

Clustering

1 *Introduction*

- *Definition*
- *Components*
- *Goals in the design of clustering algorithms*

2 *Distance Measures*

3 *Taxonomy of clustering algorithms*

K-means

4 *Comparison*

5 *Bibliography*

TRATADOS DE LÓGICA
(ORGANON)

I

Aristóteles

BIBLIOTECA CLÁSICA GREDOS

TRATADOS DE LÓGICA
(ORGANON)

I

Aristóteles

BIBLIOTECA CLÁSICA GRECOS

Categories: all the possible things that can be the subject or the predicate of a proposition


TRATADOS DE LÓGICA
(ORGANON)
I

FRANCISCO DE SMOLLETT

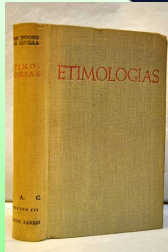
ETIMOLOGÍAS

A. C.

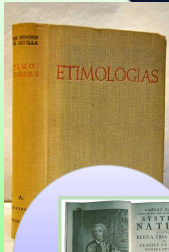
PRIMERA PARTE



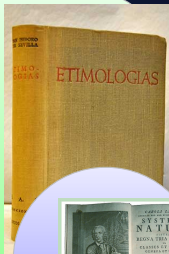
Categories: all the possible things that can be the subject or the predicate of a proposition



~~Categories: all the possible~~
summa of universal knowledge:
“The knowledge of a word’s
etymology often has an indispens-
able usefulness for interpreting
the word, for when you have
seen whence a word has origi-
nated, you understand its force
more quickly. Indeed, one’s
insight into anything is clearer
when its etymology is known.”

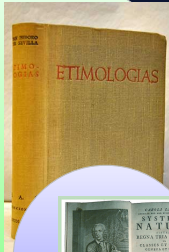


Categories: all the possible
summa of universal knowledge:
“The knowledge of a word’s
etymology often has an indispens-
able usefulness for interpreting
the word, for when you have
seen whence a word has origi-
nated, you understand its force
more quickly. Indeed, one’s
insight into anything is clearer
when its etymology is known.”



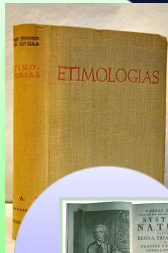
Categories: all the possible
summa of universal knowledge:
“The knowledge of a word’s
etymology often has an indispens-
able usefulness for interpreting
the word, for when you have
seen whence a word has origi-
nated, you understand its force
more quickly. Indeed, one’s

Systema naturae ⇔ binomial
nomenclature: genus + species



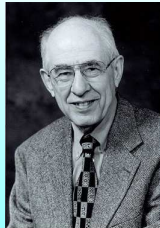
Categories: all the possible
summa of universal knowledge:
“The knowledge of a word’s
etymology often has an indispens-
able usefulness for interpreting
the word, for when you have
seen whence a word has origi-
nated, you understand its force
more quickly. Indeed, one’s

Systema naturae \Leftrightarrow binomial
nomenclature: genus + species



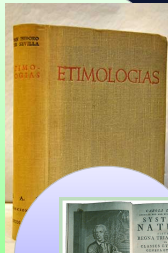
Categories: all the possible
summa of universal knowledge:
“The knowledge of a word’s
etymology often has an indispens-
able usefulness for interpreting
the word, for when you have
seen whence a word has origi-
nated, you understand its force
more quickly. Indeed, one’s

Systema naturae \Leftrightarrow binomial



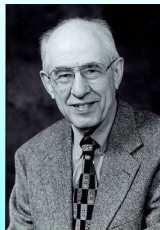
"Are thinking and referring identical with
computational states of the brain?"

Hilary Putnam



Categories: all the possible
summa of universal knowledge:
“The knowledge of a word’s
etymology often has an indispens-
able usefulness for interpreting
the word, for when you have
seen whence a word has origi-
nated, you understand its force
more quickly. Indeed, one’s

Systema naturae \Leftrightarrow binomial



"Are thinking and referring identical with
computational states of the brain?"

Hilary Putnam

SECOND EDITION

WITH A NEW SECTION: "ON ROBUSTNESS & FRAGILITY"

NEW YORK TIMES BESTSELLER

THE BLACK SWAN



The Impact of the
HIGHLY IMPROBABLE

"The most prophetic voice of all."
—GQ

Nassim Nicholas Taleb

Nassim Nicholas Taleb

Judea Pearl

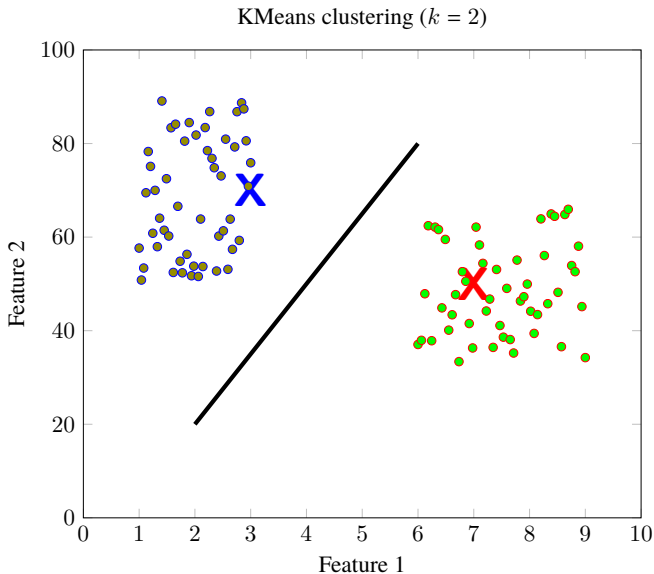
Behind any causal conclusion there must be some causal assumption, untested in observational studies^a

^aPearl 2009.

Clustering from a computational point of view

1. Proximity between items
2. What do we mean by proximity?
3. Are they really related?
4. Can we perform such a methodology in an automatic way?

Clustering vs. Classification



Classification

- ✓ There exist LABELS for some points
- ✓ Rule such that new points are assigned labels properly
- ✓ Supervised learning

Clustering

- ✓ No labels
- ✓ Points assign to clusters according to **HOW CLOSE** they are to one another
- ✓ Identify STRUCTURE in data
- ✓ Unsupervised learning

Clustering definition

- ✓ Cluster analysis \equiv organization of a “collection of patterns into clusters based on similarity” (Jain, Murty, et al. 1999)
- ✓ The definition and the scope of “cluster” in the data set are not easy to define
- ✓ According to (Jain and Dubes 1988), a cluster
 1. set of similar objects
 2. set of points aggregated in the testing environment | the distance between two points in a cluster is less than the distance between any point in the cluster and any point of other clusters
 3. clusters can be densely connected regions in a multi-dimensional space separated by loosely connected points

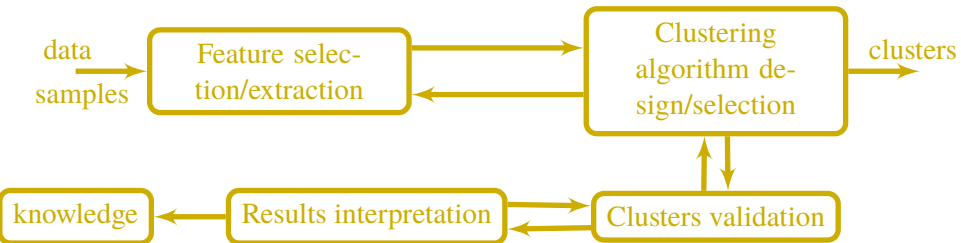
Clustering definition

- ✓ Cluster analysis \equiv organization of a “collection of patterns into clusters based on similarity” (Jain, Murty, et al. 1999)
- ✓ According to (Jain and Dubes 1988), a cluster
 1. set of similar objects \Leftarrow (minimum intra-cluster)
 2. set of points aggregated in the testing environment | the distance between two points in a cluster is less than the distance between any point in the cluster and any point of other clusters \Leftarrow (minimum inter-/intra-cluster)
 3. clusters can be densely connected regions in a multi-dimensional space separated by loosely connected points \Leftarrow (graph-/dense-based clustering)

Clustering \neq unsupervised predictive learning

- ✓ Unsupervised predictive learning
 - ✗ Vector quantization (Gersho and Gray 2012), probability density estimation, expectation maximization (Bishop 2006)
 - ✗ Accurate characterization of unobserved samples generated from the same probability distribution
- ✓ Clustering analysis
 - ✗ Unsupervised “non-predictive” learning
 - ✗ Split data sets into subsets (based on specific distance/similarity measures)
 - ✗ Not based on the “trained characterization”

Components of the clustering methodology



Goals in the design of clustering algorithms

- ¹ **Scalability** Temporal and spatial complexity should be bounded even in large datasets
- Robustness** Data outliers should be accurately detected
- Order independence** Different input data should not lead to different final results
- Minimum user-defined parameters** Configurability should be at least as possible: reduce configuration burden
- Mixed data type** Data representation should comprise numeric and/or binary and/or categorical codes
- Variability/flexibility in clusters shape** Clusters shape should be fixed
- Point proportion admissibility** Duplicating dataset + re-clustering \Rightarrow changes in the final results

¹Andreopoulos et al. 2009.

Distance measures

Definition (dx_i, x_j)

The distance between two instances x_i and x_j , which is a metric distance measure if it satisfies the following properties:

✓ *Triangle inequality*

$$dx_i, x_j \leq dx_i, x_k + dx_k, x_j, \forall x_i, x_j, x_k \in \mathcal{S}$$

✓ $dx_i, x_j = 0 \rightarrow x_i = x_j \forall x_i, x_j \in \mathcal{S}$

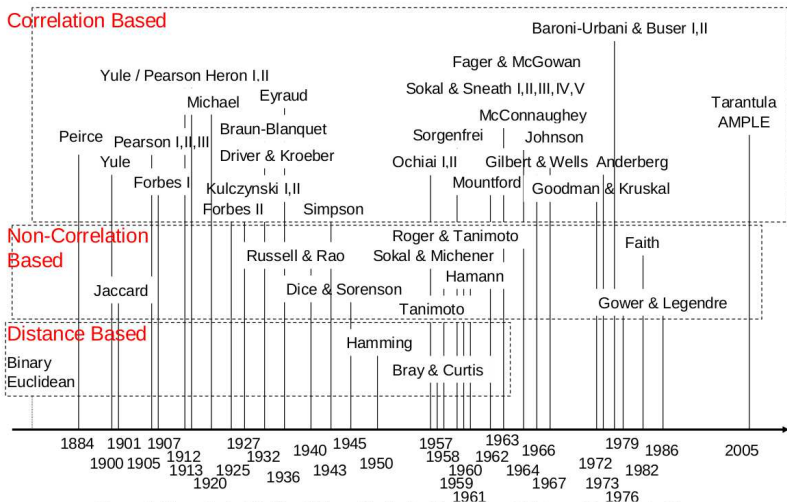


Figure 1 Chronological Table of Binary Similarity Measures and Distance Measures by Year

Mathematical definition of items proximity

Similarity measures

$$\zeta x_i, x_i = 0$$
$$\zeta x_i, x_j \neq 0, \forall i \neq j$$

Minkowski distance

⇒ Set of distance measures for numeric attributes

Definition (Minkowski metric)

$$d_{x_i, x_j} = |x_{i,1} - x_{j,1}|^g + |x_{i,2} - x_{j,2}|^g + \cdots + |x_{i,p} - x_{j,p}|^g$$
$$x_{i,k} \in a, b \subset \mathcal{R}$$

- ✓ $g = 2 \Rightarrow$ *Euclidean distance*
- ✓ $g = 1 \Rightarrow$ *Manhattan distance*
- ✓ $g = \infty \Rightarrow$ *The greatest of the paraxial distances, i.e., the Chebychev metric*

1 *Introduction*

2 *Distance Measures*

- *Minkowski distance*
- *Distances for binary attributes*

3 *Taxonomy of clustering algorithms*

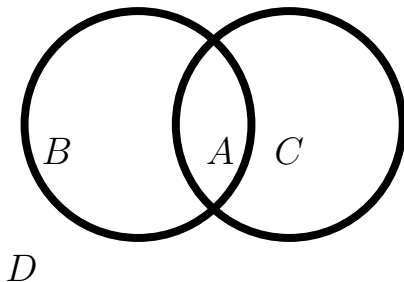
4 *Comparison*

5 *Bibliography*

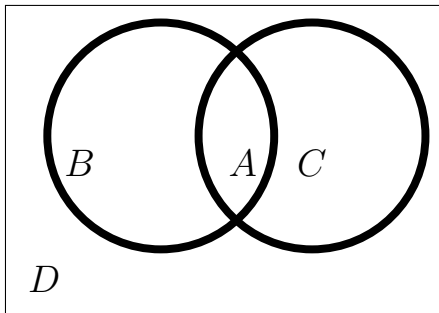
Distances for binary attributes

B stands for the attributes only contained in the original set

C stands for the attributes only contained in the test set



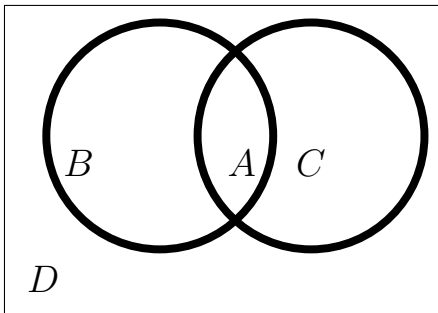
Distances for binary attributes



Definition (Jaccard coefficient)

$$\frac{A}{A + B + C}$$

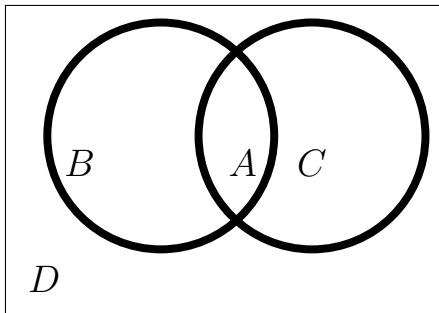
Distances for binary attributes



Definition (Russell and Rao coefficient)

$$\frac{A}{A + B + C + D}$$

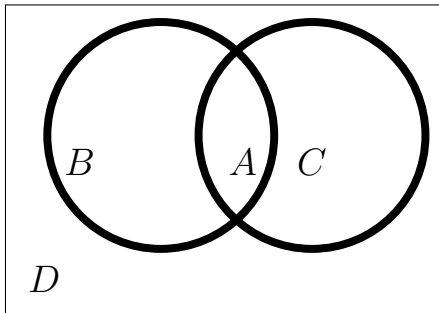
Distances for binary attributes



Definition (Dice coefficient)

$$\frac{2A}{A + B + A + C}$$


Distances for binary attributes



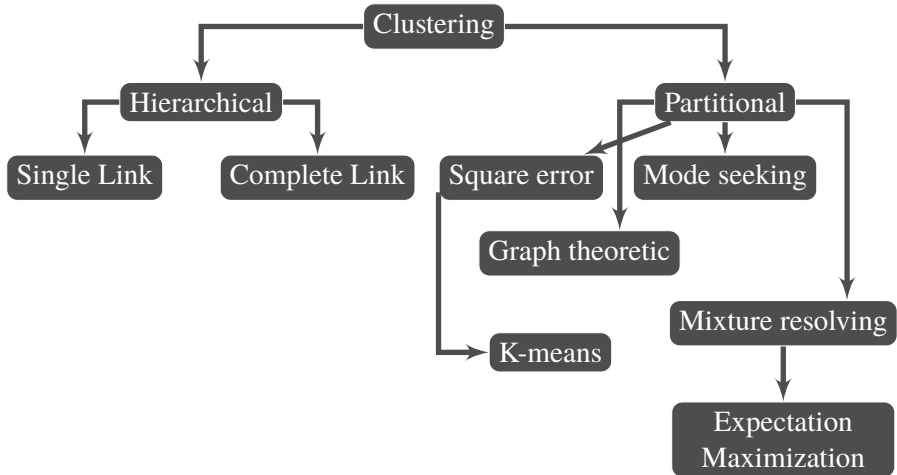
Definition (Roger and Tanimoto coefficient)

$$\frac{2A}{A + B + A + C}$$

Music dataset for the example

- ✓  prueba.tsv
- ✓ ejemplo_metricas_distancias_final.ipynb

Clustering of clustering: a taxonomy



The bottom-line of the previous taxonomy

Agglomerative vs. Divisive Algorithmic structure and operation \Rightarrow
agglomerative approach: bottom-up construction;
divisive approach: top-down (\Rightarrow Hierarchical clustering)

Monothetic vs. Polythetic Use sequential or simultaneous of features in the process (most algorithms are polythetic)

Hard vs. Fuzzy In fuzzy schemes a point can pertain to several clusters

Deterministic vs. Stochastic Deterministic objective function or random search technique

Incremental vs. Non-incremental If the size of the data is incremental or not

1 *Introduction*

2 *Distance Measures*

3 *Taxonomy of clustering algorithms*

- *Partioning clustering algorithms*
- *Hierarchical methods*

4 *Comparison*

5 *Bibliography*

Partitioning clustering algorithms

Fixed number of clusters

Numerical methods

K-means

Farthest First Traversal (FFT) k-center

K-medoids

Partion Around Medoids (PAM)

CLARA (Clustering Large Applications)

CLARANS (Clustering Large Applications Based Upon Randomized Search)

Fuzzy K-means

Discrete methods

K-modes

Fuzzy K-modes

non-negative matrix factorization (NMF) method (clustering of microarray data)

N patterns

$x_i, i \in \{0, 1, \dots, M - 1\}$

$\{x_i\} \subset \mathcal{S}$

$\mathcal{S} = \mathcal{S}_1 \mathcal{S}_2 \mathcal{S}_3 \cdots \mathcal{S}_k$

$\mathcal{S}_i \cap \mathcal{S}_j = \emptyset, \forall i \neq j$

$x_i \in \mathcal{S}_j$

K-means

Input: S set of all possible objects (with N attributes, and $\#S = M$)

Input: K number of clusters (user-defined parameter)

Output: K clusters

```
1 while Termination condition is not satisfied do
2   |   Assign instances to the closest cluster center
3   |   Update cluster centers according to the previous operation
4 end
```

K-means

Input: S set of all possible objects (with N attributes, and $\#S = M$)

Input: K number of clusters (user-defined parameter)

Output: K clusters

```
1 while Termination condition is not satisfied do
2   |   Assign instances to the closest cluster center
3   |   Update cluster centers according to the previous operation
4 end
```

- ✓ Proof of the finite convergence (Selim and Ismail 1984)
- ✓ Complexity per iteration $\sim \mathcal{O}(K \cdot M \cdot N)$

K-means

- ☺ Low complexity, ease of interpretation and implementation, adaptability to sparse data
- ☹ ↑↑ sensitivity to the initial partition
- ✓ Works well only on data sets having isotropic clusters: it is not as flexible as single link algorithms
- ✓ ↑↑ sensitivity to noisy data and outliers
- ✓ It can be only computed if the mean is properly defined (i.e., for numeric attributes)
- ✓ It requires in advance the number of clusters: no trivial when no prior knowledge is available

K-means examples

1. `ejemplo_kmeans_fundamentos.zip`
2. `ejemplo_kmeans_fundamentos.ipynb`

K-prototypes

- ✓ Based on K-means but it can applied to categorical attributes
- ✓ Similarity measure is based on the number of mismatches instead of the Euclidean distance

PAM algorithm

- ✓ Very close to K-means
- ✓ Each cluster is represented by the most centric object in the cluster (instead of the mean, which may not belong to the cluster)
- ✓ The centers of the clusters are always datapoints

✓ ejemplo_medoids.ipynb

Hierarchical methods

- ✓ The result of hierarchical methods is a dendrogram
- ✓ A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level
- ✓ The merging (agglomerative) or division (divisive) of clusters is ruled by some similarity measure, e.g., a sum of squares

Regarding the similarity measure...

Single-link clustering The distance between two clusters is assumed to be equal to the shortest distance from any member of one cluster to any member of the other cluster

Complete-link clustering The distance between two clusters is assumed to be equal to the longest distance from any member of one cluster to any member of the other cluster

Average-link clustering The distance between two clusters is equal to the average distance from any member of one cluster to any member of the other cluster

Regarding the similarity measure...

Single-link clustering The distance between two clusters is assumed to be equal to the shortest distance from any member of one cluster to any member of the other cluster

☹ “chaining effect”: a few points that form a bridge between two clusters cause the single-link clustering to unify these two clusters into one

Complete-link clustering The distance between two clusters is assumed to be equal to the longest distance from any member of one cluster to any member of the other cluster

Average-link clustering The distance between two clusters is equal to the average distance from any member of one cluster to any member of the other cluster

Regarding the similarity measure...

Single-link clustering The distance between two clusters is assumed to be equal to the shortest distance from any member of one cluster to any member of the other cluster

Complete-link clustering The distance between two clusters is assumed to be equal to the longest distance from any member of one cluster to any member of the other cluster

Average-link clustering The distance between two clusters is equal to the average distance from any member of one cluster to any member of the other cluster

☹ May cause elongated clusters to split and for portions of neighboring elongated clusters to merge

Examples hierarchical methods

- ✓ `example_distance_ngrams.ipynb`
- ✓ `http://scikit-learn.org/0.15/auto_examples/cluster/plot_lena_ward_segmentation.html`

Main advantages of hierarchical methods

Versatility For example, the single-link methods: good performance on datasets containing non-isotropic clusters, including well-separated, chain-linked and concentric clusters

Multiple partitions These methods produce not one partition, but multiple nested partitions, which allow different users to choose different partitions

Main disadvantages of hierarchical methods

- Inability to scale well** The time complexity of hierarchical algorithms is at least $\mathcal{O}M^2$ (where M stands for the number of instances).
- No back-tracking capability** It is possible to going back and undo previous steps

Comparison between different clustering methods

Algorithm	Loss function	# clusters	Cluster shape	Parameter Estimation Algorithm
K-means	Within-class squared distance from mean	Pre-determined	Isotropic	K-means
Single-link clustering	Maximum distance between a point and its nearest neighbor within a cluster	Data-dependent	Anisotropic	Greedy agglomerative clustering
Gaussian Mixture Models	$-\log PX$, (equivalent to within-class squared distance from mean)	Isotropic	Pre-determined	EM
Spectral Clustering	Balanced cut	Pre-determined	Anisotropic	Laplacian Eigenmaps + Kmeans/ thresholding eigenvector signs

Cluster algorithm	Complexity	High dimensional data
k-means	$\mathcal{O}(K \cdot M \cdot N)$ time $\mathcal{O}(M + K)$ space	No
Hierarchical clustering	$\mathcal{O}(M^2)$ time $\mathcal{O}(M^2)$ space	No
DBSCAN (Density Based Spatial Clustering)	$\mathcal{O}(M \log M)$ time	No
DENCLUE (Density Based Clustering)	$\mathcal{O}(M \log M)$ time	Yes

See pruebas_hdbscan.ipynb

DBSCAN vs Kmeans examples






✓ `example_dbscan.ipynb`

Assignment







1. ejemplo_metricas_distancias_final.ipynb

Create a pandas dataframe to compute cosine and smoothed cosine distances for a given set of artists

Some references. . . |

-  Andreopoulos, Bill, Aijun An, Xiaogang Wang, and Michael Schroeder (2009). “A roadmap of clustering algorithms: finding a match for a biomedical application”. In: *Briefings in Bioinformatics* 10.3, pp. 297–314.
-  Bishop, Christopher M (2006). *Pattern recognition and machine learning*. springer.
-  Choi, Seung-Seok, Sung-Hyuk Cha, and Charles C Tappert (2010). “A survey of binary similarity and distance measures”. In: *Journal of Systemics, Cybernetics and Informatics* 8.1, pp. 43–48.
-  Gersho, Allen and Robert M Gray (2012). *Vector quantization and signal compression*. Vol. 159. Springer Science & Business Media.
-  Jain, Anil K and Richard C Dubes (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Some references. . . II

-  Jain, Anil K, M Narasimha Murty, and Patrick J Flynn (1999). “Data clustering: a review”. In: *ACM computing surveys (CSUR)* 31.3, pp. 264–323.
-  Ling, Maurice HT (2010). “COPADS, I: Distance Coefficients between Two Lists or Sets.”. In: *Python Papers Source Codes* 2.
-  Pearl, Judea (2009). *Causality*. Cambridge university press.
-  Selim, Shokri Z and Mohamed A Ismail (1984). “K-means-type algorithms: a generalized convergence theorem and characterization of local optimality”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 1, pp. 81–87.
-  Taleb, Nassim Nicholas (2010). *The black swan:: The impact of the highly improbable fragility*. Vol. 2. Random House.
-  Xu, Rui and Donald Wunsch (2005). “Survey of clustering algorithms”. In: *IEEE Transactions on neural networks* 16.3, pp. 645–678.