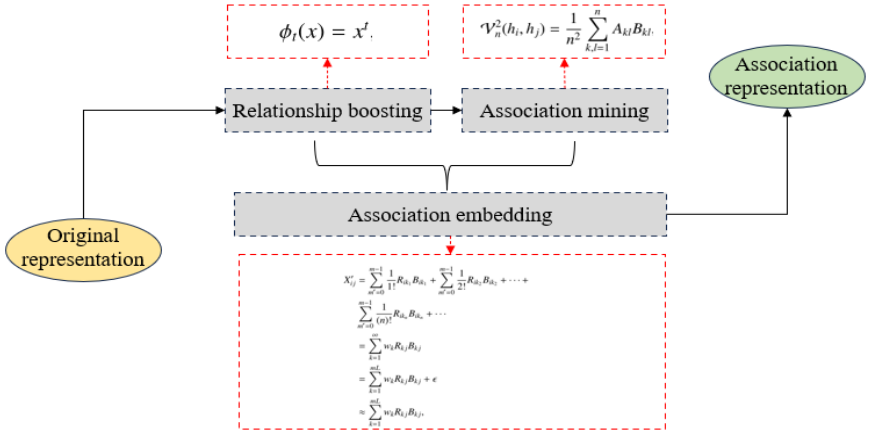
A Data Representation Method Using Distance Correlation

Xinyan LIANG , Yuhua QIAN, Qian GUO , Keyin ZHENG

Frontiers of Computer Science, DOI: 10.1007/s11704-023-3396-y

Problems & Ideas

- Problems of association–based representation learning:
  - Data representation obtained by association–based fusion (AF) is highly dimensional or sparse.
  - Pearson's correlation coefficient (pCor) used to capture the association between features by AF does not well balance effectiveness and efficiency.
- Ideas: Jointly both distance correlation and PCA techniques can well balance the effectiveness and efficiency for association-based representation learning.



Data	pCor	dCor	NMI	MIC	MIC_e
Iris	0.16	0.05	1.01	0.30	0.29
oocMer4D	0.36	5.90	75.37	71.23	70.37
Contrac	0.22	0.55	3.84	10.05	7.16
Abalone	0.26	1.24	3.71	10.82	11.08
Magic	1.02	7.50	7.07	58.84	59.11
AWArgsift-hist	178.90	618.08	3124.94	9956.50	15541.76

Left: Overall framework of AssoRep. Right: Computation time (s) of the different association mining methods.

Main Contributions

- Contributions:
  - We introduce a fresh perspective on data representation improvement through association between features, which perfects the relationship-based learning that mainly focuses on relationships among samples such as graph neural network and spectral clustering.
  - A novel distance correlation-based representation learning method is proposed, and it well balances effectiveness and efficiency compared to its counterpart AF.
  - The experimental results on 120 benchmark show that the proposed AssoRep outperforms the other methods in term of five popular evaluation metrics widely used for classification in most cases.

Data	Benchmark	AF	AF_X	$CRAM_c$	$CRAM_d$	FS_{MI}	FS_{LR}	AssoRep
Iris	0.907±0.053	0.927±0.055	0.953±0.043	0.953±0.043	0.953±0.043	0.947±0.050	0.940±0.055	**0.973±0.033**
oocMer4D	0.796±0.036	0.751±0.023	0.811±0.035	0.820±0.028	0.822±0.028	0.797±0.028	0.800±0.035	**0.837±0.020**
Contrac	0.507±0.042	**0.568±0.052**	0.566±0.058	0.519±0.035	0.517±0.035	0.507±0.030	0.519±0.041	**0.568±0.054**
Abalone	0.647±0.020	0.640±0.023	0.659±0.022	0.650±0.015	0.651±0.017	0.647±0.019	0.635±0.012	**0.662±0.021**
Magic	0.791±0.006	0.837±0.007	0.844±0.008	0.845±0.008	0.844±0.008	0.791±0.007	0.787±0.008	**0.850±0.008**
Annealing	0.873±0.027	0.893±0.017	0.910±0.024	0.910±0.024	0.911±0.021	0.880±0.024	0.863±0.017	**0.951±0.014**
ctg-10classes	0.768±0.032	0.802±0.030	0.800±0.026	0.817±0.023	0.813±0.023	0.771±0.027	0.751±0.030	**0.834±0.027**
oocTris2F	0.797±0.030	0.815±0.036	0.815±0.031	0.828±0.043	0.829±0.040	0.795±0.031	0.785±0.031	**0.836±0.030**
Mean values	0.7608(↑5.31%)	0.7791(↑3.48%)	0.7948(↑1.91%)	0.7927(↑2.012%)	0.7925↑2.14%)	0.7669(↑4.70%)	0.7600(↑5.39%)	**0.8139**
Avg. rank	6.813	5.250	3.563	3.125	3.063	6.188	6.938	1.063

Accuracy comparison between AssoRep with other feature enhancement methods