symmetry

MDPI

Article

A Novel 2D Clustering Algorithm Based on Recursive Topological Data Structure

Ismael Osuna-Galán [1,†], Yolanda Pérez-Pimentel [1,†] and Carlos Aviles-Cruz [2,*,†]

[1] Departamento de Electrónica, Polytechnic University of Chiapas, Carretera Tuxtla Gutierrez-Portillo Zaragoza Km 21+500, Suchiapa 29150, Mexico; iosuna@upchiapas.edu.mx (I.O.-G.); ypimentel@upchiapas.edu.mx (Y.P.-P.)

[2] Departamento de Electrónica, Universidad Autónoma Metropolitana, Av. San Pablo 180 Col. Reynosa Tamaulipas, Ciudad de México 02200, Mexico

* Correspondence: caviles@azc.uam.mx; Tel.: +52-55-5318-9030

† These authors contributed equally to this work.

Abstract: In the field of data science and data mining, the problem associated with clustering features and determining its optimum number is still under research consideration. This paper presents a new 2D clustering algorithm based on a mathematical topological theory that uses a pseudometric space and takes into account the local and global topological properties of the data to be clustered. Taking into account cluster symmetry property, from a metric and mathematical-topological point of view, the analysis was carried out only in the positive region, reducing the number of calculations in the clustering process. The new clustering theory is inspired by the thermodynamics principle of energy. Thus, both topologies are recursively taken into account. The proposed model is based on the interaction of particles defined through measuring homogeneous-energy criterion. Based on the energy concept, both general and local topologies are taken into account for clustering. The effect of the integration of a new element into the cluster on homogeneous-energy criterion is analyzed. If the new element does not alter the homogeneous-energy of a group, then it is added; otherwise, a new cluster is created. The mathematical-topological theory and the results of its application on public benchmark datasets are presented.

Keywords: clustering; mathematical topology; data clustering; pseudometric; grouping; synthetic datasets

Citation: Osuna-Galán, I.; Pérez-Pimentel, Y.; Aviles-Cruz, C. A Novel 2D Clustering Algorithm Based on Recursive Topological Data Structure. *Symmetry* **2022**, *14*, 781. https://doi.org/10.3390/sym14040781

Academic Editor: Juan Luis García Guirao

Received: 4 March 2022
Accepted: 7 April 2022
Published: 9 April 2022

1. Introduction

Clustering and its classification has increased significantly due to a large amount of digital information available on the internet, especially on social media. The task of grouping information that contains common characteristics or meaning and its subsequent classification is a cornerstone in areas such as data sciences, data mining, and pattern recognition. Despite the important advancement that has been done in the algorithms of clustering and its classifications, it is still under the attention of researchers.

Clustering algorithms are based on their clustering paradigm, and the amount and dimension of the data to be handled. There are several review papers reported in the literature. Saxena et al. [1] have classified them into two main groups, i.e., hierarchical and partitional algorithms. The hierarchical algorithms are further subdivided into agglomerative and divisive algorithms, while partitional algorithms are subdivided into density-based, distance-based, and model-based algorithms.

Another review paper is presented by Dong et al. [2], where the main attention was paid to analyzing the clustering algorithms from supervised and unsupervised perspective. Nevertheless, the most widely accepted classification is that of [3] in which the authors have divided the algorithm into Partitional, Hierarchical, Density-Based, Grid-Based, Spectral,

Gravitational, Neural Network-Based, and Evolutionary-Based. We consider the similar classification given in detail in Section 2.

In this research paper, a new 2D clustering algorithm is presented. The proposal is based on the energy and homogeneity of topological theory. Taking into account cluster symmetry property, from metric and mathematical-topological point of view, the analysis is done only in positive region, reducing the number of calculations in the clustering process. The proposal works according to the homogeneous-energy effect that is calculated when a cluster receives a new element. If the new element does not alter the local homogeneous energy of the cluster, then it is added to the cluster. Otherwise, a new cluster will be generated. The algorithm will terminate when all elements are assigned to a single cluster.

Our novel approach was developed to work in two-dimensional space, and it was tested and validated using the bi-dimensional *Shape* databases widely used in the data clustering literature from the Machine Learning group of School of Computing, University of Eastern Finland [4]. The Shape dataset was chosen because of its levels of complexity, spherically separable point distribution (R15, D31), embedded classes (Jain, flame, and compound), and complex distributions (spiral and pathbased). The previous data distributions are difficult to be correctly clustered by a single clustering algorithm, which is achieved in our proposed methodology.

The rest of the paper is organized as follows. In Section 2, state of the art is presented. The proposal is presented in Section 3. Topology-based theory is given in Section 4. Section 5 presents the methodology. Experimental results and discussion are given in Section 6. Finally, conclusions and perspectives are given in Section 7.

2. State of the Art

Clustering is an unsupervised learning method that classifies unlabeled data objects into several groups based on the similarities among them. The main characteristic of clustering is that prior knowledge of the data is not required. Extensive use of clustering algorithms is made in areas of data science and data mining, where the objective is to group the information that has common characteristics, as well as to define the optimal number of groups. Figure 1 shows the common clustering algorithms: Partitional, Hierarchical, Density-Based, Grid-Based, Spectral, Gravitational, Neural Network-Based, and Evolutionary Clustering [3]. There is another clustering technique based on semantic definition [5], but it has a disadvantage that it works for the supervised clustering only.

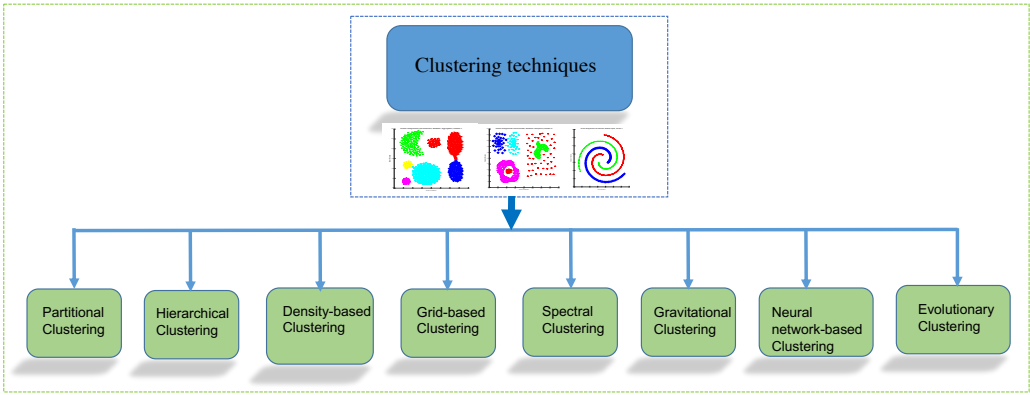


Figure 1. Taxonomy of clustering techniques.

The main works related to clustering algorithms are presented in [6–14]. Clustering algorithms are classified into *Partitional Clustering*, *Hierarchical Clustering*, *Density-based Clustering*, *Spectral Clustering*, *Gravitational Clustering*, *Evolutionary Clustering*. Their advantages and disadvantages are summarized in Table 1. Although there is a wide range of clustering algorithms available, none of the aforementioned clustering algorithm is self-sufficient for all types of clustering problems. Published clustering methodologies have been conceived for each type of database to be clustered. The reported algorithms do not take into account

the topology of the data, neither local nor global, which confines them to work with local densities or general distance criteria. This proposal defines a homogeneous-energy measure that takes into account the local and global topological properties of the data to be clustered (see Sections 3 and 4).

Table 1. Clustering algorithms with its merits and demerits.

Clustering Algorithm	Works	Advantages	Disadvantages
Partitional	[14–24]	⋆ Suitable for large datasets. ⋆ Easy and simple implementation.	⋆ Not appropriate for non-hyperspherical clusters. ⋆ Optimization problem (non-guaranteed minimum value). ⋆ Sensitive to noise and cluster initialization.
Hierarchical	[25–30]	⋆ Suitable for small databases. ⋆ Visual facility to define number of groups. ⋆ Graphic interaction.	⋆ Trouble finding the optimal number of groups automatically. ⋆ Less robust to noise.
Density-Based	[6,31–36]	⋆ Suitable for uninformed cluster shape. ⋆ Suitable for databases following a given density function. ⋆ Automatically finds the number of clusters.	⋆ Computationally heavy. ⋆ Main disadvantage is defining *a priori* density function.
Grid-Based	[37–42]	⋆ Quick handling time. ⋆ Automatic data quantity management.	⋆ User defined number of cells. ⋆ Clustering dependent on divided cells.
Spectral	[43–48]	⋆ Better representation in the frequency domain. ⋆ Preserving time-frequency correspondence.	⋆ Experience to analyze in the frequency domain. ⋆ Representation in the complex plane. ⋆ Computationally heavy.
Gravitational	[49–54]	⋆ Suitable for small databases. ⋆ Suitable for hyper-spherical separation. ⋆ Suitable for small datasets.	⋆ Optimization problem (non-guaranteed minimum value). ⋆ Computationally heavy.
Neural Network-Based	[55–60]	⋆ Suitable for large datasets. ⋆ Automatically finds the number of clusters. ⋆ Suitable for uninformed cluster shape.	⋆ Optimization problem (non-guaranteed minimum value). ⋆ Computationally heavy.

Our proposal is an attempt to overcome the main limitations of typical clustering algorithms. With the topological model based on mathematical pseudometric, unlike partitional and gravitational algorithms, we no longer have the need to establish prior knowledge of the data. Thus, it is not necessary to define any density-based function, and the use of working only with the hyper-spherical separation functions is avoided. Therefore, based on the Algebra of sets, due to not occupying excessive memory as the density-based, spectral, and gravitational algorithms do, the calculations and results are obtained faster. This research paper aims to compare the clustering and synthetic benchmark datasets.

3. Proposal

There are many clustering algorithms focused on grouping elements according to their feature distribution. As presented in the above section, there are clustering algorithms. i.e., *Partitional*, *Hierarchical*, *Density-Based*, and so on, that solve a particular clustering problem. Our proposal takes into account the topology of both local and global dataset, which makes it more general and be able to cluster any type of data. The measure of similarity between the elements to cluster is defined via a *pseudometric*, and the clustering criteria is based on *an affinity function* and a homogeneous-energy state. *An affinity function* measures how close one element is to another. While a *homogeneous-energy function* determines, for a closed system, the equilibrium locally, as well as globally. Local and global homogeneous-energies are reduced to a minimum value in equilibrium. Therefore, if we associate the elements with a homogeneous-energy function, which is obtained in principle with the affinity

function, a group of elements will be formed by common elements if they are kept below their homogeneous-energy level.

At the beginning, r-representatives are selected (either randomly or sequentially), considering all the elements of the database to form a subset of elements representing each group. Subsequently, the affinity of the r-representatives to the rest of the elements in the database is calculated. An element will be assigned to a group if its group homogeneous-energy level is low or without changes. Otherwise, a new group and a new representative will be generated.

The *representatives* will be a subset of a given group. A new element can be added to a given group if it is an affinity between that new element and the group of *representatives*, keeping its group homogeneous-energy level low or without changes.

If the homogeneous-energy in a given group is drastically altered or higher than the level, then the new element is not related to this grouping, and as a consequence, the new element is rejected. A new group is created in which the elements are labeled as "others" because they do not have similar properties to each other.

With the proposal that considers homogeneous-energy, which takes into account the local and global topologies, all the disadvantages indicated in Table 1 presented in Section 2 are overcome.

4. Theoretical Fundamentals of a Topological-Pseudometric-Based Clustering

In the context of mathematical topology, six definitions, a proposition, a theorem, and an example are provided. Definition 1 refers to pseodometry. Definition 2 deals with set theory, i.e., empty sets and subsets. Definition 3 deals with the definition of pseudometrics as applied to both representative and database elements. Definition 4 refers to the energy function. Definition 5 deals with pseudometrics around a topological neighborhood. Finally, definition 6 measures the pseudometrics to the representatives and to the subsets of groups. The example shows the energy variability having the same representative element, but with different topologically distributed neighbors.

Definition 1. *Let X be a non-empty set. The Cartesian product $X \times Y$ of sets X and Y is the set of all ordered pairs (x, y) with $x \in X$ and $y \in Y$.*

A pseudometric space (X, d) is a set X together with a non-negative real-valued function $d : X \times X \to [0, \infty)$ (called a pseudometric function) such that, for every x, y, $z \in X$.

$$\begin{aligned} d(x,y) &= d(y,x) \\ d(x,y) &\leq d(x,z) + d(z,y) \\ d(x,x) &= 0. \end{aligned} \tag{1}$$

Unlike a metric space, points in a pseudometric space need not to be distinguishable; that is, one may have $d(x, y) = 0$ for distinct values $x \neq y$.

The pseudometric topology is induced by the open balls, with $x \in X$ and $r \geq 0$,

$$B_r(x) = \{x \in X : d(x,y) < r\}, \tag{2}$$

which forms basis for the topology.

A topological space is said to be a pseudo metrizable topological space if the space can be given a pseudometric such that the pseudometric topology coincides with the given topology on the space [61].

Definition 2. *An exact cover $\mathbb{C}(X)$ of a set X is a family $\mathbb{C}(X) = \{C_i | i \in I\}$ of nonempty subsets of X such that the following conditions are satisfied:*

- *For each $i \in I$, $C_i \subset X$.*
- *For $i, j \in I$ and $i \neq j$ implies $C_i \cap C_j = \emptyset$.*
- $\bigcup_{i \in I} C_i = X$.

Definition 3. *Let (X, d) be a pseudometric space and A be a subset of X. If an element $x^* \in A$ satisfies the condition:*

$$\sum_{y \in A} d(x^*, y) \leq \sum_{y \in A} d(z, y), \tag{3}$$

for every $z \in A$ then x^* is called representative of A. The set of representatives of A is denoted by $\mathbb{R}(A)$.

The distance between point $x \in X$ and a set A can be defined as:

$$d(A, x) = \min\{d(a, x) \, : \, x \in X, a \in \mathbb{R}(A)\}. \tag{4}$$

Lemma 1. *Let (X, d) be a pseudometric space. For each finite set $A \in 2^X$, the set of representatives $\mathbb{R}(A)$ is a non-empty set.*

Proof. For the set $A = \{a_i : i = 1, \ldots, n\}$, considers,

$$D_i = \frac{1}{|A| - 1} \sum_{y \in A} d(a_i, y), \tag{5}$$

take $D^* = min\{D(a_i) : i = 1, \ldots, n\}$. There is an i^* such that $D^* = D_i$. For definition of a_{i^*} it follows,

$$\sum_{y \in A} d(a_{i^*}, y) \leq \sum_{y \in A} d(x, y). \tag{6}$$

For each $x \in A$. This proves that a_{i^*} is a representative of A. □

Definition 4. *The energy function $\mathbb{E} : 2^X \to [0, \infty)$ can be defined as follows:*

$$\mathbb{E}(A) = \frac{1}{|A| - 1} \sum_{x \in A} d(x^*, x), \tag{7}$$

where $x^* \in \mathbb{E}(A)$.

It should be noted that the energy $\mathbb{E}(A)$ of a set A is independent of the choice of the representatives of A. By definition, if x_1 and x_2 are representative of A, then, the following condition must be satisfied:

$$\frac{1}{|A| - 1} \sum_{y \in A} d(x_1, y) = \frac{1}{|A| - 1} \sum_{y \in A} d(x_2, y). \tag{8}$$

A point p preserves the energy of A if $\mathbb{E}(A \cup \{p\}) \leq \mathbb{E}(A)$.

Example 1. *Energy variations can be represented by subsets elements, having the same representative.*

Let the sets be X, Y, and Z in R^2 with the Euclidean metric (see Figure 2). Elements are marked with dots and its representative with a star. Elements are distributed in two concentric circles with radii 1, and 2. Consider the subsets Y, and Z as removing a circle from X (see Figure 2 (Y) and (Z)). In addition, it can be noted that the energy of set X, Y, and Z are 1.5, 2, and 1, respectively. It can be remarked that energy changes under subsets. Thus, for this example $Y \subseteq X$ does not imply $\mathbb{E}(Y) \leq \mathbb{E}(X)$.

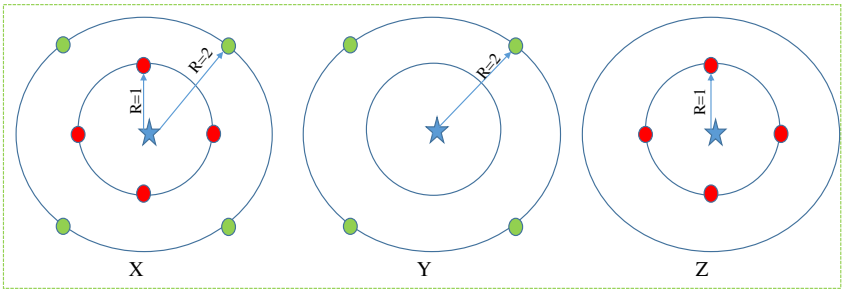


Figure 2. Subsets example of energy: X dots points distributed in two concentric circles with radii 1, and 2, and star is its *representative*. Y Subset subtracting internal circle, and Z Subset subtracting external circle.

Definition 5. *Let (X, d) be a pseudometric space and $\delta \geq 0$. An exact cover $\mathbb{C}(X) = \{C_i : i \in I\}$ of X is called a δ-exact cover if $\mathbb{E}(C_i) \leq \delta$ for each $i \in I$.*

Definition 6. *Let (X, d) be a pseudometric space, A be a subset of X, $\mathbb{R}(A)$ be the set of representatives of A, and $\lambda = \mathbb{E}(A)$. The set (see Equation (9)) given below is called a star cover of A.*

$$st(A, \lambda) = \bigcup_{x^* \in R(A)} B_\lambda(x^*). \tag{9}$$

Proposition 1. *Let (X, d) be a pseudometric space, $A \in 2^X$, and $\mathbb{R}(A)$, $\mathbb{E}(A)$ be the set of representatives, and energy of A, respectively. The star cover $st(A, \lambda)$ with $\lambda = \mathbb{E}(A)$ satisfies $\mathbb{E}(st(A, \lambda)) \leq \mathbb{E}(A)$.*

Proof. Consider $x^* \in \mathbb{E}(A)$ and $p \in B_{\mathbb{E}(A)}(x^*) \setminus A$, then $d(x^*, p) \leq \mathbb{E}(A)$. Let $A_1 = A \cup \{p\}$.

$$\begin{aligned}
\mathbb{E}(A_1) &= \frac{1}{|A_1| - 1} \sum_{y \in A_1} d(x^*, y) \\
&= \frac{1}{|A|} \sum_{y \in A} d(x^*, y) + \frac{1}{|A|} \sum_{p \in A} d(x^*, p) \\
&\leq \frac{1}{|A|} \sum_{y \in A} d(x^*, y) + \frac{1}{|A|} \mathbb{E}(A) \\
&= \frac{|A| - 1}{|A|(|A| - 1)} \sum_{y \in A} d(x^*, y) + \frac{1}{|A|} \mathbb{E}(A) \\
&\leq \frac{|A| - 1}{|A|} \mathbb{E}(A) + \frac{1}{|A|} \mathbb{E}(A) = \mathbb{E}(A).
\end{aligned} \tag{10}$$

It can construct the succession $\{A_j : j = 1, \ldots, n\}$ which satisfied $\mathbb{E}(A_m) \leq \mathbb{E}(A_n)$ for $n \leq m$ and that $A_j \to st(A, \lambda)$. Therefore $\mathbb{E}(st(A, \lambda)) \leq \mathbb{E}(A)$. □

Theorem 1. *Let (X, d) be a pseudometric space, and $\delta \geq 0$. There exists an exact cover $\mathbb{C}(X) = D \cup \{C_i : i \in I\}$ of X where $\{C_i : i \in I\}$ is δ exact cover of $X \setminus D$, and the element D does not preserve the energy of C_j for each $j \in I$.*

Proof. It will construct a family of sets $\{F_i\}_{i=1}^k = \{D_k \cup \{C_1, C_2, \ldots, C_k\}\}_{i=1}^k$ which satisfies the following properties for each $i, j = 1, \ldots, n$.

$$\begin{aligned}
&(i) \quad C_i \cap C_j = 0 \ \forall i \neq j \\
&(ii) \quad D_k \cap C_j = 0 \ \forall j = 1, \ldots, k \\
&(iii) \quad \mathbb{E}(C_i) \leq \delta.
\end{aligned} \tag{11}$$

Stand $x_1, x_2 \in X$. On the condition $d(x_1, x_2) < \delta$ then $C_1 = \{x_1, x_2\}$. If not, define $D_1 = \{x_1, x_2\}$. On the other hand, taking n and m to be two integers, and without loss

of generality, next condition $d(x_m, x_n) < \delta$ is fulfilled for $m \leq n$. Let $C^1 = \{x_m, x_n\}$ and $D_1 = \{x_1, x_2, \ldots, x_{m-1}, x_{m+1,\ldots,x_{n-1}}\}$, Subsequently, take $C_1 = st(C^1, \delta)$ and considering $F_1 = D_1 \cap \{C_1\}$. Taking into account the proposition 1 then inequalities stand:

$$\mathbb{E}(C_1) = \mathbb{E}(st(C^1, \delta)) \leq \mathbb{E}(C^1) < \delta, \tag{12}$$

consequently, the condition $C_1 = st(C^1, \delta)$ is satisfied. It has been shown that properties *(i)*, and *(ii)* are completely satisfied. □

5. Methodology

Proposed clustering methodology is shown in Figure 3. The main steps of the methodology are described below.

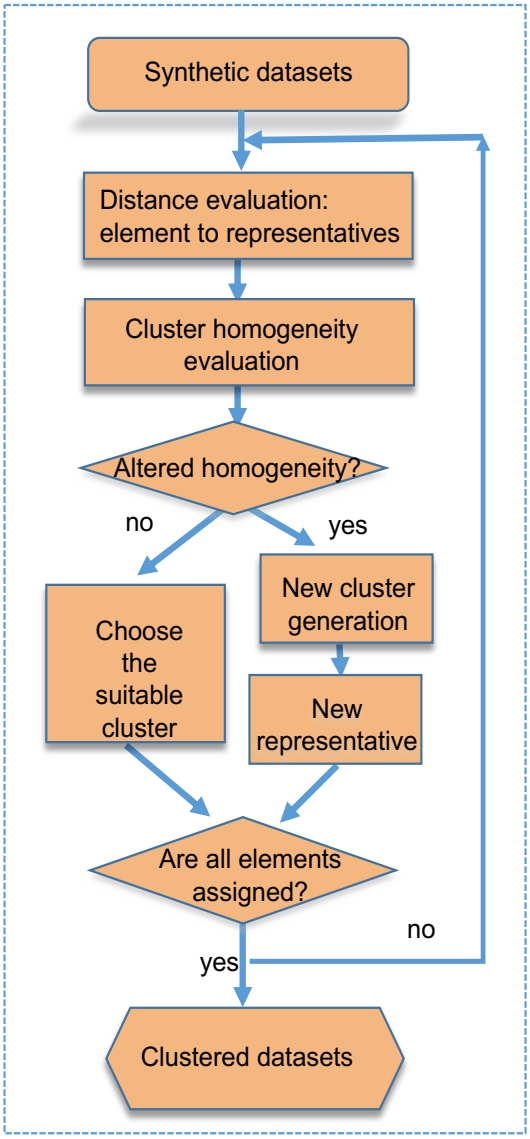


Figure 3. General scheme of proposed methodology.

1. Dataset are read: First, datasets are read as ascii files, where the first column represents the information on the x-axis, while the second column represents the information on the y-axis.
2. The manhattan distance is calculated on the totality of the data: A measure amount of all elements is defined by its manhattan distance, in this case, there is a two-dimension manhattan distance.

3. Local and general pseudometry is evaluated: Based on pseudometry defined in Definitions 1–3, local and general topology are taken into account in order to measure the cluster energy, defined in Definition 4.
4. The appropriate cluster is chosen: If the energy of the cluster is not affected by the new element, then it is integrated to the cluster, otherwise, a new one is generated.
5. The homogeneous-energy of the clusters is evaluated: At this stage, each cluster energy is calculated in order to update the cluster energy.

If all the elements are assigned to a cluster, then the algorithm ends.

5.1. Datasets

In order to test the proposed algorithm and its robustness in the automatic generation of clusters, synthetic datasets of two-dimensional points were used. Datasets were taken from Machine Learning group of School of Computing, University of Eastern Finland [4]. The dataset is shown in Figure 4 and its characteristics are given in Table 2. There were 8 datasets tested throughout the topological algorithm, that are: *Aggregation*, *Compound*, *Pathbase*, *Spiral*, *D31*, *R15*, *Jain*, and *Flame*. The Shape dataset was chosen because of its levels of complexity, spherically separable point distribution (R15, D31), embedded classes (Jain, flame, and compound), and complex distributions (spiral and pathbased). The previous data distributions are difficult to be correctly clustered by a single clustering algorithm. In our topological-pseudometric-based clustering proposal, it is possible.

Table 2. Synthetic datasets for clustering task.

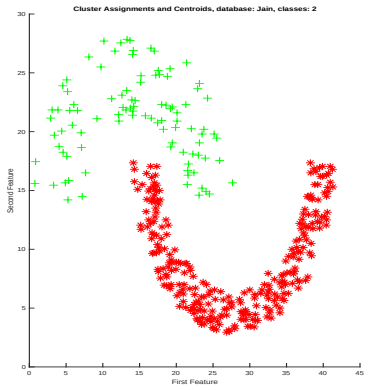
Dataset	Classes	Samples
Aggregation	K = 7	N = 788
Compound	K = 6	N = 399
Pathbase	K = 3	N = 300
Spiral	K = 3	N = 312
D31	K = 31	N = 3100
R15	K = 15	N = 600
Jain	K = 2	N = 373
Flame	K = 2	N = 240

5.2. Metric and Distance

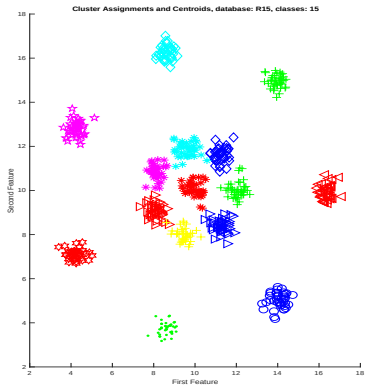
Consider the set Γ of points (P). Points correspond to a synthetic dataset (see Figure 4). According to Section 4, a pseudometric ρ is defined for two given points as follows:

$$\begin{aligned} \rho : \Gamma^2 &\to R, \\ \rho(P_1, P_2) &= d(P_1, P_2). \end{aligned} \tag{13}$$

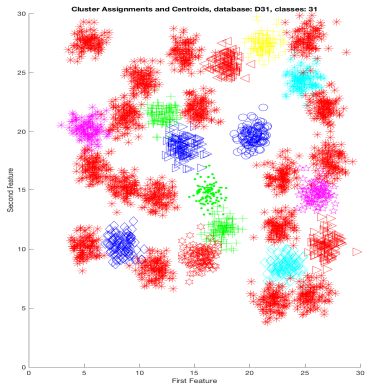
A pseudometric has been built in the space of the points. Thus, Theorem 1 will always allow to create clusters. The Manhattan distance is used in all the tests.



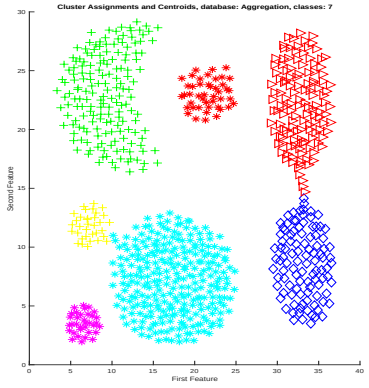
(**a**) Jain dataset.



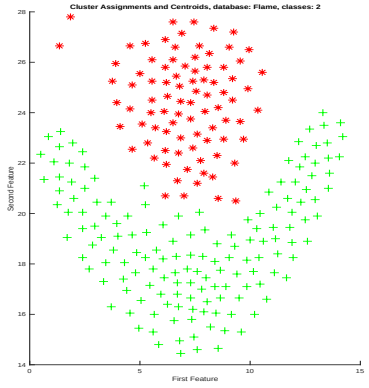
(**b**) R15 dataset.



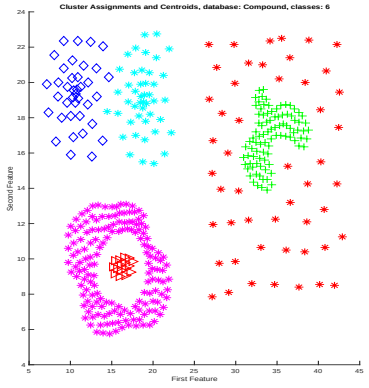
(**c**) D31 dataset.



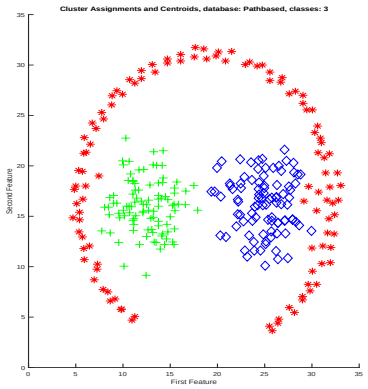
(**d**) Aggregation dataset.



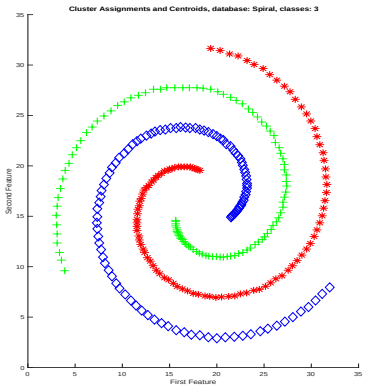
(**e**) Flame dataset.



(**f**) Compound dataset.



(**g**) Pathbased dataset.



(**h**) Spiral dataset.

Figure 4. Original 2-Dimensional *Shape* datasets.

5.3. Add the Object into the Suitable Cluster

Starting from the *representatives* in each cluster, the new homogeneous-energy measure is calculated for each new element to be added to the cluster. The homogeneities are calculated using *theorem 1* and *proposition 1* (see Section 4) for each *representatives*. Thus, the new element is assigned to the cluster which is not affected in its homogeneous-energy. Otherwise, a new cluster is created and the new element is taken as its *representatives*.

6. Experimental Results

Experiments on the D-Dimension were conducted on *shape* dataset in order to test the efficiency of the clusterings produced by our topological algorithm on a varied collection of synthetic datasets (see Table 2). The goal in this set of experiments is to show how topological clustering can be used to improve the quality and robustness of widely-used clustering datasets benchmark. Eight synthetic datasets were used, from separate distributions and compacted classes to spiral or circular distributions, which are extremely complex to cluster: *Aggregation*, *Compound*, *Pathbase*, *Spiral*, *D31*, *R15*, *Jain*, and *Flame*. The eight datasets were divided into three groups according to their distribution-shape and difficulty of grouping. Clustering error is defined as [62]:

$$error = \frac{incorrect\ clustering}{Total\ of\ cluster\ elements} \times 100 \tag{14}$$

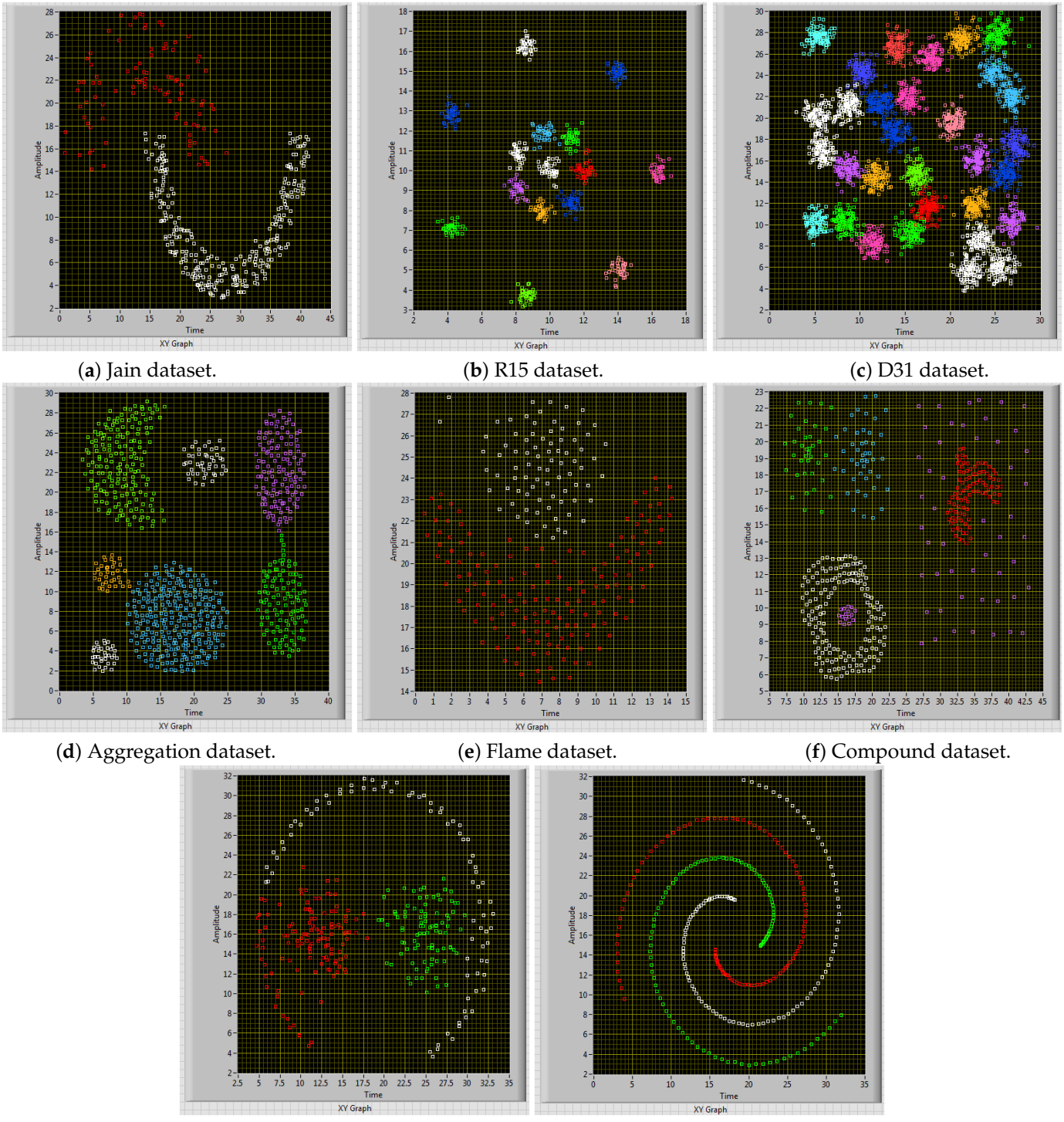
- **Easy:** The proposed algorithm works very well for the *Jain*, *D31*, *R15*, and *Aggregation* datasets (see Figure 5a–d). The clustering error is less than 2%. The clustering problem is on *Aggregation* dataset (in the union of the classes of the right side; see Figure 5d).
- **Medium:** Results are good for *Flame* and *Compound* datasets (see Figure 5e,f). The clustering error is less than 5%.
- **Complex:** For the third group, the results of the *Pathbased* and spiral datasets are good (see Figure 5g,h). The clustering error is less than 10% just for *Pathbase* dataset. The result for *Spiral* dataset is 0%.

The proposed algorithm was implemented in LabVIEW software, which is oriented to work easily with hardware, obtaining accurate and fast measurements and results. LabView allows to run algorithms from reading datasets from files, to the graphical display of information.

The algorithm begins taking *representatives* randomly from the whole dataset. The distances among *representatives* allow to define the parameter δ as intermediate values. The δ value is set at 0.2% of the distance between two representatives. After setting the initial parameters, the pseudometric grouping theorem is applied. Thus, there are $k - groups$ as a result. We impose the criteria that groups containing less than 10 percent of the entire dataset are annulled, and the elements are re-integrated into the Q set. For the next iterations, the delta value is increased by 0.2%. The algorithm ends when the Q set is empty.

Considering local and global topologies has allowed us to define a more robust algorithm than those reported in the literature. Better results have been obtained in this research work for different characteristics of the databases to be clustered.

The proposal overcomes the problems reported in existing algorithms and allows not having the limitations of (1) *a prior* knowledge of the data and (2) using only spherical separator functions. Another important advantage of this proposed algorithm is the use of Algebra of sets which helps in obtaining the results faster and without excessive memory consumption (as *density-based*, *spectral*, and *gravitational* algorithms do).



(**a**) Jain dataset. (**b**) R15 dataset. (**c**) D31 dataset.

(**d**) Aggregation dataset. (**e**) Flame dataset. (**f**) Compound dataset.

(**g**) Pathbased dataset. (**h**) Spiral dataset.

Figure 5. Clustered 2-Dimensional Shape datasets.

7. Conclusions

The new 2D clustering algorithm based on a mathematical topological theory was presented in this research paper. The proposed theory of a pseudometric-based clustering model and its application in synthetic datasets worked as expected. Thus, this new method based on topology theory has successfully worked for the clustering of easy and complex datasets. The proposal also takes into account the local and global topological properties of the data to be clustered in a definition of homogeneous-energy measurement.

The proposal overcomes the problems reported in existing algorithms and without the need for (1) *a prior* knowledge of the data and (2) using only spherical separator functions. Another advantage of the proposal is that since the proposed algorithm is based on Algebra of sets, the computational results are faster and without excessive memory consumption.

Because of the theoretical development, there is now a theorem (Theorem 1) that can be applied in any space that defines a pseudometric.

Based on the results obtained, the clustering of n-dimensional databases will be explored, as well as the application of the proposal to large databases.

Author Contributions: Conceptualization, I.O.-G. and Y.P.-P.; methodology, I.O.-G. and C.A.-C.; software, I.O.-G. and Y.P.-P.; validation, I.O.-G. and C.A.-C.; formal analysis, I.O.-G. and C.A.-C.; investigation, I.O.-G., Y.P.-P. and C.A.-C.; writing—original draft preparation, C.A.-C.; writing—review and editing, C.A.-C.; supervision, I.O.-G. and C.A.-C.; project administration, C.A.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O.; Tiwari, A.; Er, M.; Ding, W.; Lin, C.T. A review of clustering techniques and developments. *Neurocomputing* **2017**, *267*, 664–681. [CrossRef]
2. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
3. Zhao, Y.; Tarus, S.; Yang, L.; Sun, J.; Ge, Y.; Wang, J. Privacy-preserving clustering for big data in cyber-physical-social systems: Survey and perspectives. *Inf. Sci.* **2020**, *515*, 132–155. [CrossRef]
4. Fränti, P.; Sieranoja, S. K-means properties on six clustering benchmark datasets. *Appl. Intell.* **2018**, *48*, 4743–4759. [CrossRef]
5. Wan, S.P.; Yan, J.; Dong, J.Y. Personalized individual semantics based consensus reaching process for large-scale group decision making with probabilistic linguistic preference relations and application to COVID-19 surveillance. *Expert Syst. Appl.* **2022**, *191*, 116328. [CrossRef]
6. Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data Sci.* **2015**, *2*, 165–193. [CrossRef]
7. Khandare, A.; Alvi, A.S. Clustering Algorithms: Experiment and Improvements. In *Computing and Network Sustainability*; Vishwakarma, H., Akashe, S., Eds.; Springer: Singapore, 2017; pp. 263–271.
8. Patibandla, R.S.M.L.; Veeranjaneyulu, N. Survey on Clustering Algorithms for Unstructured Data. In *Intelligent Engineering Informatics*; Bhateja, V., Coello Coello, C.A., Satapathy, S.C., Pattnaik, P.K., Eds.; Springer: Singapore, 2018; pp. 421–429.
9. Osman, M.M.A.; Syed-Yusof, S.K.; Abd Malik, N.N.N.; Zubair, S. A survey of clustering algorithms for cognitive radio ad hoc networks. *Wirel. Netw.* **2018**, *24*, 1451–1475. [CrossRef]
10. Bindra, K.; Mishra, A.; Suryakant. Effective Data Clustering Algorithms. In *Soft Computing: Theories and Applications*; Ray, K., Sharma, T.K., Rawat, S., Saini, R.K., Bandyopadhyay, A., Eds.; Springer: Singapore, 2019; pp. 419–432.
11. Djouzi, K.; Beghdad-Bey, K. A Review of Clustering Algorithms for Big Data. In Proceedings of the 2019 International Conference on Networking and Advanced Systems (ICNAS), Annaba, Algeria, 26–27 June 2019; pp. 1–6. [CrossRef]
12. Ahmad, A.; Khan, S.S. Survey of State-of-the-Art Mixed Data Clustering Algorithms. *IEEE Access* **2019**, *7*, 31883–31902. [CrossRef]
13. Zhang, J.; Liang, Q.; Wang, H. Uniformities on strongly topological gyrogroups. *Topol. Its Appl.* **2021**, *302*, 107776. [CrossRef]
14. Telikani, A.; Tahmassebi, A.; Banzhaf, W.; Gandomi, A. Evolutionary Machine Learning: A Survey. *ACM Comput. Surv.* **2022**, *54*, 161. [CrossRef]
15. Jinyin, C.; Xiang, L.; Haibing, Z.; Xintong, B. A novel cluster center fast determination clustering algorithm. *Appl. Soft Comput.* **2017**, *57*, 539–555. [CrossRef]
16. Schubert, E.; Rousseeuw, P. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2019; Volume 11807.
17. Liu, X.; Zhu, X.; Li, M.; Wang, L.; Zhu, E.; Liu, T.; Kloft, M.; Shen, D.; Yin, J.; Gao, W. Multiple Kernel k-means with Incomplete Kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1191–1204. [CrossRef] [PubMed]
18. Kanika; Rani, K.; Sangeeta; Preeti. Visual Analytics for Comparing the Impact of Outliers in k-Means and k-Medoids Algorithm. In Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 4–6 February 2019; pp. 93–97. [CrossRef]
19. Gupta, T.; Panda, S.P. A Comparison of K-Means Clustering Algorithm and CLARA Clustering Algorithm on Iris Dataset. *Int. J. Eng. Technol.* **2019**, *7*, 4766–4768.

20. Li, Y.; Cai, J.; Yang, H.; Zhang, J.; Zhao, X. A Novel Algorithm for Initial Cluster Center Selection. *IEEE Access* **2019**, *7*, 74683–74693. [CrossRef]
21. Zhang, Y.; Bai, X.; Fan, R.; Wang, Z. Deviation-Sparse Fuzzy C-Means With Neighbor Information Constraint. *IEEE Trans. Fuzzy Syst.* **2019**, *27*, 185–199. [CrossRef]
22. Tang, Y.; Ren, F.; Pedrycz, W. Fuzzy C-Means clustering through SSIM and patch for image segmentation. *Appl. Soft Comput.* **2020**, *87*, 105928. [CrossRef]
23. Garcia, M.; Igusa, K. Continuously triangulating the continuous cluster category. *Topol. Appl.* **2020**, *285*, 107411. [CrossRef]
24. Osuna-Galán, I.; Pérez-Pimentel, Y.; Avilés-Cruz, C.; Villegas-Cortez, J. Topology: A Theory of a Pseudometric-Based Clustering Model and Its Application in Content-Based Image Retrieval. *Math. Probl. Eng.* **2019**, *2019*, 4540731. [CrossRef]
25. Lim, J.; Jun, J.; Kim, S.H.; McLeod, D. A Framework for Clustering Mixed Attribute Type Datasets. In Proceedings of the 4th International Conference on Emerging Databases-Technologies, Applications, and Theory (EDB 2012), Seoul, Korea, 23–25 August 2012.
26. Nazari, Z.; Kang, D.; Asharif, M.; Sung, Y.; Ogawa, S. A new hierarchical clustering algorithm. In Proceedings of the 2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 28–30 November 2015; pp. 148–152.
27. Rashedi, E.; Mirzaei, A.; Rahmati, M. Optimized aggregation function in hierarchical clustering combination. *Intell. Data Anal.* **2016**, *20*, 281–291. [CrossRef]
28. Yao, W.; Dumitru, C.; Loffeld, O.; Datcu, M. Semi-supervised Hierarchical Clustering for Semantic SAR Image Annotation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 1993–2008. [CrossRef]
29. Pitolli, G.; Aniello, L.; Laurenza, G.; Querzoni, L.; Baldoni, R. Malware family identification with BIRCH clustering. In Proceedings of the 2017 International Carnahan Conference on Security Technology (ICCST), Madrid, Spain, 23–26 October 2017; pp. 1–6. [CrossRef]
30. Cao, X.; Su, T.; Wang, P.; Wang, G.; Lv, Z.; Li, X. An Optimized Chameleon Algorithm Based on Local Features. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing (ICMLC 2018), Macau, China, 26–28 February 2018; ACM: New York, NY, USA, 2018; pp. 184–192. [CrossRef]
31. Yokoyama, S.; Bogardi-Meszoly, A.; Ishikawa, H. EBSCAN: An entanglement-based algorithm for discovering dense regions in large geo-social data streams with noise. In Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Bellevue, WA, USA, 3–6 November 2015.
32. Rehioui, H.; Idrissi, A.; Abourezq, M.; Zegrari, F. DENCLUE-IM: A New Approach for Big Data Clustering. *Procedia Comput. Sci.* **2016**, *83*, 560–567. [CrossRef]
33. Kumar, K.M.; Reddy, A.R.M. A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. *Pattern Recognit.* **2016**, *58*, 39–48. [CrossRef]
34. Behzadi, S.; Ibrahim, M.A.; Plant, C. Parameter Free Mixed-Type Density-Based Clustering. In *Database and Expert Systems Applications*; Hartmann, S., Ma, H., Hameurlain, A., Pernul, G., Wagner, R.R., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 19–34.
35. Matioli, L.C.; Santos, S.; Kleina, M.; Leite, E.A. A new algorithm for clustering based on kernel density estimation. *J. Appl. Stat.* **2018**, *45*, 347–366. [CrossRef]
36. Shu, Z.; Yang, S.; Wu, H.; Xin, S.; Pang, C.; Kavan, L.; Liu, L. 3D Shape Segmentation Using Soft Density Peak Clustering and Semi-Supervised Learning. *CAD Comput.-Aided Des.* **2022**, *145*. [CrossRef]
37. Rashad, M.A.; El-Deeb, H.; Fakhr, M.W. Document Classification Using Enhanced Grid Based Clustering Algorithm. In *New Trends in Networking, Computing, E-Learning, Systems Sciences, and Engineering*; Elleithy, K., Sobh, T., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 207–215.
38. Wagner, T.; Feger, R.; Stelzer, A. A fast grid-based clustering algorithm for range/Doppler/DoA measurements. In Proceedings of the 2016 European Radar Conference (EuRAD), London, UK, 5–7 October 2016; pp. 105–108.
39. Lalitha, K.; Thangarajan, R.; Udgata, S.K.; Poongodi, C.; Sahu, A.P. GCCR: An Efficient Grid Based Clustering and Combinational Routing in Wireless Sensor Networks. *Wirel. Pers. Commun.* **2017**, *97*, 1075–1095. [CrossRef]
40. Deng, C.; Song, J.; Sun, R.; Cai, S.; Shi, Y. Gridwave: A grid-based clustering algorithm for market transaction data based on spatial-temporal density-waves and synchronization. *Multimed. Tools Appl.* **2018**, *77*, 29623–29637. [CrossRef]
41. Chen, J.; Lin, X.; Xuan, Q.; Xiang, Y. FGCH: A fast and grid based clustering algorithm for hybrid data stream. *Appl. Intell.* **2019**, *49*, 1228–1244. [CrossRef]
42. Yang, Y.; Zhu, Z. A Fast and Efficient Grid-Based K-means++ Clustering Algorithm for Large-Scale Datasets. In *The Fifth Euro-China Conference on Intelligent Data Analysis and Applications*; Krömer, P., Zhang, H., Liang, Y., Pan, J.S., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 508–515.
43. Menendez, H.; Camacho, D. GANY: A genetic spectral-based Clustering algorithm for Large Data Analysis. In Proceedings of the 2015 IEEE Congress on Evolutionary Computation (CEC), Sendai, Japan, 25–28 May 2015; pp. 640–647.
44. Shang, R.; Zhang, Z.; Jiao, L.; Wang, W.; Yang, S. Global discriminative-based nonnegative spectral clustering. *Pattern Recognit.* **2016**, *55*, 172–182. [CrossRef]
45. Alamdari, M.; Rakotoarivelo, T.; Khoa, N. A spectral-based clustering for structural health monitoring of the Sydney Harbour Bridge. *Mech. Syst. Signal Process.* **2017**, *87*, 384–400. [CrossRef]

46. Tian, L.; Du, Q.; Kopriva, I.; Younan, N. Spatial-spectral Based Multi-view Low-rank Sparse Sbuspace Clustering for Hyperspectral Imagery. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 8488–8491.
47. Nemade, V.; Shastri, A.; Ahuja, K.; Tiwari, A. Scaled and Projected Spectral Clustering with Vector Quantization for Handling Big Data. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 2174–2179.
48. Ma, L.; Zhang, Y.; Leiva, V.; Liu, S.; Ma, T. A new clustering algorithm based on a radar scanning strategy with applications to machine learning data. *Expert Syst. Appl.* **2022**, *191*. [CrossRef]
49. Dowlatshahi, M.; Nezamabadi-Pour, H. GGSA: A Grouping Gravitational Search Algorithm for data clustering. *Eng. Appl. Artif. Intell.* **2014**, *36*, 114–121. [CrossRef]
50. Kumar, V.; Chhabra, J.; Kumar, D. Automatic cluster evolution using gravitational search algorithm and its application on image segmentation. *Eng. Appl. Artif. Intell.* **2014**, *29*, 93–103. [CrossRef]
51. Nikbakht, H.; Mirvaziri, H. A new algorithm for data clustering based on gravitational search algorithm and genetic operators. In Proceedings of the 2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP), Mashhad, Iran, 3–5 March 2015; pp. 222–227.
52. Sheshasaayee, A.; Sridevi, D. Fuzzy C-means algorithm with gravitational search algorithm in spatial data mining. In Proceedings of the 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 26–27 August 2016; Volume 1, pp. 1–5.
53. Deng, Z.; Qian, G.; Chen, Z.; Su, H. Identifying Tor Anonymous Traffic Based on Gravitational Clustering Analysis. In Proceedings of the 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 26–27 August 2017; Volume 2, pp. 79–83.
54. Alswaitti, M.; Ishak, M.; Isa, N. Optimized gravitational-based data clustering algorithm. *Eng. Appl. Artif. Intell.* **2018**, *73*, 126–148. [CrossRef]
55. Yuqing, S.; Junfei, Q.; Honggui, H. Structure design for RBF neural network based on improved K-means algorithm. In Proceedings of the 2016 Chinese Control and Decision Conference (CCDC), Yinchuan, China, 28–30 May 2016; pp. 7035–7040.
56. Amin, H.; Deabes, W.; Bouazza, K. Clustering of user activities based on adaptive threshold spiking neural networks. In Proceedings of the 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), Milan, Italy, 4–7 July 2017; pp. 1–6.
57. Abavisani, M.; Patel, V. Deep Multimodal Subspace Clustering Networks. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 1601–1614. [CrossRef]
58. Ren, Z.; Chen, J.; Ye, L.; Wang, C.; Liu, Y.; Zhou, W. Application of RBF Neural Network Optimized Based on K-Means Cluster Algorithm in Fault Diagnosis. In Proceedings of the 2018 21st International Conference on Electrical Machines and Systems (ICEMS), Jeju, Korea, 7–10 October 2018; pp. 2492–2496.
59. Kimura, M. AutoClustering: A feed-forward neural network based clustering algorithm. In Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, 17–20 November 2019; Volume 2018, pp. 659–666.
60. Cheng, Y.; Yu, S.; Liu, J.; Han, Z.; Li, Y.; Tang, Y.; Wu, C. Representation Learning Based on Autoencoder and Deep Adaptive Clustering for Image Clustering. *Math. Probl. Eng.* **2021**, *2021*, 3742536. [CrossRef]
61. Engelking, R. *General Topology*; Springer International Publishing: Cham, Switzerland, 1989.
62. Balcerzak, M.; Leonetti, P. On the relationship between ideal cluster points and ideal limit points. *Topol. Its Appl.* **2019**, *252*, 178–190. [CrossRef]