Frontiers of Computer Science

Heterogeneous Clustering via Adversarial Deep Bayesian Generative Model

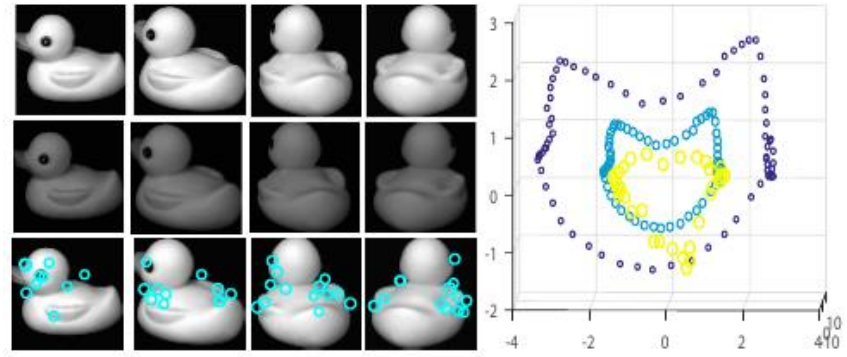
Xulun YE, Jieyu ZHAO

Frontiers of Computer Science, DOI: 10.1007/s11704-022-1376-2

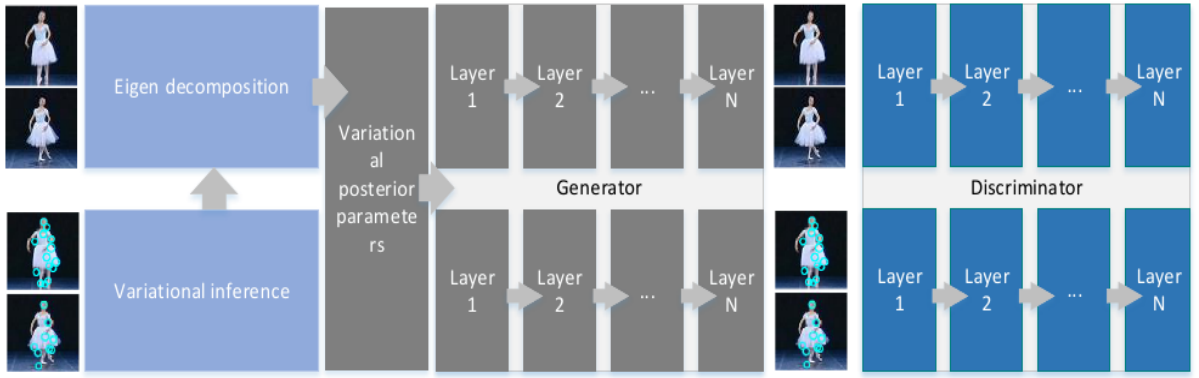
Problem

- 1) Heterogeneous features: A sample with different features may act like different clusters in the feature space;
- 2) Unknown cluster number: Many DNN-based methods require a predefined cluster number which is usually not available in many real applications.

Idea



1) Data with different feature spaces shares a same data relationship structure. We then extend this idea to measure the similarity between different features;



2) Dirichlet process with a feature metric-restricted hierarchical sample generation process can be used to estimate the cluster number.

Experiments

We validate our model on both the synthetic dataset and the real dataset.

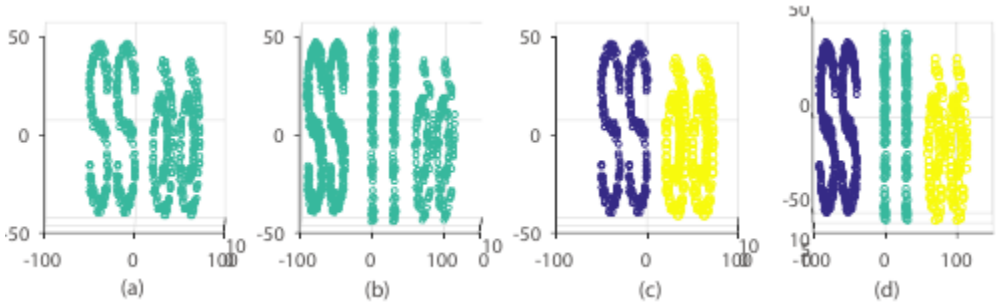


Illustration of BHAC clustering result on the synthetic dataset. a) and b) demonstrate the original dataset without class labels. c) and d) are the BHAC results.

Method	COIL20	COIL100	BALLET	USPS	MNIST	COIL20	COIL100	BALLET	USPS	MNIST
	Original Image+Noise Rotated Image					Original Image+SIFT+TNT				
BHAC	**0.51**	**0.54**	**0.45**	**0.49**	0.47	**0.61**	**0.55**	0.23	**0.41**	**0.41**
DP-space	0.07	0.02	0.07	0.02	0.01	0.02	0.02	0.01	0.05	0.02
SCAMS	0.26	0.24	0.15	0.07	0.24	0.08	0.03	0.04	0.05	0.11
AutoSC-N	0.22	0.18	0.17	0.18	0.21	0.01	0.12	0.13	0.09	0.13
CFSFDP	0.17	0.21	0.29	0.21	0.34	0.36	0.17	0.19	0.21	0.21
DPM	0.05	0.12	0.03	0.03	0.26	0.06	0.04	0.03	0.07	0.11
GFMM	0.01	0.06	0.04	0.08	0.04	0.04	0.02	0.02	0.06	0.03
BLRASC	0.41	0.03	0.11	0.30	0.37	0.29	0.22	0.27	0.24	0.03
ACIDS	0.32	0.39	0.26	0.32	0.42	0.33	0.13	**0.39**	0.10	0.17
ClusterGAN	0.27	0.17	0.04	0.23	0.47	0.11	0.13	0.01	0.14	0.04
Spectral-Net	0.29	0.32	0.35	0.28	0.39	0.34	0.17	0.36	0.22	0.21
	Original Image+Noise Rotated Image+DAMA					Original Image+SIFT+CDLS				
DP-space	0.04	0.12	0.02	0.07	0.03	0.06	0.01	0.02	0.02	0.04
SCAMS	0.26	0.25	0.14	0.21	0.27	0.11	0.02	0.02	0.01	0.02
AutoSC-N	0.31	0.37	0.11	0.23	0.31	0.31	0.07	0.06	0.14	0.12
CFSFDP	0.21	0.24	0.17	0.31	0.37	0.39	0.12	0.27	0.21	0.17
DPM	0.17	0.09	0.04	0.06	0.31	0.31	0.02	0.03	0.05	0.13
GFMM	0.12	0.16	0.02	0.06	0.02	0.03	0.01	0.01	0.01	0.04
BLRASC	0.39	0.42	0.27	0.34	0.44	0.37	0.21	0.11	0.17	0.12
ACIDS	0.41	0.44	0.27	0.27	0.51	0.42	0.11	0.36	0.13	0.14
ClusterGAN	0.36	0.32	0.05	0.38	**0.58**	0.07	0.04	0.07	0.08	0.02
Spectral-Net	0.31	0.40	0.22	0.35	0.54	0.34	0.09	0.29	0.12	0.17
Ground truth	20.0	100.0	44.0	10.0	10.0	20.0	100.0	44.0	10.0	10.0
Estimated Number	31.2	127.6	60.4	20.2	21.2	25.1	125.2	7.1	30.1	17.1

Clustering accuracy (NMI) and the estimated cluster number on the real world datasets.