DIP-MEANS: AN INCREMENTAL CLUSTERING METHOD FOR ESTIMATING THE NUMBER OF CLUSTERS

Argyris Kalogeratos and Aristidis Likas
akaloger@cs.uoi.gr arly@cs.uoi.gr

Computer Science
University of Ioannina

Neural Information
Processing Systems

CONTRIBUTION

We propose the robust ***dip-dist criterion*** for cluster structure evaluation under a simple but fundamental assumption: each cluster to admit a *unimodal* distribution. Our novel criterion does not require the actual data vectors. It applies a statistical hypothesis test (SHT), the Hartigans' ***dip test*** [1], on the distribution of the pairwise distances (or similarities) between a reference point of the set, termed 'viewer', to the rest of members. Dip-dist is incorporated in an efficient incremental clustering method called ***dip-means*** and straightforwardly extended in ***kernel dip-means*** which is applicable in kernel space.

MOTIVATION

Clustering is very broadly applied, however, the number of clusters k is usually set with ad hoc criteria (AHC), e.g. Silhouette or Information Criteria (BIC, AIC, etc). Any attempt to address the problem requires assumptions about *what the clusters we seek look like* (shape, density distribution) and, definitely, it is of great value for any assumption to be verifiable with a theoretically sound SHT.

ℹ Existing methods, either following the AHC approach such as ***x-means***, or using an SHT such as ***g-means*** & ***projected g-means***, are lacking generality since they make or imply Gaussianity assumptions.

THE DIP-DIST CRITERION

■ **What's a cluster, anyway?** We only assume that the empirical density distribution of a cluster to be unimodal. Hartigans' *dip test* is a powerful unimodality SHT.

■ **Is this all about an SHT?** The novel idea is to examine the distribution of pairwise distances between a 'viewer' datapoint and the objects of a set for unimodality.

■ **And the benefits are...**

✔ Many unimodal distributions are identifiable, e.g. Uniform, Gaussian, Student-t, etc.

✔ Unimodality SHT is applied on the 1d ecdf features.

✔ The actual data vectors are not required. Potential use in kernel space, or on not strictly numerical vectors.

✔ ... robust & efficient cluster structure evaluation.

Algorithm for dip-dist criterion [$O(bn\log n+n^2)$]:

1. Compute the ecdf U_n^r and the respective $dip(U_n^r)$, r=1...b, for the Uniform sample distributions.

2. Compute $F_n^{(x_i)}(t)=\frac{1}{n}\sum_{x_j\in c}\{Dist(x_i,x_j)\le t\}$ and $dip(F_n^{(x_i)})$, i=1...n, for all datapoint viewers in set c.

3. Do the SHT for each viewer using a significance level α and p-value $P^{(x_i)}=\#[dip(F_n^{(x_i)})\le dip(U_n^r)]\ /\ b$, r=1...b.

4. If there exist enough *split viewers* (v) in the set, we assign $score_c=\frac{1}{|v|}\sum_{x_i\in v} dip(F_n^{(x_i)})$, otherwise $score_c=0$.

DIP-DIST EXAMPLE

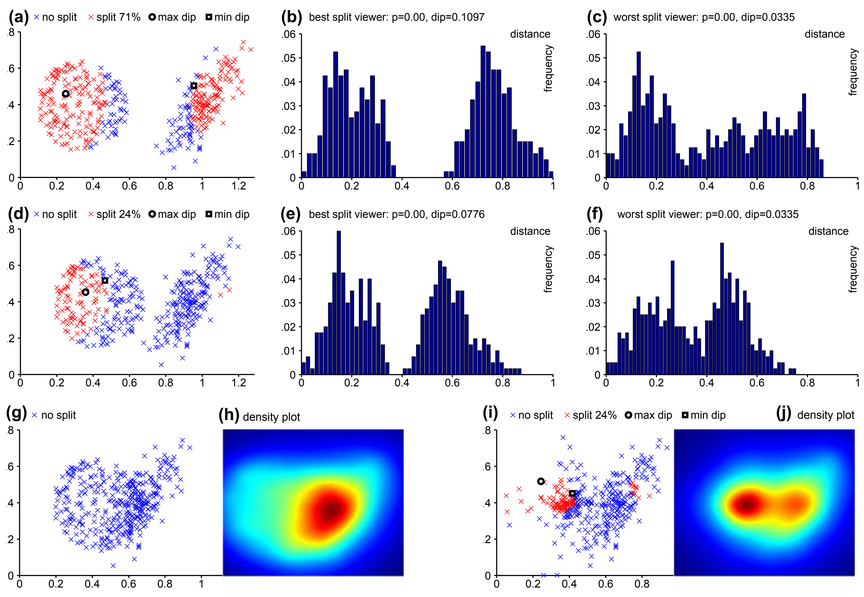

Fig. 1: 2d synthetic data with two structures of 200 datapoints each. The split viewers are denoted in red color. **(a)** A Uniform spherical & elliptic Gaussian structure. **(b, c)** The histograms of pairwise distances of the strongest and weakest split viewer. **(d)** The two structures come closer; the split viewers are reduced, so does the dip value for the best of them, which indicates that the two structures became less distinguishable. **(g)** The structures are no longer distinguishable as the density map in **(h)** shows one mode. **(i)** The Uniform spherical is replaced with a structure generated from a Student-t distribution.

DIP-MEANS CLUSTERING ALGORITHM

Dip-means(X, k_{init}, α, v_{thd})

1: $k \leftarrow k_{init}$
2: $\{C, M\} \leftarrow$ kmeans(X, k)
3: **do while** changes in cluster number occur
4: **for** j=1,...,k
5: $score_j \leftarrow$ unimodalityTest(c_j, α, v_{thd})
6: **end for**
7: **if** $\max_j(score_j) > 0$
8: target $\leftarrow \text{argmax}_j\ (score_j)$
9: $\{m_L, m_R\} \leftarrow$ splitCluster(c_{target})
10: M $\leftarrow \{M\text{-}m_{target}, m_L, m_R\}$
11: $\{C, M\} \leftarrow$ kmeans(X, M)
12: **end if**
13: **end do**
14: **return** $\{C, M\}$

■ It is an incremental method that combines three individual components:
- a local search clustering technique [k-means]
- a cluster structure evaluation and selection criterion [dip-dist]
- a cluster splitting procedure, [10 trials of 2-means]

■ In each incremental iteration...
- to avoid overestimation of k only the candidate with max *score* is split.
- the k+1 clusters are refined using k-means.

■ The procedure terminates when no split candidates are identified.

ℹ ***Kernel dip-means*** uses kernel k-means and a modified splitting procedure.

APPLICATION ON SYNTHETIC DATA

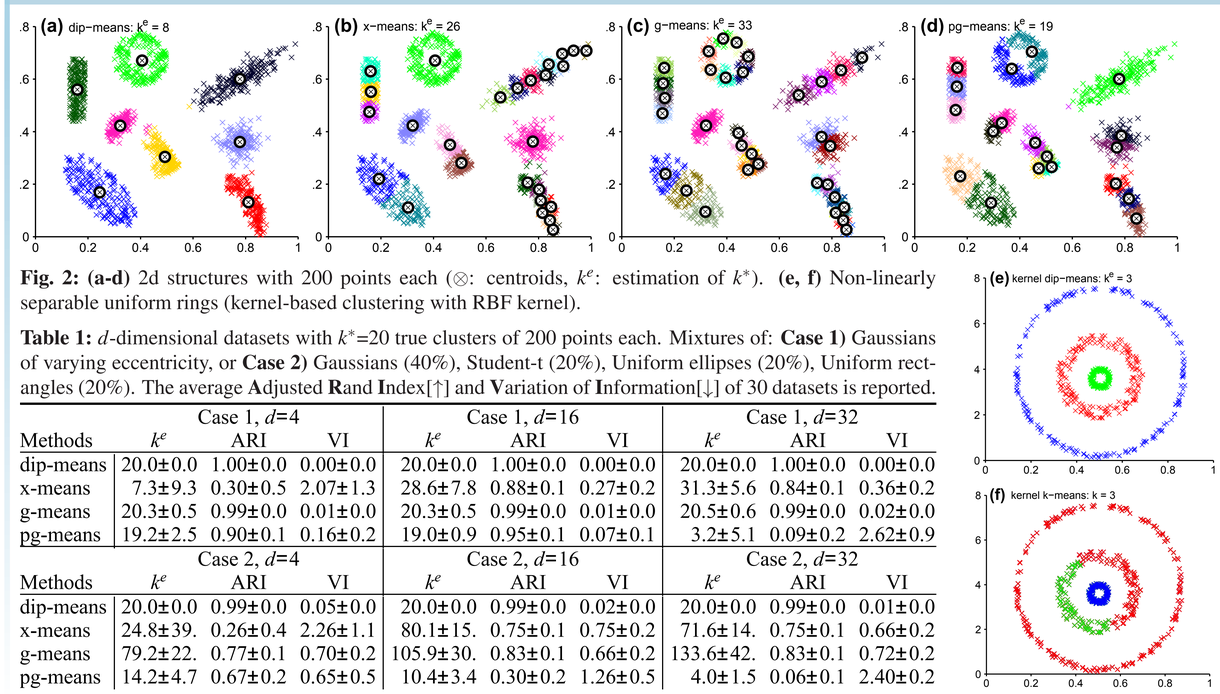

Fig. 2: (a-d) 2d structures with 200 points each (⊗: centroids, k^e: estimation of k^*). **(e, f)** Non-linearly separable uniform rings (kernel-based clustering with RBF kernel).

Table 1: d-dimensional datasets with k^*=20 true clusters of 200 points each. Mixtures of: **Case 1)** Gaussians of varying eccentricity, or **Case 2)** Gaussians (40%), Student-t (20%), Uniform ellipses (20%), Uniform rectangles (20%). The average **A**djusted **R**and **I**ndex[↑] and **V**ariation of **I**nformation[↓] of 30 datasets is reported.

	Case 1, d=4			Case 1, d=16			Case 1, d=32		
Methods	k^e	ARI	VI	k^e	ARI	VI	k^e	ARI	VI
dip-means	20.0±0.0	1.00±0.0	0.00±0.0	20.0±0.0	1.00±0.0	0.00±0.0	20.0±0.0	1.00±0.0	0.00±0.0
x-means	7.3±9.3	0.30±0.5	2.07±1.3	28.6±7.8	0.88±0.1	0.27±0.2	31.3±5.6	0.84±0.1	0.36±0.2
g-means	20.3±0.5	0.99±0.0	0.01±0.0	20.3±0.5	0.99±0.0	0.01±0.0	20.5±0.6	0.99±0.0	0.02±0.0
pg-means	19.2±2.5	0.90±0.1	0.16±0.2	19.0±0.9	0.95±0.1	0.07±0.1	3.2±5.1	0.09±0.2	2.62±0.9
	Case 2, d=4			Case 2, d=16			Case 2, d=32		
Methods	k^e	ARI	VI	k^e	ARI	VI	k^e	ARI	VI
dip-means	20.0±0.0	0.99±0.0	0.05±0.0	20.0±0.0	0.99±0.0	0.02±0.0	20.0±0.0	0.99±0.0	0.01±0.0
x-means	24.8±39.	0.26±0.4	2.26±1.1	80.1±15.	0.75±0.1	0.75±0.2	71.6±14.	0.75±0.1	0.66±0.2
g-means	79.2±22.	0.77±0.1	0.70±0.2	105.9±30.	0.83±0.1	0.66±0.2	133.6±42.	0.83±0.1	0.72±0.2
pg-means	14.2±4.7	0.67±0.2	0.65±0.5	10.4±3.4	0.30±0.2	1.26±0.5	4.0±1.5	0.06±0.1	2.40±0.2

CLUSTERING REAL-WOLD DATASETS

	$PD3_{te}$ (k^*=3)			$PD4_{te}$ (k^*=4)			$PD10_{te}$ (k^*=10)		
Methods	k^e	ARI	VI	k^e	ARI	VI	k^e	ARI	VI
dip-means	3	**0.879**	**0.332**	4	**0.626**	**0.545**	7	0.343	**1.587**
x-means	155	0.031	3.792	194	0.039	3.723	515	0.041	3.825
g-means	21	0.226	1.800	36	0.209	2.049	73	0.295	1.961
pg-means	4	0.835	0.359	10	0.576	0.954	13	**0.447**	1.660
	$PD3_{tr}$ (k^*=3)			$PD4_{tr}$ (k^*=4)			$PD10_{tr}$ (k^*=10)		
Methods	k^e	ARI	VI	k^e	ARI	VI	k^e	ARI	VI
dip-means	3	**0.963**	**0.116**	4	**0.522**	**0.841**	9	0.435	**1.452**
x-means	288	0.018	4.378	381	0.020	4.372	942	0.024	4.387
g-means	52	0.106	2.641	58	0.143	2.464	149	0.160	2.605
pg-means	5	0.655	0.740	8	0.439	1.320	14	**0.494**	1.504
	Coil3 (k^*=3)			Coil4 (k^*=4)			Coil5 (k^*=5)		
Methods	k^e	ARI	VI	k^e	ARI	VI	k^e	ARI	VI
dip-means	3	**1.000**	**0.000**	5	**0.912**	**0.173**	4	**0.772**	**0.308**
x-means	8	0.499	0.899	11	0.499	0.951	15	0.601	0.907
g-means	7	0.669	0.650	12	0.502	0.977	18	0.434	1.204

Table 2: Bold indicates the best value for the results in real datasets. **Pendigits** (UCI) contains 16-dimensional vector representing written digits from 0-9. We used the training $\mathbf{PD}_{tr}$ and testing set $\mathbf{PD}_{te}$ with 7494 and 3498 instances, respectively and subsets that contain the digits $\{0,2,4\}$ ($\mathbf{PD3}_{tr}$ and $\mathbf{PD3}_{te}$) and $\{3,6,8,9\}$ ($\mathbf{PD4}_{tr}$ and $\mathbf{PD4}_{te}$). **Coil-100** contains 72 images taken from different angles for each one of the 100 included objects. We used tree subsets **Coil3**, **Coil4**, **Coil5**, with images from 3, 4 and 5 objects, respectively. The images are represented by the *Bag of Visual Words* model using 1000 visual words.

References

[1] J.A. Hartigan and P. M. Hartigan. The dip test of unimodality. *The Annals of Statistics*, 13(1), pp. 70–84, 1985.