Noname manuscript No.
(will be inserted by the editor)

Clustering Mixed-Type Data: a benchmark study on KAMILA and K-prototypes - APPENDIX

**Jarrett Jimeno · Madhumita Roy ·
Cristina Tortora**

Received: date / Accepted: date

This appendix contains the figures and the results of the paper: Jimeno J., Roy M., and Tortora C. (2020) Clustering Mixed-Type Data: a benchmark study on KAMILA and K-prototypes. In *Studies in Classification, Data Analysis, and Knowledge Organization.*

J. Jimeno • M. Roy • C. Tortora
Department of Mathematics and Statistics
San Jose State University
One Washington Square
San Jose, CA 95192
E-mail: jarrett.jimeno@sjsu.edu, madhumita.roy@sjsu.edu, cristina.tortora@sjsu.edu

.

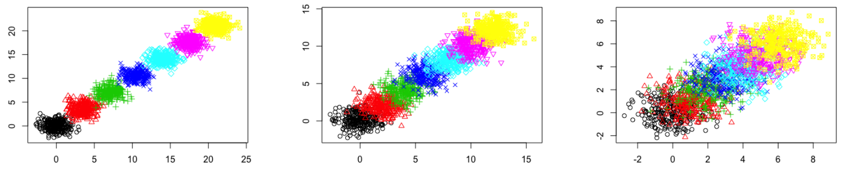

Fig. 1: *Two dimensional simulated data sets with 7 clusters and 30% (left), 60% (center), and 80% (right) overlap.*

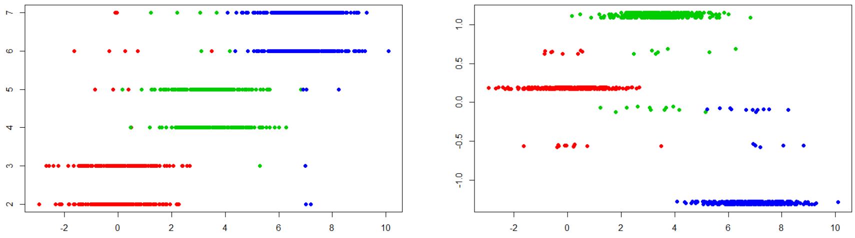

Fig. 2: *(left) First 2 dimensions of a 3 dimensional mixed-type data set with 1 numeric and 2 nominal variables, 3 clusters, 30% overlap, and 1000 observations. (right) The same data set with the nominal variables processed through MCA.*

Table 1: *Average and standard deviation of ARI of each clustering method against the true labels for 2 clusters.*

Number of Clusters: 2	Cluster Overlap	30%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		1.000	0.000	1.000	0.000	1.000	0.000
K-prototypes		1.000	0.000	1.000	0.000	1.000	0.000
K-means		1.000	0.000	1.000	0.000	1.000	0.000
C-means		1.000	0.000	1.000	0.000	1.000	0.000
PD		1.000	0.000	1.000	0.000	1.000	0.000
Student-*t*		1.000	0.000	1.000	0.000	1.000	0.000

Number of Clusters: 2	Cluster Overlap	60%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		1.000	0.000	1.000	0.000	1.000	0.000
K-prototypes		1.000	0.000	1.000	0.000	1.000	0.000
K-means		1.000	0.000	1.000	0.000	1.000	0.000
C-means		1.000	0.000	1.000	0.000	1.000	0.000
PD		1.000	0.000	1.000	0.000	1.000	0.000
Student-*t*		1.000	0.000	1.000	0.000	1.000	0.000

Number of Clusters: 2	Cluster Overlap	80%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		1.000	0.000	1.000	0.000	1.000	0.000
K-prototypes		0.994	0.003	1.000	0.000	1.000	0.000
K-means		0.221	0.438	0.222	0.441	0.222	0.441
C-means		0.661	0.495	0.666	0.500	0.667	0.500
PD		0.995	0.003	1.000	0.000	1.000	0.000
Student-*t*		0.998	0.002	1.000	0.000	1.000	0.000

Table 2: *Average and standard deviation of the number of algorithm iterations of each clustering method against the true labels for 2 clusters.*

Number of Clusters: 2	Cluster Overlap	30%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		3.040	0.052	3.025	0.026	3.020	0.026
K-prototypes		2.600	0.699	2.500	0.527	2.600	0.516
K-means		1.000	0.000	1.000	0.000	1.000	0.000
C-means		14.14	0.129	13.83	0.225	13.83	0.250
PD		8.800	0.422	8.000	0.000	8.000	0.000
Student-t		10.80	0.422	7.100	0.316	18.30	1.947

Number of Clusters: 2	Cluster Overlap	60%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		3.050	0.053	3.045	0.044	3.065	0.053
K-prototypes		2.600	0.699	2.600	0.699	2.400	0.516
K-means		1.000	0.000	1.000	0.000	1.000	0.000
C-means		24.14	0.296	24.22	0.317	23.96	0.277
PD		16.00	0.000	15.00	0.000	14.70	0.483
Student-t		4.000	0.000	4.000	0.000	4.000	0.000

Number of Clusters: 2	Cluster Overlap	80%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		3.505	0.132	3.465	0.178	3.385	0.120
K-prototypes		3.400	0.516	3.300	0.483	3.400	0.516
K-means		1.000	0.000	1.000	0.000	1.000	0.000
C-means		135.3	6.376	113.3	4.174	109.2	4.463
PD		13.20	0.422	14.00	0.000	15.00	0.000
Student-t		4.000	0.000	4.000	0.000	4.000	0.000

Table 3: *Average and standard deviation of ARI of each clustering method against the true labels for 5 clusters.*

Number of Clusters: 5	Cluster Overlap	30%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		1.000	0.000	1.000	0.000	1.000	0.000
K-prototypes		1.000	0.000	1.000	0.000	1.000	0.000
K-means		1.000	0.000	1.000	0.000	1.000	0.000
C-means		0.780	0.166	0.848	0.116	0.848	0.116
PD		0.850	0.178	0.889	0.134	0.896	0.125
Student-t		1.000	0.000	1.000	0.000	1.000	0.000

Number of Clusters: 5	Cluster Overlap	60%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		1.000	0.000	1.000	0.000	1.000	0.000
K-prototypes		1.000	0.000	1.000	0.000	1.000	0.000
K-means		1.000	0.000	1.000	0.000	1.000	0.000
C-means		0.960	0.119	1.000	0.000	1.000	0.000
PD		1.000	0.000	1.000	0.000	1.000	0.000
Student-t		1.000	0.000	1.000	0.000	1.000	0.000

Number of Clusters: 5	Cluster Overlap	80%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		1.000	0.000	1.000	0.000	1.000	0.000
K-prototypes		0.994	0.003	0.994	0.003	0.994	0.003
K-means		0.837	0.184	0.837	0.184	0.837	0.184
C-means		0.675	0.228	0.675	0.228	0.675	0.228
PD		0.809	0.228	0.809	0.228	0.809	0.228
Student-t		1.000	0.000	1.000	0.000	1.000	0.000

Table 4: *Average and standard deviation of the number of algorithm iterations of each clustering method against the true labels for 5 clusters.*

Number of Clusters: 5	Cluster Overlap	30%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		3.105	0.463	3.060	0.341	3.585	0.595
K-prototypes		4.200	3.225	3.700	1.567	3.200	0.632
K-means		1.900	0.316	1.900	0.316	1.700	0.483
C-means		22.95	2.223	23.42	3.016	22.97	3.088
PD		182.8	66.09	84.70	76.07	71.00	0.000
Student-t		9.600	1.506	10.30	1.160	6.700	0.483

Number of Clusters: 5	Cluster Overlap	60%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		3.905	0.227	3.780	0.441	4.375	0.459
K-prototypes		4.000	1.633	3.800	0.632	3.700	1.059
K-means		2.100	0.568	2.000	0.471	1.900	0.316
C-means		31.71	0.507	30.09	0.412	30.03	0.650
PD		79.50	33.78	136.7	42.61	145.0	25.28
Student-t		4.100	0.316	4.900	0.316	5.000	0.000

Number of Clusters: 5	Cluster Overlap	80%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		6.375	0.414	6.375	0.414	6.375	0.414
K-prototypes		6.700	6.273	6.700	6.273	6.700	6.273
K-means		2.800	0.422	2.800	0.422	2.800	0.422
C-means		2309	2994	2309	2994	2309	2994
PD		202.8	44.94	202.8	44.94	202.8	44.94
Student-t		4.000	0.000	4.000	0.000	4.000	0.000

Table 5: *Average and standard deviation of ARI of each clustering method against the true labels for 7 clusters.*

Number of Clusters: 7	Cluster Overlap	30%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		1.000	0.000	0.963	0.073	0.979	0.062
K-prototypes		1.000	0.000	0.979	0.063	0.979	0.064
K-means		1.000	0.000	1.000	0.000	1.000	0.000
C-means		0.858	0.114	0.850	0.119	0.839	0.135
PD		0.914	0.101	0.877	0.150	0.915	0.101
Student-t		1.000	0.000	0.980	0.060	1.000	0.000

Number of Clusters: 7	Cluster Overlap	60%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		1.000	0.000	0.941	0.089	0.917	0.098
K-prototypes		1.000	0.000	0.979	0.064	1.000	0.000
K-means		1.000	0.000	1.000	0.000	1.000	0.000
C-means		0.921	0.156	1.000	0.000	0.961	0.117
PD		0.960	0.119	1.000	0.000	1.000	0.000
Student-t		1.000	0.000	0.962	0.076	0.939	0.092

Number of Clusters: 7	Cluster Overlap	80%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		0.938	0.092	0.915	0.100	0.938	0.093
K-prototypes		0.993	0.004	1.000	0.000	1.000	0.003
K-means		0.837	0.184	0.773	0.147	0.758	0.148
C-means		0.687	0.054	0.708	0.073	0.676	0.021
PD		0.848	0.180	0.857	0.170	0.857	0.170
Student-t		1.000	0.000	0.957	0.084	0.980	0.060

Table 6: *Average and standard deviation of the number of algorithm iterations of each clustering method against the true labels for 7 clusters.*

Number of Clusters: 7	Cluster Overlap	30%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		2.880	0.400	2.600	0.378	2.805	0.372
K-prototypes		4.900	3.604	3.700	0.483	4.300	1.889
K-means		2.000	0.471	2.000	0.471	2.000	0.000
C-means		43.12	9.789	36.73	9.276	42.26	6.244
PD		48.60	8.708	51.70	5.926	56.80	11.24
Student-t		11.40	0.843	16.30	1.252	8.300	0.948

Number of Clusters: 7	Cluster Overlap	60%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		3.780	0.309	3.580	0.437	3.810	0.410
K-prototypes		4.000	1.414	3.700	0.823	3.800	0.633
K-means		2.000	0.000	2.000	0.000	1.900	0.316
C-means		36.74	2.827	34.82	2.201	34.61	4.128
PD		96.60	58.20	85.00	34.48	70.90	27.93
Student-t		4.900	0.316	5.000	0.000	5.000	0.000

Number of Clusters: 7	Cluster Overlap	80%					
	Variable Proportion continuous : nominal	1:3		1:1		3:1	
Clustering Method		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA		7.735	0.660	6.430	0.282	6.530	0.296
K-prototypes		5.200	0.155	4.400	0.699	4.500	1.000
K-means		2.900	0.422	2.600	0.843	2.200	0.422
C-means		5901.	1.252e+04	4433.	4236.	6676.	1262.
PD		127.8	24.08	145.5	24.26	122.0	9.787
Student-t		4.000	0.000	4.300	0.483	5.000	0.000

Table 7: *Average ARI with corresponding standard deviation on simulated data sets with correlated clusters, skewed clusters, and clusters with fewer observations.*

Fixed Parameters	60% Overlap, 3:1 Variable Ratio, 5 Clusters					
Variable Parameters	Correlated Continuous Variables		Skew t Continuous Variables		Small Sample Size	
Clustering Method	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA	0.828	0.012	0.283	0.011	1.000	0.000
K-prototypes	0.827	0.013	0.309	0.011	1.000	0.000
K-means	0.828	0.014	0.249	0.010	1.000	0.000
C-means	0.836	0.011	0.252	0.009	1.000	0.000
PD	0.460	0.021	0.266	0.005	0.940	0.133
Student's t	0.819	0.009	0.722	0.003	1.000	0.000

Table 8: *Average iteration count with corresponding standard deviation on simulated data sets with correlated clusters, skewed clusters, and clusters with fewer observations.*

Fixed Parameters	60% Overlap, 3:1 Variable Ratio, 5 Clusters					
Variable Parameters	Correlated Continuous Variables		Skew t Continuous Variables		Small Sample Size	
Clustering Method	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
KAMILA	10.43	0.863	21.50	1.428	3.895	0.283
K-prototypes	12.60	4.993	18.60	5.641	3.545	0.522
K-means	2.200	0.422	82.71	9.990	1.909	0.302
C-means	82.71	9.990	188.1	29.20	26.42	0.653
PD	41.00	4.570	67.80	7.315	169.6	89.08
Student’s t	51.60	3.062	79.60	35.58	4.714	0.665