Statistical and Computational Trade-Offs in Imbalanced Kernel Clustering

Jing Zhang, Yucong Dai, Hong Tao, Chenping Hou

Frontiers of Computer Science, DOI: 10.1007/s11704-026-52144-2

Problems & Ideas

- Problems of existing methods for imbalanced clustering:
  - Quadratic scaling renders kernel-based imbalanced clustering impractical for large-scale nonlinear data.
  - Existing accelerated kernel methods lack specialized statistical guarantees and fail to establish optimal risk bounds for imbalanced clustering.
- Ideas: A framework of fast Imbalanced Kernel k-Means, exploring both computational demands and statistical analysis.

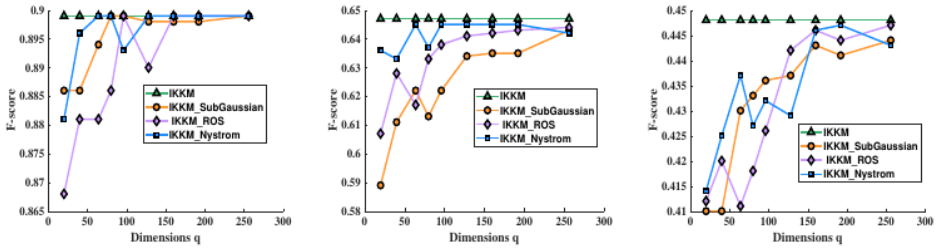


(e) Pavia Center (f) Pavia Center

Illustration of a large-scale HSI dataset with severe class imbalance, highlighting the challenges of large-scale imbalanced clustering.

Main Contributions

- Contributions:
  - Establish the first excess clustering risk bound for IKKM, providing a statistical learning guarantee for imbalanced kernel clustering;
  - By integrating sketching techniques into the IKKM framework, we propose a fast IKKM with the excess clustering risk upper bound. This is the first theoretical guarantee for sketching-based IKKM



(a) pageblocks (b) a9a (c) connect-4

Figure 1 Comparing the F-score of clustering across various dimensions (q) using the IKKM, IKKM_S, IKKM_R, and IKKM_N methods on three datasets.

Table 4 The average ACC results of methods with 100% ratio on each dataset.

Datasets	KKM	SGKKM	ROSKKM	NKKM	IKKM	IKKM_S	IKKM_R	IKKM_N
pageblocks	.899(.000)	.899(.000)	.899(.000)	.899(.001)	.899(.000)	.899(.000)	.899(.000)	.899(.000)
a9a	.715(.002)	.681(.032)	.691(.021)	.711(.009)	.716(.001)	.693(.015)	.695(.047)	.714(.004)
w8a	.778(.003)	.784(.003)	.771(.011)	.780(.002)	.925(.000)	.922(.003)	.924(.003)	.925(.000)
shuttle	.785(.002)	.785(.000)	.785(.000)	.786(.000)	.793(.016)	.786(.000)	.793(.017)	.792(.009)
connect-4	.384(.001)	.377(.022)	.372(.011)	.378(.017)	.442(.018)	.438(.019)	.437(.015)	.440(.006)
SVHN	.123(.001)	.124(.001)	.123(.003)	.123(.000)	.134(.001)	.134(.001)	.134(.004)	.135(.001)
skin-nonskin	—	.550(.002)	.548(.002)	.550(.001)	—	.565(.006)	.563(.017)	.560(.004)
covtype	—	.315(.006)	.313(.009)	.299(.028)	—	.346(.009)	.335(.020)	.340(.003)

Table 3 The training time cost (second) of proposed methods with different ratios on each dataset.

Dataset	Ratio	KKM	SGKKM	ROSKKM	NKKM	IKKM	IKKM_S	IKKM_R	IKKM_N
pageblocks	20%	0.018	0.008	0.013	0.006	0.270	0.092	0.102	0.075
	60%	0.080	0.013	0.026	0.009	2.809	0.170	0.193	0.158
	100%	0.225	0.020	0.038	0.013	7.434	0.215	0.231	0.207
a9a	20%	0.588	0.043	0.050	0.024	11.457	0.223	0.253	0.195
	60%	7.545	0.144	0.152	0.056	79.098	0.665	0.751	0.624
	100%	20.185	0.198	0.266	0.119	230.179	1.588	1.822	1.571
w8a	20%	1.019	0.024	0.069	0.016	23.162	0.266	0.309	0.253
	60%	15.325	0.091	0.269	0.060	199.229	1.470	1.726	1.453
	100%	102.964	0.566	0.754	0.386	525.731	2.293	2.516	2.136
shuttle	20%	1.298	0.025	0.100	0.024	34.491	0.476	0.495	0.433
	60%	15.325	0.103	0.284	0.086	298.976	1.795	2.160	1.712
	100%	32.599	0.163	0.403	0.126	820.829	3.064	3.580	3.057
connect-4	20%	9.441	0.159	0.187	0.151	38.961	0.402	0.459	0.388
	60%	66.844	0.740	0.958	0.632	357.776	1.992	2.233	1.936
	100%	288.554	1.341	1.698	1.284	1050.991	3.260	3.443	2.918
SNHN	20%	20.110	0.460	0.518	0.367	49.401	0.720	0.773	0.657
	60%	332.081	0.828	1.755	0.782	477.741	2.749	3.209	2.733
	100%	838.442	3.072	4.113	4.657	1368.900	4.924	5.384	4.474
skin_nonskin	20%	79.457	0.238	0.263	0.202	476.049	2.013	2.169	1.959
	60%	—	0.892	0.963	0.654	—	6.860	7.690	6.734
	100%	—	1.237	1.621	1.190	—	11.851	12.934	11.175
Covertype	20%	—	4.353	5.093	4.057	—	7.483	8.150	7.068
	60%	—	15.220	16.962	14.160	—	19.827	20.528	17.290
	100%	—	24.409	26.408	20.948	—	27.591	28.463	26.990