Clusterability, Model Selection and Evaluation

Kaixun Hua – Data Mining Research Lab

Advisor: Prof. Dan A. Simovici

UMass Boston

Content

Introduction

Clustering is the prototypical unsupervised learning activity which consists in identifying cohesive and well-differentiated groups of records in data.

- increasing needs of clustering massive datasets;
  running clustering algorithms is expensive (especially for hierarchical and spectral clustering);
- data exist without any obvious clustering structure;
  however, if a clustering algorithm is applied, an irrelevant clustering structure may be returned;
- no ground truth in many practical clustering tasks (data is not labeled);
  different clustering algorithms give different (often implicit) measures of clustering quality;
- ambiguity exists for picking correct number of clusters;
  in practical, it is even harder for datasets with heavily imbalanced cluster structures.

Introduction

Our works tend to accomplish the following tasks:

- Deciding whether it is worth to do clustering on a dataset
- Improving the clustering result by twisting the distance space of dataset
- Determining the number of clusters in a dataset
- Unsupervised evaluation of clustering result

Clusterability Concept

A data set is clusterable if such groups exist; however, due to the variety in data distributions and the inadequate formalization of certain basic notions of clustering, determining data clusterability before applying specific clustering algorithms is a difficult task.

- Data clusterability is the existence of clustering (grouping) structure in data. This means that data can be partitioned in groups containing similar objects such that the groups are well-differentiated.
- We seek a measure of clusterability that quantifies the degree of how much inherent cluster structure the data possess.
- If a dissimilarity defined on a data set is close to an ultrametric it is natural to assume that the data set is clusterable.

Ultrametrics

Let $S \subseteq \mathbb{R}^k$ be a finite k-dimensional data set. An ultrametric is a mapping $d : S \times S \to \mathbb{R}_{\geq 0}$, which satisfies the following properties:

- Identity: $d(x, x) = 0$;
- Symmetry: $d(x, y) = d(y, x)$
- Triangle Inequality:

$$d(x, y) \leq \max\{d(y, z), d(x, z)\}, \forall x, y, z \in S, \tag{1}$$

r-spheric clustering

Definition

A *closed sphere* in (S, d) is a set $B[x, r]$ defined by

$$B[x, r] = \{y \in S \mid d(x, y) \leqslant r\}.$$

When (S, d) is an ultrametric space two spheres having the same radius r in (S, d) are either disjoint or coincide.

Definition

The collection of closed spheres of radius r in S, $\mathcal{C}_r = \{B[x, r] \mid r \in S\}$ is a partition of S; we refer to this partition as an *r-spheric clustering* of (S, d).

Every r-spheric clustering in an ultrametric space is a *perfect clustering* (all of its in-cluster distances are smaller than all of its between-cluster distances).

A Special Matrix Product

Let $\mathbb{P}_\infty = \{x \in \mathbb{R} \mid x \geqslant 0\} \cup \{\infty\}$, we define "$\vee$" and "$\wedge$" be the binary operation on $\mathbb{P}_\infty$ as follows:

Definition

$x \vee y = \min\{x, y\}$ and $x \wedge y = \max\{x, y\}$

Suppose $A \in \mathbb{P}_\infty^{m \times n}$ and $B \in \mathbb{P}_\infty^{n \times p}$,
We define a new product of two matrices as follows:

Definition

$C = A \otimes B \in \mathbb{P}_\infty^{m \times p}$ such that,

$$c_{ij} = \bigvee_{k=1}^{n} (a_{ik} \wedge b_{kj}) = \min\{\max\{a_{ik}, b_{kj}\} \mid 1 \leqslant k \leqslant n\} \tag{2}$$

Ultrametricity and Matrix Product

Definition

A is an *ultrametric matrix* if A is symmetric, $a_{ii} = 0$ and $a_{ij} \leqslant \max\{a_{ik}, a_{kj}\}$ for $1 \leqslant i, j, k \leqslant n$.

If we define $A \preccurlyeq B$ if $a_{ij} \geqslant b_{ij}$, we have the following consequence:

Theorem

If $A \in \mathbb{P}^{n \times n}$ is a dissimilarity matrix there exists $m \in \mathbb{N}$ such that

$$A \preccurlyeq A^2 \preccurlyeq \cdots \preccurlyeq A^m = A^{m+1} = \cdots = A^{m+d}, \forall d > 0$$

and A^m is an ultrametric matrix.

Ultrametricity

The *ultrametricity* of a matrix $A \in \mathbb{P}^{n \times n}$ is defined as follows:

Definition

Let $A \in \mathbb{P}^{n \times n}$ be the dissimilarity matrix of S, and $m(A)$ is the least integer that A^m is the ultrametric matrix, then the *ultrametricity* $\mathbf{u}(A) = \frac{n}{m}$

We refer to $m(A)$ as the *stabilization power* of the matrix A.
If $m(A) = 1$, A is ultrametric itself and $u(A) = n$.

The Definition of Clusterability [SH19]

Conjecture: a dissimilarity space (D, d) is more clusterable if the dissimilarity is closer to an ultrametric, hence if $m(A_D)$ is small.

Definition

The *clusterability of a data set* D is the number

$$\text{clust}(D) = \frac{n}{m(A_D)},$$

where $n = |D|$, A_D is the dissimilarity matrix of D and $m(A_D)$ is the stabilization power of A_D.

The lower the stabilization power, the closer A is to an ultrametric matrix, and thus, the higher the clusterability of the data set.

Empirical Study

Lattice-like Toy Data Generation:

- Generate series of datasets by assigning data points on the positions with integer pairs.
- Create dissimilarity matrix by Manhattan distance
- Move data points to different locations to generate distinct structured clusterings.

Real Data Set:

- Iris, Swiss, Faithful, Rivers, Trees
- USAJudgeRatings, USArrests, Attitude, Cars

Experiments - Lattice Toy Data

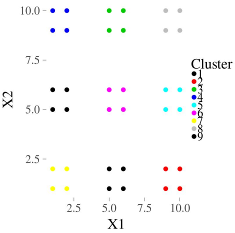

Figure 1: $k = 9$

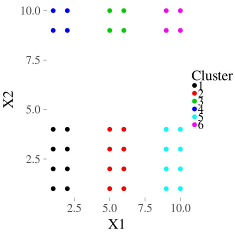

Figure 2: $k = 6$

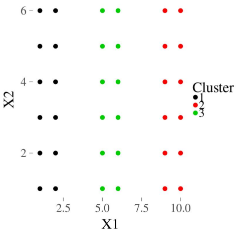

Figure 3: $k = 3$

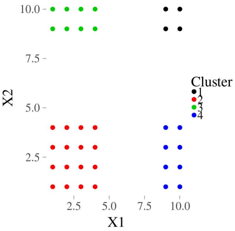

Figure 4: $k = 4$

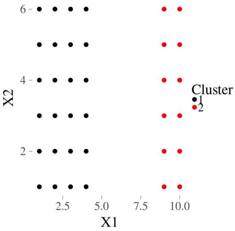

Figure 5: $k = 2$

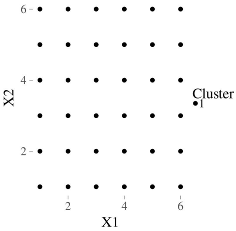

Figure 6: $k = 1$

Histogram of Original Distance

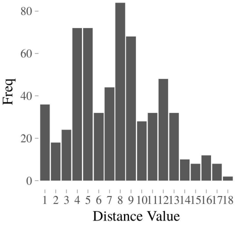

Figure 7: $k = 9$

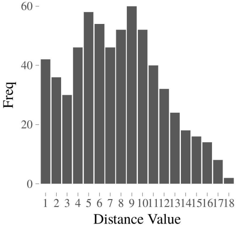

Figure 8: $k = 6$

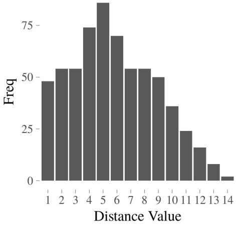

Figure 9: $k = 3$

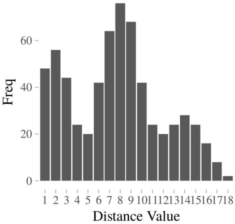

Figure 10: $k = 4$

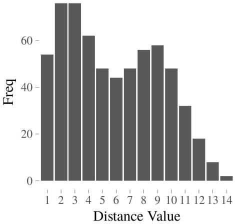

Figure 11: $k = 2$

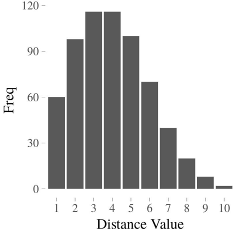

Figure 12: $k = 1$

Histogram of Distance after Power Operation

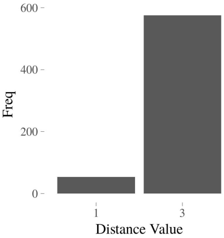

Figure 13: $m = 6$

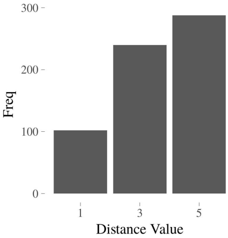

Figure 14: $m = 4$

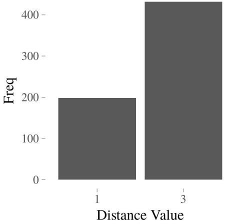

Figure 15: $m = 5$

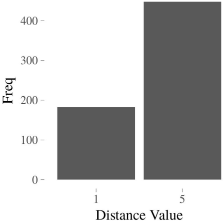

Figure 16: $m = 5$

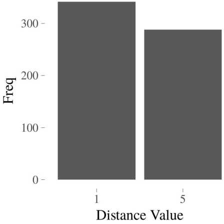

Figure 17: $m = 7$

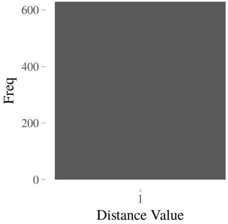

Figure 18: $m = 9$

Distance Collapse

Given dataset with 4 perfect-uniform cluster and generated with the same scheme above:

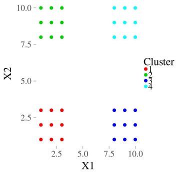

Figure 19: Original dataset with four clusters

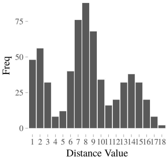

Figure 20: Histogram of distinct value in the original matrix

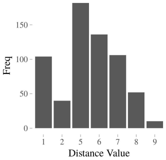

Figure 21: Histogram of distinct value in the matrix after one multiplication

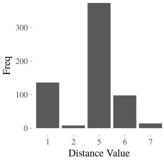

Figure 22: Histogram of distinct value in the matrix after two multiplication

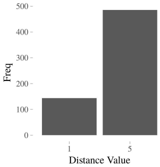

Figure 23: Histogram of distinct value in the matrix after three multiplication

Validation on Real Data Sets

Table 1: All clusterable datasets have values greater than 5 for their clusterability; all non-clusterable datasets have values no larger than 5.

Dataset	n	Dip	Silv.	$m(A_D)$	clust(D)
iris	150	0.0000	0.0000	14	10.7
swiss	47	0.0000	0.0000	6	7.8
faithful	272	0.0000	0.0000	31	8.7
rivers	141	0.2772	0.0000	22	6.4
attitude	30	0.9040	0.9449	6	5
trees	31	0.3460	0.3235	7	4.4
USAJudgeRatings	43	0.9938	0.7451	10	4.3
USArrests	50	0.9394	0.1897	15	3.3
cars	50	0.6604	0.9931	15	3.3

Clustering by Elevating Clusterability

- We can improve the quality of clustering result by increasing the ultrametricity of its dissimilarity matrix.
- By definition, the new dissimilarity matrix will be more clusterable.
- Better performance can be achieved on the powered dissimilarity matrix(ultrametric distance matrix)

Entangled spirals dataset

Clustering by promoting ultrametricity (clusterability)
k-medoids clustering algorithm are performed on two dissimilarity matrices:

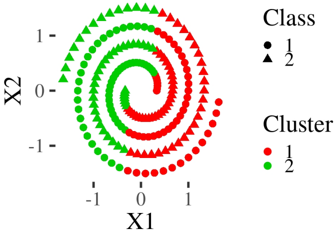

Figure 24: Clustering Result on Spiral dataset based on original dissimilarity matrix

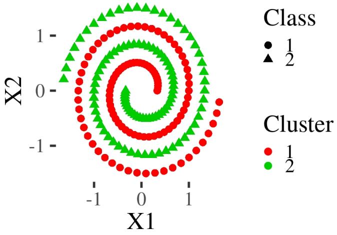

Figure 25: Clustering Result on Spiral dataset based on the maximum ultrametricity matrix

Entangled spiral dataset

Distance matrix of dataset with two entangled spirals with total of 200 data points

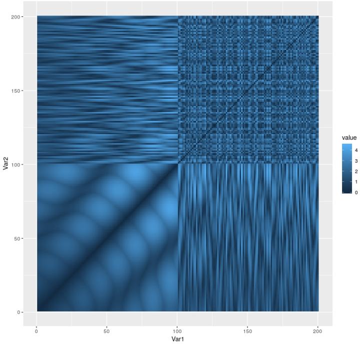

Figure 26: Original Distance matrix on Spiral dataset

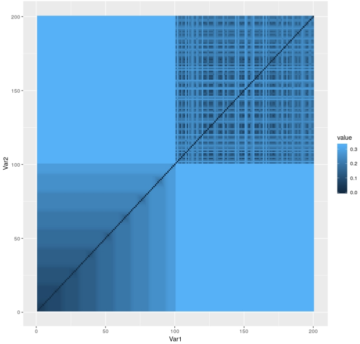

Figure 27: Maximum ultrametricity Distance matrix on Spiral dataset

Model Selection

Difficulties in model selection in clustering:

- most clustering algorithms need a parameter k that specifies the number of clusters to detect;
- the definition of an optimal model is ambiguous;
- clustering is even more difficult if the clusters are heavily imbalanced.

Generalized Partitional Entropy

Definition

A *partition* of set S is a non-empty collection of pairwise disjoint and non-empty subsets of S referred to as *blocks*, $\pi = \{B_1, B_2, \ldots B_n \mid \bigcup_{i=1}^{n} B_i = S\}$

The set of partitions of a set S is denoted as $\text{PART}(S)$

Definition

If $\pi = \{B_1, B_2, \ldots B_n \mid \bigcup_{i=1}^{n} B_i = S\} \in \text{PART}(S)$ is a partition of a set S and $\beta > 0$, then its β-entropy, H_β, is given by:

$$H_\beta(\pi) = \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{i=1}^{n} \left(\frac{|B_i|}{|S|}\right)^{\beta}\right) \quad (3)$$

Some Special β

Shannon Entropy:

$$\lim_{\beta \to 1} H_\beta(\pi) = -\sum_{i=1}^{n} \frac{|B_i|}{|S|} \log \frac{|B_i|}{|S|} \tag{4}$$

Gini Index:

$$H_2(\pi) = 2\left(1 - \sum_{i=1}^{n} \left(\frac{|B_i|}{|S|}\right)^2\right). \tag{5}$$

Conditional Entropy and Metric on PART(S)

Definition

If $\pi = \{B_1, B_2, \ldots B_n\} \in \text{PART}(S)$ and $C \subseteq S$, The *trace of* π on C is the partition $\pi_C \in \text{PART}(C)$ given by

$$\pi_C = \{B_i \cap C \mid B_i \in \pi, B_i \cap C \neq \emptyset\}$$

Theorem

If $\pi = \{B_1, B_2, \ldots B_n\}$ *and* $\sigma = \{C_1, C_2, \ldots C_n\}$ *are two partitions in* PART(S), *then*

$$\begin{aligned} H_\beta(\pi \wedge \sigma) &= H_\beta(\sigma) + \sum_{j=1}^{m} \left(\frac{|C_j|}{|S|}\right)^\beta H_\beta(\pi_{C_j}) \\ &= H_\beta(\pi) + \sum_{j=1}^{m} \left(\frac{|B_j|}{|S|}\right)^\beta H_\beta(\sigma_{B_j}) \end{aligned}$$

Conditional Entropy and Metric on PART(S)

Definition

The *conditional β-entropy* $H_\beta(\pi|\sigma)$ is defined as

$$H_\beta(\pi|\sigma) = H_\beta(\pi \wedge \sigma) - H_\beta(\sigma)$$

Theorem

The function $d_\beta : \mathrm{PART}(S) \times \mathrm{PART}(S) \to \mathbb{R}$ *defined by*

$$d_\beta(\pi, \sigma) = H_\beta(\pi|\sigma) + H_\beta(\sigma|\pi)$$

is a metric on PART(S).

Imbalanced Partitions

Let $h_\beta : [0,1] \longrightarrow \mathbb{R}$ be defined by $h_\beta(x) = \frac{x - x^\beta}{1 - 2^{1-\beta}}$ where $\beta > 0$ and $\beta \neq 1$.

Theorem

h_β is a concave function for $\beta > 0$ and $\beta \neq 1$.

We can rewrite the β-entropy as follows

$$\begin{aligned} H_\beta(\pi) &= \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{i=1}^{n} \left(\frac{|B_i|}{|S|} \right)^\beta \right) \\ &= \sum_{i=1}^{n} h_\beta \left(\frac{|B_i|}{|S|} \right), \end{aligned}$$

Behavior of function $h_\beta(x)$

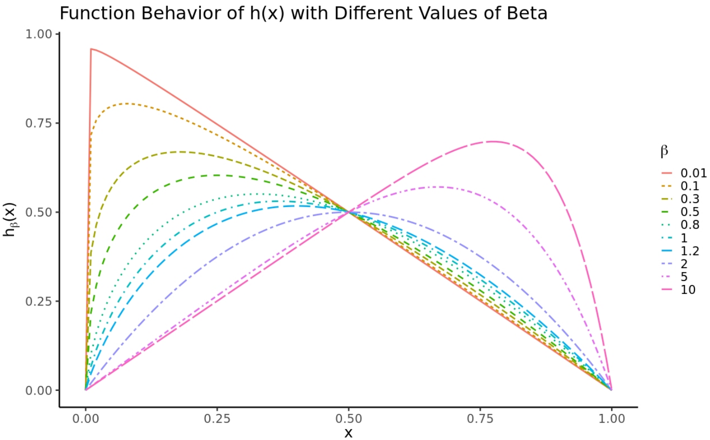

Figure 28: *Behavior of Function $h_\beta(x)$ with different β. Here,* $x = \frac{|B_i|}{|S|} \in [0, 1], i \in [1, n]$

Sum of Square-Errors

Let S be the set of objects to be clustered. We assume that S is a subset of $\mathbb{R}^n$ equipped with the Euclidean metric.

Definition

The *center* $\mathbf{c}_C$ *of a subset* C of S is defined as $\mathbf{c}_C = \frac{1}{|C|} \sum\{\mathbf{o} \mid \mathbf{o} \in C\}$. For a partition $\pi = \{C_1, C_2, \ldots, C_m\}$ of S the *sum of square errors* sse of π is defined as

$$\text{sse}(\pi) = \sum_{i=1}^{m} \sum_{\mathbf{o} \in C_i} d^2(\mathbf{o}, \mathbf{c}_{C_i}). \tag{6}$$

Current Approaches

Intuitively, the optimal choice of k will strike a balance between the cohesion of data, and sum of square errors:

- Elbow Method
- AIC: $\text{argmin}_k[-2L(k) + 2kd]$
- BIC: $\text{argmin}_k[-2L(k) + ln(n)kd]$

where k is the number of clusters, $L(\cdot)$ is the likelihood function of model with parameter k, d represents the dimension and n is the data size.

Dual Criteria Compromise

We aim to look for the optimal model that minimize both the model distortion and model complexity simultaneously [HS18, HS19].

	π	ι_S	$\cdots$	ω_S
Model Complexity	$\mathcal{H}_\beta(\pi)$	$\frac{1-n^{1-\beta}}{1-2^{1-\beta}}$	$\searrow$	0
Model Distortion	$\text{sse}(\pi)$	0	$\nearrow$	$\sum_{\mathbf{o}\in S} \parallel \mathbf{o} - \mathbf{c} \parallel^2$

- ι_S has the most balanced clusters and it is the least cohesive clustering;
- ω_S is the least balanced cluster but it is the most cohesive clustering.

Multi-objective Optimization and Pareto Optimal

- Decisions should be taken in the presence of trade-offs between two conflicting objectives.
- Model selection can be treated as a multi-objective optimization problem.

Definition

Let $\pi, \sigma \in \text{PART}(S)$. The partition σ *dominates* π if $H(\sigma) \leqslant H(\pi)$ and $\text{sse}(\sigma) \leqslant \text{sse}(\pi)$.
A partition $\tau \in \text{PART}(S)$ is *Pareto optimal* if there is no other partition that dominates τ.

If a partition π is Pareto optimal, then it is no worse than another partitions from the point of view of ($H(\pi)$ and $\text{sse}(\pi)$) and is better in at least one of these criteria.

Pareto Front

Definition

The set of partitions that are not dominated by other partitions is the *Pareto front*.

It allow us to define a natural number of clusters using the Pareto front of the following bi-criterial problem.
Let $\mathbf{F} : \mathrm{PART}(S) \longrightarrow \mathbb{R}^2$, where

$$\mathbf{F}(\pi) = (H(\pi), \mathrm{sse}(\pi))$$

where $\pi \in \mathrm{PART}(S)$.

Pareto Front

Examples for Iris and Libras dataset. We apply k-means clustering algorithm. Both are normalized into $[0, 1]$.

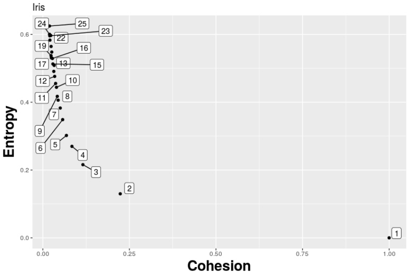

Figure 29: Pareto Front for Iris Dataset

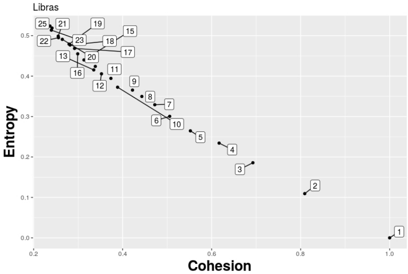

Figure 30: Pareto Front for Libras Dataset

Hypervolume

A popular indicator for multi-objective optimization problem. It estimates the closeness of the estimated solutions to the true Pareto front.

Definition

The *hypervolume* that corresponds to a partition π is

$$\mathrm{HV}(\pi) = (H(\iota_S) - H(\pi))(\mathrm{sse}(\omega_S) - \mathrm{sse}(\pi))$$

The optimal partition for a dataset is obtained as

$$\pi_{opt} = \operatorname*{argmax}_{\pi} \mathrm{HV}(\pi)$$

k-means, Hierarchical Clustering and Contour Curves [HS19]

- If a natural clustering structure exists, two different clustering algorithms will generate similar clustering results with optimal number of clusters.
- We evaluate partitional models with the contour curves of the distance between partitions generated from k-means and ward-linkage hierarchical clustering algorithm.
- The sink on the contour map can be an indicator of the “natural” number of clusters.

Examples of the contours of *Iris* dataset and an artificial dataset with 10 Gaussian Distributed clusters.

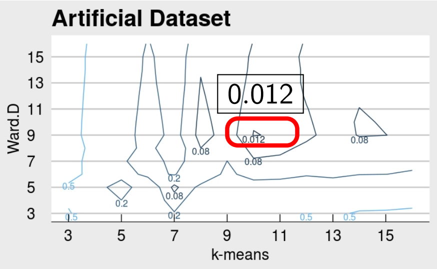

Figure 31: 10-cluster Artificial Dataset

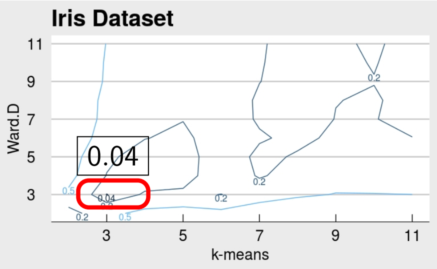

Figure 32: *Iris* Dataset

Empirical Study

Synthetic datasets for testing:

- clusters that are well separated;
- clusters that are well separated but closer with each other;
- clusters that have different density;
- clusters that have different sizes and number of points;
- clusters that overlap.

Real datasets for testing:

- *Iris Data*
- *Wine Recognition Data*
- *LIBRAS Movement Database*
- *Pen-Based Recognition of Handwritten Digits*
- *E. Coli Dataset*
- *Vowel Recognition*
- *Poker Dataset*

Empirical Study–Synthetic datasets

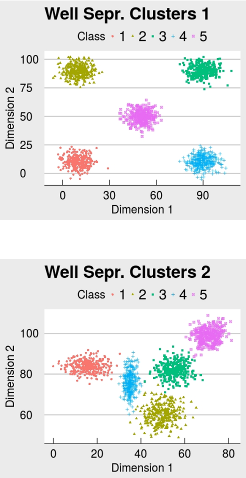

Figure 33: Data Structure

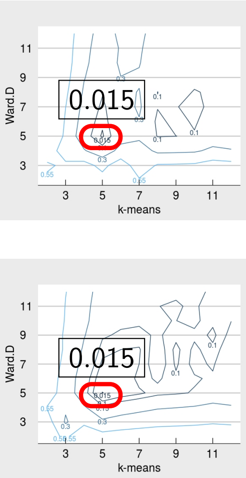

Figure 34: Contour Map

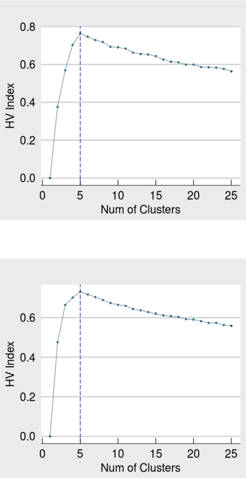

Figure 35: HV-index

Empirical Study–Synthetic datasets

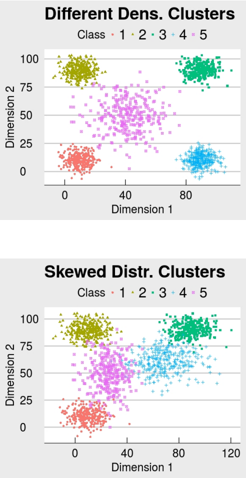

Figure 36: Data Structure

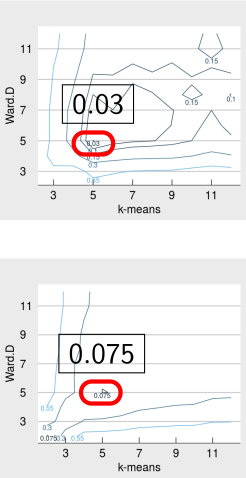

Figure 37: Contour Map

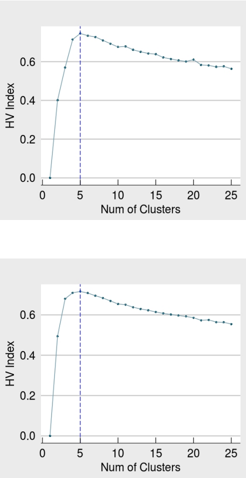

Figure 38: HV-index

Empirical Study–Synthetic datasets

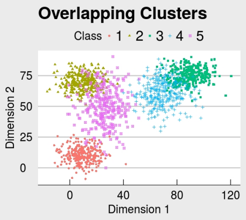

Figure 39: Data Structure

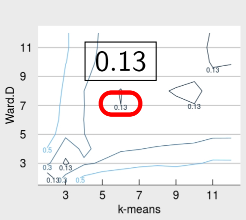

Figure 40: Contour Map

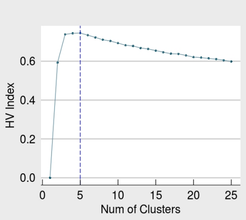

Figure 41: HV-index

Empirical Study–Results

Table 2: Comparison between the number of clusters for datasets; g represents the number of clusters obtained by using the log-likelihood function of Gaussian Mixture Model while k represents those numbers when using the sum of squared errors.

Data Sets	β	natural number of clusters(CPU Times[seconds])							
		Gap Stat.	Jump Mthd.	Pred. Strgth.	AIC(g/k)	BIC(g/k)	RIM	HV Index	Cntr. Mthd.
Well Sep. I(5)	1.00	5(3.92)	5(0.87)	3(2.90)	8(1.23)/30(0.29)	8(1.14)/30(0.34)	12(976)	**5**(0.92)	5
Well Sep. II(5)	1.00	5(4.04)	5(0.92)	5(2.82)	13(1.19)/30(1.11)	5(1.23)/30(1.12)	6(977)	**5**(0.90)	5
Diff. Dens.(5)	1.00	5(4.13)	5(0.97)	5(2.96)	5(1.30)/30(0.31)	5(1.11)/30(0.37)	4(968)	**5**(0.95)	5
Skw. Dist.(5)	1.00	5(4.17)	30(1.06)	5(3.05)	6(1.49)/30(0.32)	5(1.13)/30(0.33)	3(968)	**5**(0.99)	5
Ovrlp.(5)	0.95	3(4.26)	3(1.09)	**5**(2.87)	6(1.34)/30(0.41)	5(1.19)/30(0.41)	1(960)	**5**(0.97)	3/6
Iris(3)	1.00	4(0.65)	24(0.33)	3(1.60)	30(0.11)/5(0.48)	30(0.13)/4(0.53)	25(962)	**3**(0.55)	3
Wine(3)	1.0	1(1.22)	28 (0.93)	**3** (2.01)	30(0.59)/30(0.26)	7(0.50)/30(0.50)	19(964)	4 (0.65)	8
Libras(15)	1.00	6(9.65)	30(1.96)	2(5.52)	30(1.66)/2(1.27)	30(1.42)/1(1.09)	**13**(964)	**13**(1.95)	15/16
Ecoli(8)	0.9	6 (1.90)	25 (1.32)	3 (1.96)	30(0.51)/2(0.12)	11(0.38)/1(0.41)	**9**(967)	**7**(0.65)	7
Vowel(11)	0.8	4 (5.67)	29 (1.53)	4 (2.9)	30(1.21)/27(0.32)	30(1.07)/19(0.33)	5(983)	**9**(1.35)	13
PenDigits(10)	1.20	22(206.2)	29(19.41)	6(25.10)	30(7.52)/30(5.53)	30(7.16)/30(5.38)	-	**9**(9.27)	15
Poker(1-9)(9)	1.4	4 (1889)	29 (1574)	2 (2080)	30(256)/30(926)	30(240)/30(915)	-	**10**(477)	-

Empirical Study–Imbalanced Clustering Structure

β selection for imbalanced data sets: the more imbalanced the data clusters are, the lower β we should choose.
Three datasets are used for experiments; during the experiments a portion of one cluster from each dataset is eliminated:

- skewed distribution synthetic dataset;
- *Iris data*;
- *Wine recognition data*.

Range of β that yields correct k clusters for the modified dataset:

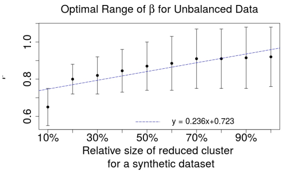

Figure 42: $k = 5$, Synthetic Data

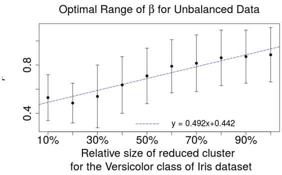

Figure 43: $k = 3$, Iris Data

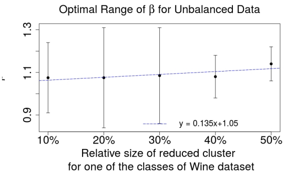

Figure 44: $k = 3$, Wine Data

Reference

Kaixun Hua and Dan A Simovici.
Dual criteria determination of natural clustering structures in data.
In *International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. IEEE, 2018.

Kaixun Hua and Dan A Simovici.
On finding natural clustering structures in data.
In *Transactions on Pattern Analysis and Machine Intelligence*. Under Review, 2019.

Dan A Simovici and Kaixun Hua.
Data ultrametricity and clusterability.
In *International Conference on Mathematical Models & Computational Techniques in Science & Engineering*. To be published, 2019.

Thank you