Appendix - Clustering approaches for mixed-type data: A comparative review

Badih GHATTAS & Alvaro SANCHEZ SAN BENITO

December 2024

A Appendix

	$K=2$		$K=4$		$K=6$	
Methods	N=300	N=1200	N=300	N=1200	N=300	N=1200
KAMILA	0.037	0.026	0.156	0.152	0.130	0.125
K-prototypes	0.050	0.027	0.133	0.127	0.146	0.157
PDQ	0.005	0.004	0.009	0.001	0.018	0.007
Convex KM	0.050	0.036	0.146	0.137	0.142	0.144
MBN	0.144	0.295	0.223	0.223	0.092	0.132
LCM	0.045	0.031	0.171	0.145	0.168	0.170

Table 1: Mean ARI values for the BN classifier (M3) simulation model varying the number of clusters (K) and sample size (N). The dataset contains 3 continuous and 3 categorical variables.

	$K=2$		$K=4$		$K=6$	
Methods	N=300	N=1200	N=300	N=1200	N=300	N=1200
KAMILA	0.621	0.693	0.663	0.738	0.626	0.697
K-prototypes	0.561	0.568	0.522	0.530	0.407	0.367
PDQ	0.130	0.086	0.123	0.041	0.086	0.081
Convex KM	0.556	0.582	0.575	0.563	0.478	0.476
MBN	0.074	0.426	0.168	0.211	0.155	0.188
LCM	0.396	0.390	0.643	0.688	0.678	0.682

Table 2: Mean ARI values for the mixture of BN (M4) simulation model varying the number of clusters (K) and sample size (N). The dataset contains 3 continuous and 3 categorical variables.

	K=2											
	O = 30%						O = 60%					
	N=700			N=1400			N=700			N=1400		
Method/Cont.Prop	33%	50%	66%	33%	50%	66%	33%	50%	66%	33%	50%	66%
KAMILA	1.00	1.00	1.00	1.00	1.00	1.00	0.999	0.998	0.998	0.998	0.998	0.999
K-prototypes	1.00	1.00	1.00	1.00	1.00	1.00	0.980	0.994	0.997	0.986	0.993	0.994
PDQ	0.996	0.996	0.996	0.992	0.992	0.994	0.779	0.705	0.703	0.745	0.697	0.684
Convex KM	1.00	1.00	1.00	1.00	1.00	1.00	0.994	0.998	0.998	0.995	0.998	0.998
MBN	0.253	0.130	0.320	0.624	0.621	0.517	0.168	0.126	0.135	0.216	0.308	0.311
LCM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	K=5											
KAMILA	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.997	0.998	0.996	0.998	0.999
K-prototypes	1.000	1.000	1.000	1.000	1.000	1.000	0.973	0.987	0.990	0.975	0.978	0.992
PDQ	1.00	1.00	1.00	0.999	1.00	1.00	0.926	0.946	0.963	0.904	0.938	0.961
Convex KM	1.000	1.000	1.000	1.000	1.000	1.000	0.993	0.998	0.997	0.991	0.996	0.998
MBN	0.087	0.061	0.058	0.042	0.059	0.075	0.962	0.942	0.893	0.982	0.980	0.955
LCM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	K=10											
KAMILA	0.977	0.910	0.913	0.936	0.898	0.949	0.895	0.876	0.913	0.897	0.955	0.970
K-prototypes	0.915	0.928	0.899	0.914	0.900	0.873	0.866	0.808	0.836	0.838	0.828	0.841
PDQ	0.981	1.00	1.00	0.995	1.00	1.00	0.636	0.680	0.816	0.593	0.685	0.792
Convex KM	0.961	0.960	0.947	0.986	0.961	0.934	0.955	0.960	0.993	0.951	0.971	0.962
MBN	0.628	0.644	0.629	0.762	0.762	0.734	0.287	0.250	0.173	0.516	0.452	0.377
LCM	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.98	0.99	0.99	1.00	1.00

Table 3: Mean ARI values from the multivariate Gaussian simulation model (M1) varying the number of clusters (K), sample size (N), cluster overlap (0) and proportion of continuous variables in the dataset. The dimension is equal to 12.