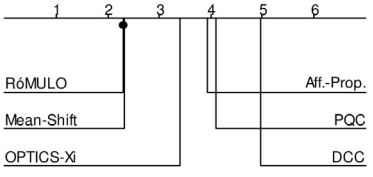
b) Global statistics.



c) Boxplot chart.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	2.30		N/A
Aff.-Prop.	3.92		0.00
Mean-Shift	2.33	0.00	0.15
OPTICS-Xi	3.40		< 0.01
PQC	4.11		0.00
DCC	4.96		0.00

d) Statistical analysis.



e) Critical diference diagram.

Table 12
Classification power: Accuracy.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Kaggle	0.51	0.42	0.40	0.45	0.18	0.41
OpenML	0.51	0.48	0.35	0.43	0.18	0.29
Sckit-Learn	0.48	0.72	0.03	0.30	0.12	0.22
Synthetic	0.60	0.73	0.56	0.39	0.16	0.31
UCI	0.65	0.64	0.48	0.58	0.37	0.23

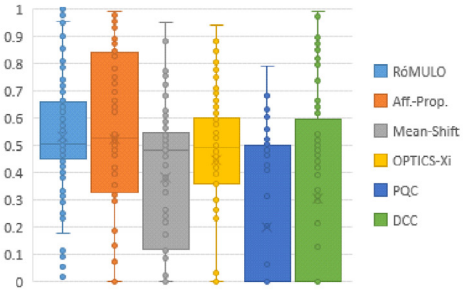
a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	0.02	0.00	0.00	0.00	0.00	0.00
25th percentile	0.46	0.35	0.12	0.36	0.00	0.00
Median	0.65	0.83	0.54	0.60	0.50	0.56
75th percentile	1.00	0.99	0.95	0.94	0.79	0.99
Maximum	1.00	0.99	0.95	0.94	0.79	0.99
Average	0.53	0.52	0.38	0.45	0.20	0.31
Standard deviation	0.21	0.32	0.25	0.23	0.26	0.35
Variance	0.04	0.10	0.06	0.05	0.07	0.12
Conf. Int. at 95%	0.04	0.06	0.05	0.05	0.05	0.07

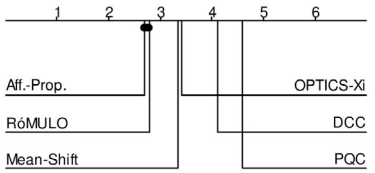
b) Global statistics.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	2.81	0.00	0.20
Aff.-Prop.	2.71		N/A
Mean-Shift	3.36		0.02
OPTICS-Xi	3.43		< 0.01
PQC	4.58		0.00
DCC	4.13		0.00

d) Statistical analysis.



c) Boxplot chart.



e) Critical diference diagram.

Regarding the Adjusted Rand coefficient, cf. Table 10, RóMULO ranks at the first position and it is followed by the competitors based on Affinity-Propagation, Mean-Shift, OPTICS-Xi, DCC, and PQC. RóMULO attains an average Adjusted Rand coefficient of 0.04, with a minimum of −0.12 and a maximum of 1.00. The minimum occurs in a dataset from the "OpenML" data lake that has seven 1024-dimensional data with three classes. The other competitors attained similar values for this coefficient, ranging from the competitor based on Affinity-Propagation, which attained exactly the same value, to the competitor based on DCC, which attained a value of 0.01. Note that this is a particularly difficult dataset[3] because there are too few data, but they are highly dimensional and the number of sample data per class is very low; simply put, this dataset provides little evidence to learn a good classifier. The best value was attained on a dataset from the "Kaggle" data lake with roughly three hundred 21-dimensional data with two classes. The other competitors attained a value that was nearly 0.00 with this dataset, which is peculiar since there are many pairs of attributes that result in perfect linear classifiers. Such attributes are quickly discovered by RóMULO; unfortunately, GSPPCA selects five attributes whose values are too interwoven and make it difficult for the competitors to make the classes apart. The statistical analysis confirms that RóMULO ranks at position 2.46 and it is followed by the competitors based on Affinity-Propagation at position 2.83, Mean-Shift at position 3.33, OPTICS-Xi at position 3.56, DCC at position 4.25, and PQC at position 4.59. Note that the p-value returned by the statistical tests is below the significance level in every comparison, which supports the hypothesis that RóMULO ranks the first regarding the Adjusted Rand coefficient.

Regarding the Fowlkes-Mallows coefficient, cf. Table 11, RóMULO ranks at the first position and it is closely followed by the competitor based on Mean-Shift; then come the competitors based on OPTICS-Xi, Affinity-Propagation, PQC, and DCC. RóMULO attains an average Fowlkes-Mallows coefficient of 0.38, with a minimum of 0.01 and a maximum of 1.00. The worst value is attained with a dataset from the "OpenML" data lake that provides 21427 data with 18 attributes and 287 classes. The competitors based on Affinity-Propagation, PQC, and DCC timed out with this dataset; only the competitors based on Mean-Shift and OPTICS-Xi could process it and they attained a Fowlkes-Mallows coefficient of 0.11 and 0.08, respectively.

[3] The datasets in which RóMULO performed the worst were explored as follows: a ROC analysis was performed using Random Forest (tree learning), J4.8 (rule learning), Naive Bayes (Bayesian learning), and a five-layer Perceptron (neural leaning); *k*-means was also run by setting *k* to the number of classes in the corresponding dataset. The data were considered too deeply interwoven if the average ROC coefficient was close to 0.50, which basically indicates that the classifiers learnt make random predictions, or if the effectiveness coefficients were very poor. Such datasets are particularly difficult to deal with because they require a data engineer who can really understand their meaning to perform some attribute engineering.

Table 13
Classification power: Time.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Kaggle	47.61	47.61	36.02	35.93	163.92	81.82
OpenML	64.14	64.14	46.13	32.41	150.38	127.44
Sckit-Learn	182.82	182.82	240.50	240.43	303.12	300.41
Synthetic	55.43	55.43	54.53	54.30	240.09	187.04
UCI	0.61	0.61	0.32	0.09	75.01	147.86

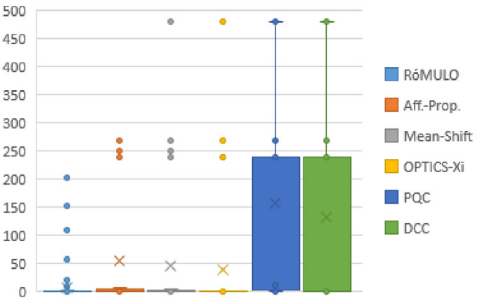
a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	0.01	< 0.01	< 0.01	< 0.01	0.09	0.03
25th percentile	0.06	0.03	0.04	0.02	2.23	0.24
Median	1.61	5.64	1.89	0.62	240.07	240.01
75th percentile	203.35	267.69	480.00	480.00	480.00	480.00
Maximum	203.35	267.69	480.00	480.00	480.00	480.00
Average	7.27	56.37	46.58	39.24	157.61	132.80
Standard deviation	28.13	102.20	107.17	101.74	145.17	137.97
Variance	791.20	10,444.45	11,486.10	10,350.70	21,075.43	19,036.92
Conf. Int. at 95%	5.51	20.03	21.01	19.94	28.45	27.04

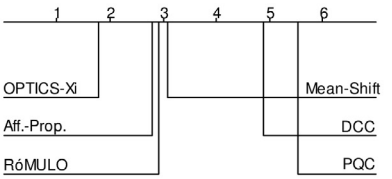
b) Global statistics.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	2.92		< 0.01
Aff.-Prop.	2.79		< 0.01
Mean-Shift	3.08	0.00	0.00
OPTICS-Xi	1.79		N/A
PQC	5.53		0.00
DCC	4.90		0.00

d) Statistical analysis.



c) Boxplot chart.



e) Critical diference diagram.

The dataset was inspected and the conclusion was that the attributes had very few different values, which made it really difficult to make the 287 classes apart; note that it was not a problem with the amount of data, which is relatively large, but the few attributes and values in comparison with the relatively large number of classes. The best value was attained with the same dataset that maximised the Adjusted Rand coefficient; the competitors behaved similarly regarding the Fowlkes-Mallows coefficient. Basically, RóMULO could easily find the subspace of attributes that helps make the classes apart, but GSPPCA selected five attributes that did not help in this task. The statistical analysis confirms that RóMULO and the competitor based on Mean-Shift share the first position in the ranking because the p-value from their comparison is above the significance level; then come the competitors that are based on OPTICS-Xi, Affinity-Propagation, PQC, and DCC, whose comparisons result in statistically-significant differences because the returned p-values are zero or nearly zero.

Regarding the Accuracy coefficient, cf. Table 12, the competitor that is based on Affinity-Propagation ranks at the first position and it is followed by RóMULO and the competitors that are based on Mean-Shift, OPTICS-Xi, DCC, and PQC. RóMULO attains an average Accuracy coefficient of 0.53, with a minimum of 0.02 and a maximum of 1.00. The worst value is attained in a dataset from the "OpenML" data lake that has 1 480 data with 10 000 attributes and 50 classes. The competitors based on PQC and DCC timed out on this dataset and the competitors based on Mean-Shift and OPTICS-Xi attained the same accuracy as RóMULO on it; only the competitor based on Affinity-Propagation could attain an accuracy of 0.40. The dataset was inspected and the conclusion was that its data are, again, far too interwoven to make the classes clearly apart using clustering; neither was RóMULO able to find any good subspaces of attributes. The best value was attained with the same dataset as was the case regarding the previous coefficients. The statistical analysis confirms that the competitor based on Affinity-Propagation ranks at the first position and it is closely followed by RóMULO, which is statistically indistinguishable from it because the p-value returned by Hommel's test is above the significance level; then come the competitors that are based on Mean-Shift, OPTICS-Xi, DCC, and PQC, whose comparisons result in statistically-significant differences because the returned p-value is below the significance level in each case.

Regarding efficiency, cf. Table 13, RóMULO ranked at the third position in these experiments; the results indicate that the competitor based on OPTICS-Xi is the most efficient one, which is closely followed by the one based on Affinity-Propagation; the competitors that are based on Mean-Shift, DCC, and PQC rank behind RóMULO. From the previous experimentation, it is clear that the search strategy of RóMULO has a negative impact on its efficiency, but it generally finds better clusterings; the

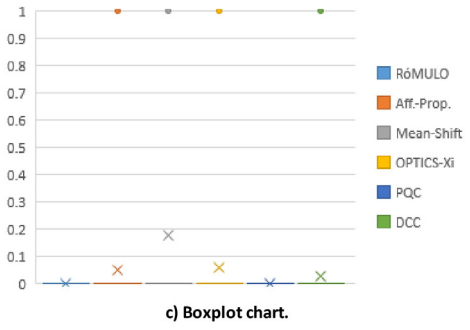
Table 14
Classification power: Degradation.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Kaggle	0.00	0.05	0.19	0.00	0.00	0.05
OpenML	0.00	0.00	0.17	0.08	0.00	0.04
Sckit-Learn	0.00	0.00	0.50	0.00	0.00	0.00
Synthetic	0.00	0.11	0.33	0.22	0.00	0.00
UCI	0.00	0.23	0.00	0.00	0.00	0.00

a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	0.00	0.00	0.00	0.00	0.00	0.00
25th percentile	0.00	0.00	0.00	0.00	0.00	0.00
Median	0.00	0.00	0.00	0.00	0.00	0.00
75th percentile	0.00	1.00	1.00	1.00	0.00	1.00
Maximum	0.00	1.00	1.00	1.00	0.00	1.00
Average	0.00	0.05	0.18	0.06	0.00	0.03
Standard deviation	0.00	0.22	0.39	0.24	0.00	0.17
Variance	0.00	0.05	0.15	0.06	0.00	0.03
Conf. Int. at 95%	0.00	0.04	0.08	0.05	0.00	0.03

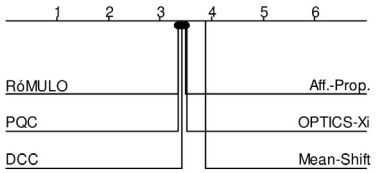
b) Global statistics.



c) Boxplot chart.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	3.34	0.33	N/A
Aff.-Prop.	3.49		0.46
Mean-Shift	3.88		0.01
OPTICS-Xi	3.52		0.46
PQC	3.34		0.50
DCC	3.43		0.50

d) Statistical analysis.



e) Critical diference diagram.

experimentation to check how powerful it is for classification purposes confirms the previous finding, since it can work in reasonable times and attain very good classification scores. It takes from a minimum of 0.01 minutes on a small dataset from the "OpenML" data lake with 85 data that have four attributes and four classes to a maximum of 57.93 minutes on a large dataset from the "Kaggle" data lake with 51 047 data that have 874 attributes and two classes; the other competitors timed out on this dataset where RóMULO could attain an Adjusted Rand Coefficient of 0.00, a Fowlkes-Mallows coefficient of 0.22, and an Accuracy coefficient of 0.43.

Tables 14 and 15 report on the ratio of degraded clusterings and timeouts, respectively. Note that RóMULO and the competitor based on PQC are the only proposals that attain 0.00 degradation ratio in this experimentation; in the case of RóMULO, this is again due to its ability to quickly discard degraded clusterings; in the case of the competitor based on PQC, the result is not that surprising since it was the proposal that resulted in less degraded ratios in the previous experimentation. The statistical analysis confirms that RóMULO is closely followed by the proposals based on PQC, DCC, Affinity-Propagation, and OPTICS-Xi, which are indistinguishable from it because the p-values returned for these comparisons are clearly above the significance level; the last position in the rank is for the competitor based on Mean-Shift and the difference was confirmed to be statistically significant because the p-value returned was clearly below the significance level. Regarding the ratio of timeouts, RóMULO is again the only proposal that could process all of the datasets within the time constraints that were set; then come the competitors based on OPTICS-Xi with a ratio of 0.09, Mean-Shift with a ratio of 0.12, Affinity-Propagation with a ratio of 0.16, DCC with a ratio of 0.48, and PQC with a ratio of 0.58. The statistical analysis makes it clear that RóMULO and the competitors based on OPTICS-Xi, Mean-Shift, and Affinity-Propagation are statistically indistinguishable because the p-value returned by Hommel's test is clearly above the significance level; it also makes it clear that the competitors based on DCC and PQC rank below the previous group because the p-value computed is zero in both cases and this implies that the differences in rank are very significant.

The conclusion is that RóMULO ranks the first regarding the Adjusted Rand, the Fowlkes-Mallows, and the Accuracy coefficients. Regarding the first coefficient, the differences in rank are statistically significant with regard to all of the other competitors; regarding the second and the third coefficient RóMULO shares the first position in the rank with the competitors based on Mean-shift and Affinity-Propagation, respectively. It is closely followed by the competitors based on Affinity-Propagation, Mean-Shift, and OPTICS-Xi, and then come the competitors based on DCC and PQC. It performs quite well regarding efficiency and does not result in any degraded clusterings or timeouts.

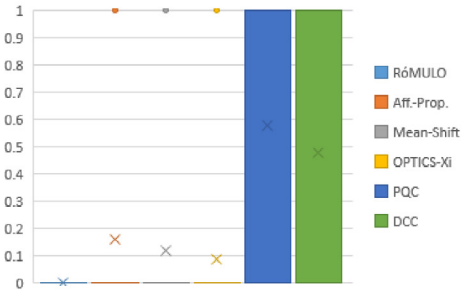
Table 15
Classification power: Timeouts.

Data lake	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Kaggle	0.00	0.19	0.14	0.14	0.67	0.33
OpenML	0.00	0.21	0.13	0.08	0.57	0.47
Sckit-Learn	0.00	0.25	0.50	0.50	0.75	0.75
Synthetic	0.00	0.00	0.00	0.00	0.78	0.56
UCI	0.00	0.00	0.00	0.00	0.31	0.62

a) Average results per data lake.

Statistic	RóMULO	Aff.-Prop.	Mean-Shift	OPTICS-Xi	PQC	DCC
Minimum	0.00	0.00	0.00	0.00	0.00	0.00
25th percentile	0.00	0.00	0.00	0.00	0.00	0.00
Median	0.00	0.00	0.00	0.00	1.00	1.00
75th percentile	0.00	1.00	1.00	1.00	1.00	1.00
Maximum	0.00	1.00	1.00	1.00	1.00	1.00
Average	0.00	0.16	0.12	0.09	0.58	0.48
Standard deviation	0.00	0.37	0.33	0.29	0.50	0.50
Variance	0.00	0.14	0.11	0.08	0.25	0.25
Conf. Int. at 95%	0.00	0.07	0.06	0.06	0.10	0.10

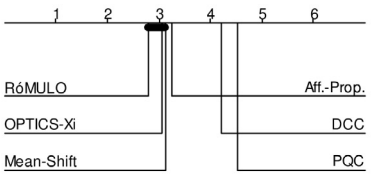
b) Global statistics.



c) Boxplot chart.

Proposal	Rank	Im.-Dav.	Hommel
RóMULO	2.79		N/A
Aff.-Prop.	3.27		0.02
Mean-Shift	3.15	< 0.01	0.05
OPTICS-Xi	3.06		0.07
PQC	4.53		0.00
DCC	4.23		0.00

d) Statistical analysis.



e) Critical diference diagram.

5. Conclusions

This article introduces a new clustering proposal called RóMULO. It is intended to help data engineers extract new knowledge from their data lakes by finding compact and isolated clusters in a subspace of informative attributes. It explores a research niche that has not been studied before: using a genetic meta-heuristic to perform multi-way single-subspace automatic clustering.

RóMULO was confronted with five strong competitors that combine the state-of-the-art attribute selection proposal with three classical single-way clustering proposals, a recent quantum-inspired one, and a recent deep-learning one. They were explored regarding their ability to find compact and isolated clusterings and the extent to which those clusterings can be considered good classifications. The first exploration was performed using the Silhouette, the Davies-Bouldin, and the Caliński-Harabasz coefficients; the second one was performed using the Adjusted Rand, the Fowlkes-Mallows, and the Accuracy coefficients. Furthermore, it was very efficient regarding CPU time and it did not result in any degraded clusterings or timeouts. The conclusions regarding the rankings were supported by means of the corresponding statistical analyses at the standard significance level.

Future research includes exploring how an engineer can include some background knowledge into the clustering process, e.g., to flag some data as belonging to the same or different clusters or to provide custom attribute transformers in the case of non-numeric attributes. Exploring whether some ideas based on collaborative filtering might help improve the efficiency in the case of large datasets is also worth exploring. Much attention shall also be paid to exploring other meta-heuristics, e.g., coral-reef or quantum optimisation, and deep learning.

CRediT authorship contribution statement

Patricia Jiménez: Conceptualization, Methodology, Validation, Resources, Data curation, Writing – review & editing. **Juan C. Roldán:** Conceptualization, Methodology, Validation, Resources, Data curation, Writing – review & editing. **Rafael Corchuelo:** Conceptualization, Methodology, Software, Validation, Resources, Data curation, Writing – original-draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors were partially supported by the Spanish National Research Council (TIN2016-75394-R, PID2020-112540RB-C44) and the Andalusian Research Council (P18-RT-1060, US-1381375). The work by Juan C. Roldán was also supported by a PIF grant from the University of Seville. The authors are very grateful to Dinamic Area, S.L. for providing an ideal industrial context and the machinery required to perform the experimental evaluation.

References

[1] Charu C. Aggarwal, An introduction to cluster analysis, in: Data clustering: algorithms and applications, CRC Press, 2013, pp. 1–28.
[2] Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle, Saeed Ur Rehman, Research on particle swarm optimization based clustering: a systematic review of literature and techniques, Swarm Evol. Comput. 17 (2014) 1–13, https://doi.org/10.1016/j.swevo.2014.02.001.
[3] Daniel Aloise, Amit Deshpande, Pierre Hansen, Preyas Popat, NP-hardness of Euclidean sum-of-squares clustering, Mach. Learn. 75 (2) (2009) 245–248, https://doi.org/10.1007/s10994-009-5103-0.
[4] Amazon, Inc. Data lakes on Amazon Web Services. URL: https://aws.amazon.com/en/solutions/implementations/data-lake-solution, 2021. Accessed: 2021-10-12..
[5] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, Íñigo Perona, An extensive comparative study of cluster validity indices, Pattern Recogn. 46 (1) (2013) 243–256, https://doi.org/10.1016/j.patcog.2012.07.021.
[6] Hans-Georg Beyer, Hans-Paul Schwefel, Evolution strategies: a comprehensive introduction, Nat. Comput. 1 (1) (2002) 3–52, https://doi.org/10.1023/A:1015059928466.
[7] Chin-Wei Bong, Mandava Rajeswari, Multi-objective nature-inspired clustering and classification techniques for image segmentation, Appl. Soft Comput. 11 (4) (2011) 3271–3282, https://doi.org/10.1016/j.asoc.2011.01.014.
[8] Charles Bouveyron, Pierre Latouche, Pierre-Alexandre Mattei, Bayesian variable selection for globally sparse probabilistic PCA, Electron. J. Stat. 12 (2) (2018) 3036–3070, https://doi.org/10.1214/18-EJS1450.
[9] Zhaohong Deng, Kup-Sze Choi, Yizhang Jiang, Jun Wang, Shitong Wang, A survey on soft subspace clustering, Inf. Sci. 348 (2016) 84–106, https://doi.org/10.1016/j.ins.2016.01.101.
[10] Michel M. Deza, Elena Deza, Encyclopedia of distances. Springer, 4th edition,., 2016.
[11] Tansel Dökeroglu, Ender Sevinç, Tayfun Kucukyilmaz, Ahmet Cosar, A survey on new generation meta-heuristic algorithms, Comput. Ind. Eng. 137 (2019), https://doi.org/10.1016/j.cie.2019.106040.
[12] Raúl V. Casaña, Paulo J.G. Eslava, Sandra Ortega-Martorell Lisboa, Ian H. Jarman, José D. Martín-Guerrero, Probabilistic quantum clustering, Knowl.-Based Syst. 194 (2020), https://doi.org/10.1016/j.knosys.2020.105567 105567.
[13] Elliackin Figueiredo, Mariana Macedo, Hugo Valadares Siqueira, Clodomir J. Santana, Anu Gokhaled, Carmelo J.A. Bastos-Filhoa, Swarm intelligence for clustering: a systematic review with new perspectives on data mining, Eng. Appl. AI 82 (2019) 313–329, https://doi.org/10.1016/j.engappai.2019.04.007.
[14] Adán J. García, Wilfrido Gómez-Flores, Automatic clustering using nature-inspired meta-heuristics: a survey, Appl. Soft Comput. 41 (2016) 192–213, https://doi.org/10.1016/j.asoc.2015.12.001.
[15] Paolo Lo Giudice, Lorenzo Musarella, Giuseppe Sofo, Domenico Ursino, An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake, Inf. Sci. 478 (2019) 606–626, https://doi.org/10.1016/j.ins.2018.11.052.
[16] Raymond Greenlaw, Sanpawat Kantabutra, Survey of clustering: algorithms and applications, J. Inf. Retrieval Res. 3 (2) (2013) 1–29, https://doi.org/10.4018/ijirr.2013040101.
[17] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. ACM Comput. Surv. 51(5): 93:1–93:42, 2019. doi: 10.1145/3236009..
[18] Jiawei Han, Micheline Kamber, Jian Pei, Cluster analysis: basic concepts and methods, in: Data Mining, Morgan Kaufmann, third edition, 2012, pp. 443–495, https://doi.org/10.1016/B978-0-12-381479-1.00010-1.
[19] Thomas Helmuth, Lee Spector, James Matheson, Solving uncompromising problems with Lexicase selection, IEEE Trans. Evol. Comput. 19 (5) (2015) 630–643, https://doi.org/10.1109/TEVC.2014.2362729.
[20] Christian Hennig, What are the true clusters?, Pattern Recogn Lett. 64 (2015) 53–62, https://doi.org/10.1016/j.patrec.2015.04.009.
[21] Eduardo R. Hruschka, Ricardo J.G.B. Campello, Alex A. Freitas, and André C.P.L.F. de Carvalho. A survey of evolutionary algorithms for clustering. IEEE Trans. Syst. Man Cybern. Part C 39 (2): 133–155, 2009. doi: 10.1109/TSMCC.2008.2007252..
[22] IBM, Inc. Data lake solutions. URL: https://www.ibm.com/analytics/data-lake, 2021. Accessed: 2021-10-12..
[23] Anil K. Jain, Data clustering: 50 years beyond k-means, Pattern Recogn. Lett. 31 (8) (2010) 651–666, https://doi.org/10.1016/j.patrec.2009.09.011.
[24] Rezaul Karim, Oya Beyan, Achille Zappa, Ivan G. Costa, Dietrich Rebholz-Schuhmann, Michael Cochez, Stefan Decker, Deep learning-based clustering approaches for bioinformatics, Briefings Bioinform. 22 (1) (2021) 393–415, https://doi.org/10.1093/bib/bbz170.
[25] Meike Klettke, Hannes Awolin, Uta Störl, Daniel Müller, and Stefanie Scherzinger. Uncovering the evolution history of data lakes. In BigData, pages 2462–2471, 2017. doi: 10.1109/BigData.2017.8258204..
[26] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. TKDD, 3 (1): 1:1–1:58, 2009. doi: 10.1145/1497577.1497578..
[27] William la Cava, Thomas Helmuth, Lee Spector, Jason H. Moore, A probabilistic and multi-objective analysis of Lexicase selection and <texmath type="inline">epsilon </texmath>-Lexicase selection, Evol. Comput. J. (2019) 1–26, https://doi.org/10.1162/evco_a_00224.
[28] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, Huan Liu, Feature selection: a data perspective, ACM Comput. Surv. 50 (6) (2018) 94:1–94:45, https://doi.org/10.1145/3136625.
[29] Boris Lorbeer, Ana Kosareva, Bersant Deva, Dzenan Softic, Peter Ruppel, Axel Küpper, Variations on the clustering algorithm Birch, Big Data Res. 11 (2018) 44–53, https://doi.org/10.1016/j.bdr.2017.09.002.
[30] José M. Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, José C. Riquelme-Santos, An approach to validity indices for clustering techniques in Big Data, Prog. AI 7 (2) (2018) 81–94, https://doi.org/10.1007/s13748-017-0135-3.
[31] José M. Luna-Romera, María Martínez-Ballesteros, Jorge García-Gutiérrez, José C. Riquelme-Santos, External clustering validity index based on chi-squared statistical test, Inf. Sci. 487 (2019) 1–17, https://doi.org/10.1016/j.ins.2019.02.046.
[32] S. Manochandar, M. Punniyamoorthy, Ramasamy K. Jeyachitra, Development of new seed with modified validity measures for k-means clustering, Comput. Ind. Eng. 141 (2020), https://doi.org/10.1016/j.cie.2020.106290 106290.

[33] Christian Mathis, Data lakes, Datenbank-Spektrum 17 (3) (2017) 289–293, https://doi.org/10.1007/s13222-017-0272-7.
[34] Microsoft, Inc. Data lakes. URL: https://azure.microsoft.com/en-us/solutions/data-lake, 2021. Accessed: 2021-10-12..
[35] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, Jun Long, A survey of clustering with deep learning from the perspective of network architecture, IEEE Access 6 (2018) 39501–39514, https://doi.org/10.1109/ACCESS.2018.2855437.
[36] Satyasai J. Nanda, Ganapati Panda, A survey on nature inspired meta-heuristic algorithms for partitional clustering, Swarm Evol. Comput. 16 (2014) 1–18, https://doi.org/10.1016/j.swevo.2013.11.003.
[37] Frank Neumann, Carsten Witt, Bioinspired computation in combinatorial optimization, Springer, 2010, https://doi.org/10.1007/978-3-642-16544-3, ISBN 978-3-642-16543-6.
[38] Satoshi Oyama, Katsumi Tanaka. How many objects? Determining the number of clusters with a skewed distribution. In ECAI, pages 771–772, 2008. doi: 10.3233/978-1-58603-891-5-771..
[39] Lakshmi Patibandla and Naralasetti Veeranjaneyulu. Survey on clustering algorithms for unstructured data. In Intelligent Engineering Informatics, volume 695. Springer, 2018. doi: 10.1007/978-981-10-7566-7_41..
[40] Witold Pedrycz, Collaborative fuzzy clustering, Pattern Recognit. Lett. 23 (14) (2002) 1675–1686, https://doi.org/10.1016/S0167-8655(02)00130-7.
[41] Christoph Quix. Data lakes: a solution or a new challenge for big data integration? In DATA, page 7, 2016..
[42] Sandeep Rana, Sanjay Jasola, Rajesh Kumar, A review on particle swarm optimization algorithms and their applications to data clustering, Artif. Intell. Rev. 35 (3) (2011) 211–222, https://doi.org/10.1007/s10462-010-9191-9.
[43] Sohil Atul Shah and Vladlen Koltun. Deep continuous clustering. CoRR, abs/1803.01449, 2018. URL: http://arxiv.org/abs/1803.01449..
[44] Yinghua Shen, Witold Pedrycz, Collaborative fuzzy clustering algorithm: some refinements, Int. J. Approx. Reason. 86 (2017) 41–61, https://doi.org/10.1016/j.ijar.2017.04.004.
[45] David J. Sheskin, Handbook of parametric and nonparametric statistical procedures, Chapman & Hall/CRC Press (2020).
[46] Kelvin Sim, Vivekanand Gopalkrishnan, Arthur Zimek, Gao Cong, A survey on enhanced subspace clustering, Data Min. Knowl. Discov. 26 (2) (2013) 332–397, https://doi.org/10.1007/s10618-012-0258-x.
[47] Xu. Dongkuan, Yingjie Tian, A comprehensive survey of clustering algorithms, Ann. Data Sci. 2 (2) (2015) 165–193, https://doi.org/10.1007/s40745-015-0040-1.
[48] Xu. Rui, Donald C. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Networks 16 (3) (2005) 645–678, https://doi.org/10.1109/TNN.2005.845141.