D2R2: Diffusion-based Representation with Random Distance Matching for Tabular Few-shot Learning

Ruoxue Liu*, Linjiajie Fang*, Wenjia Wang †, Bing-Yi Jing †

The Hong Kong University of Science and Technology
The Hong Kong University of Science and Technology(Guangzhou)
Southern University of Science and Technology

*Equal contribution. †Corresponding authors

Motivation

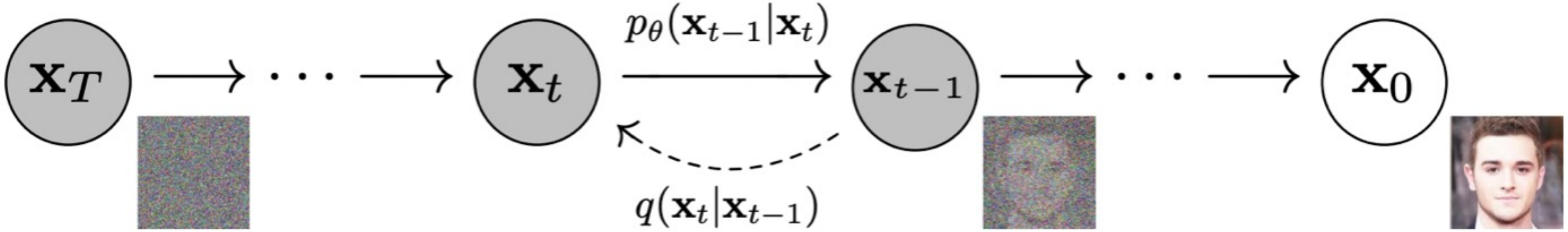
Challenges

- Scarcity of labeled training samples.
- Diverse feature types: numerical & categorical.
- Prototype classification: high variance & multi-modal samples.

Motivations

- Diffusion models are remarkable generators. But can we leverage diffusion models to extract semantic representations for tackling few-shot learning in complex tabular data?

Diffusion Process



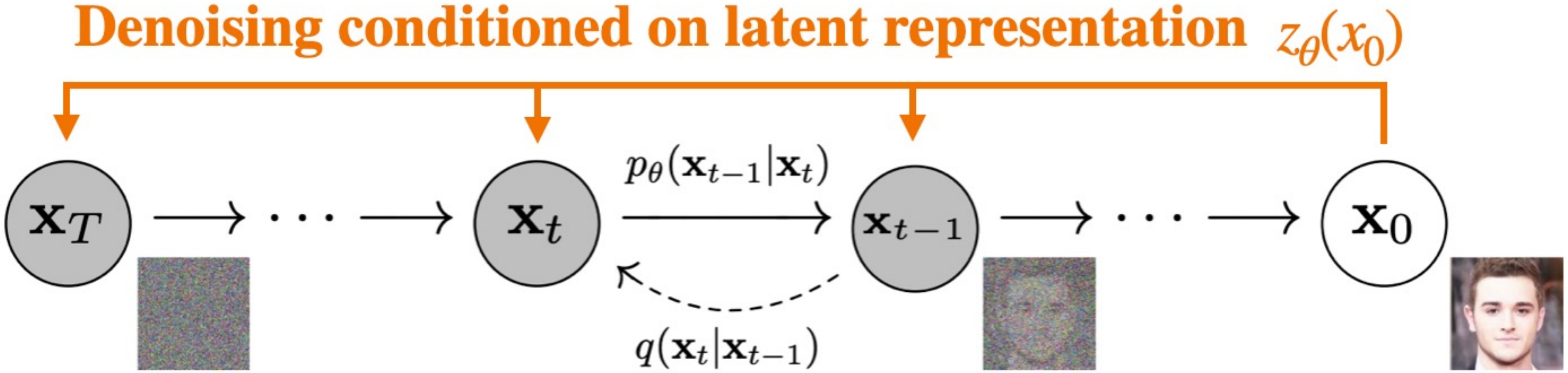
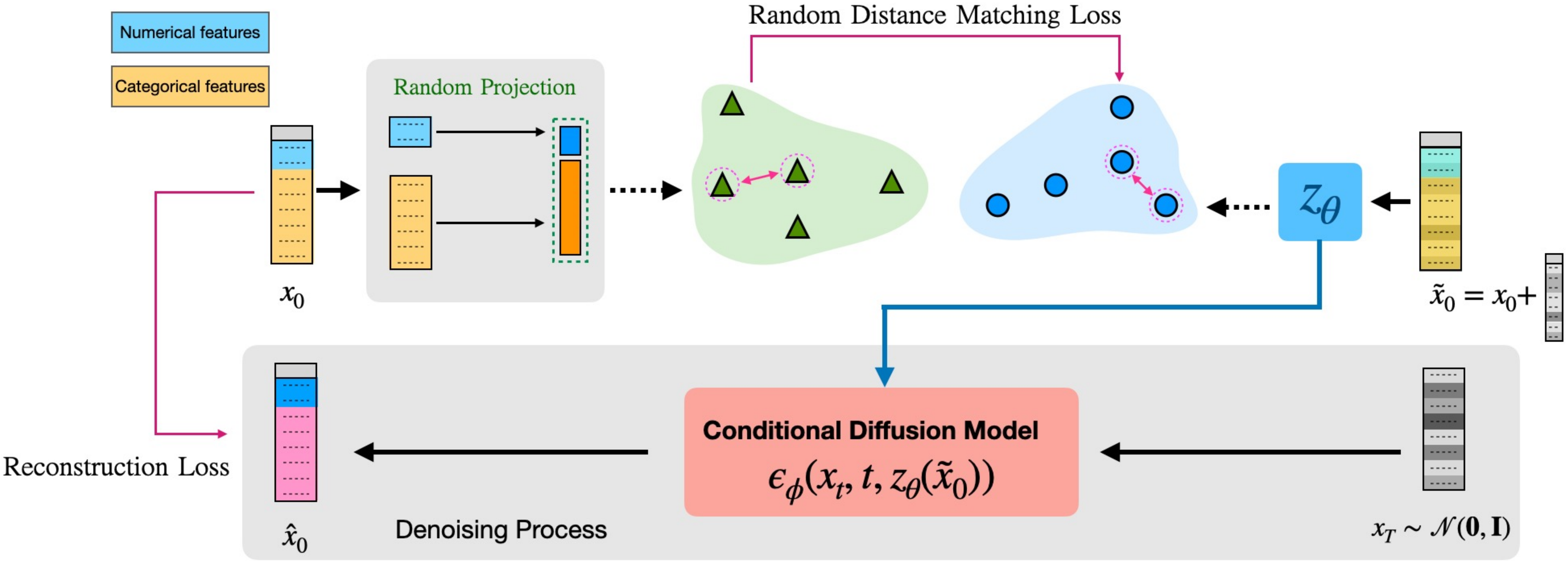


Image source: Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

Our contributions

- We are the first to propose a specifically designed diffusion method to learn semantic knowledge for tabular data.
- We propose D2R2, which adapts well to mixture types of tabular data using diffusion models and random distance matching.
- We introduce a novel classifier with instance-wise iteration prototypes, which is able to construct highly accurate and stable prototypes.
- D2R2 outperforms baselines by a significant margin.

D2R2: Overall Framework



$$\mathcal{L}_{D2R2}(\phi, \theta) = \mathcal{L}_{recon}(\phi, \theta) + \alpha \cdot \mathcal{L}_{rdm}(\theta)$$

Diffusion-based representation learning

$$\mathcal{L}_{recon}(\phi,\theta) = \mathbf{E}_{t,\mathbf{x}_0,\mathbf{x}_t} || \epsilon_\phi(\mathbf{x}_t, t, z_\theta(\tilde{\mathbf{x}}_0)) - \epsilon ||_2^2$$

Learnable embedding

Timestep

Noisy samples

$$x_0 \rightarrow z_\theta(x_0) \rightarrow \hat{x}_0$$

Information bottleneck

Random Distance Matching

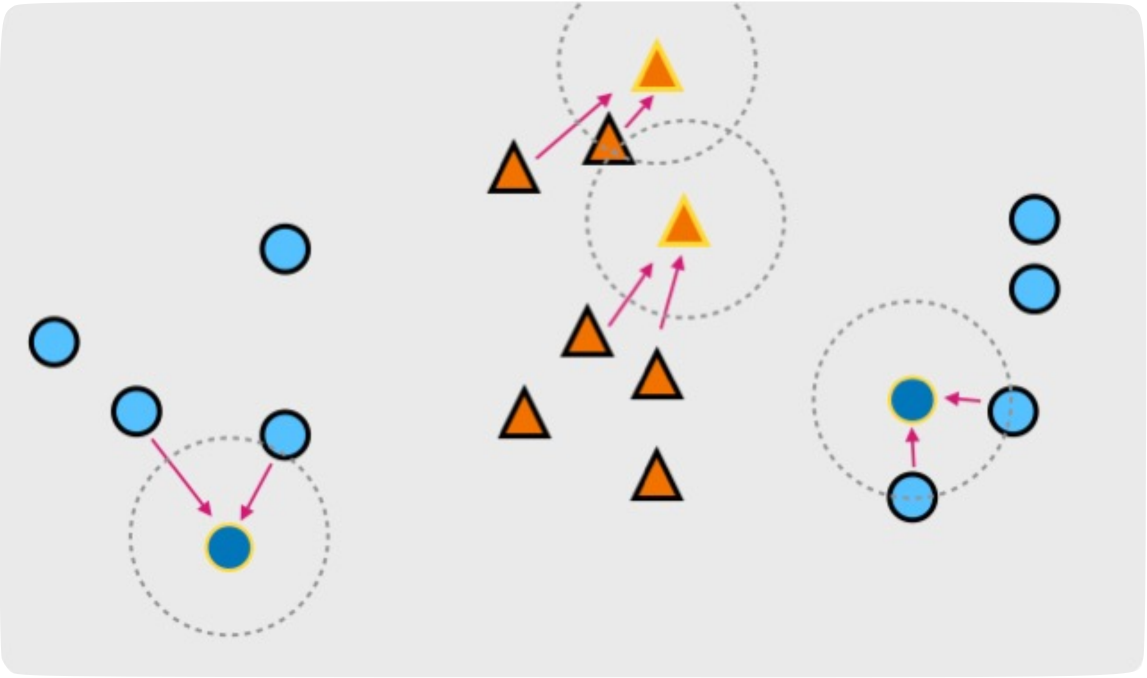
$$\mathcal{L}_{rdm}(\theta) = \mathbf{E}_{\mathbf{x}_0, \mathbf{x}_0', W} ||d(z_\theta(\mathbf{x}_0), z_\theta(\mathbf{x}_0')) - d(W\mathbf{x}_0, W\mathbf{x}_0')||^2$$

- Incorporate distance information for downstream classification tasks, which is missing from the diffusion representation learning.

- Handle mixed feature types through different projection distributions.

$$W\mathbf{x}_0 := concat(W_{num}[\mathbf{x}_0]_{num}, W_{cat}[\mathbf{x}_0]_{cat})$$

Instance-wise Prototype Iteration



- Initialize prototypes as the embedding of labeled instances.
- Run soft K-means iteration on instances.
- Predict the label by the nearest prototype.

Experiments & Results

Method	cmc	diabetes	dna	income	karkunen	optdigits	pixel	nomao	brest
				#shot=1					
CatBoost	36.03	56.74	39.15	57.55	53.24	58.30	54.74	63.62	69.71
TabPFN	35.37	53.35	-	-	46.02	55.74	-	-	-
KNN	35.39	58.50	42.20	51.45	54.61	65.60	60.79	63.51	71.87
Mean Teacher(*)	35.58	58.05	46.58	60.63	54.57	66.10	61.02	64.23	71.92
ICT(*)	36.53	58.08	46.55	61.83	58.37	69.12	60.88	-	-
Pseudo-Label(*)	34.97	57.03	44.26	60.52	49.44	61.50	56.12	62.39	69.92
MPL(*)	35.13	57.39	44.22	60.85	47.66	61.52	56.01	64.28	71.33
SubTab	36.23	58.22	46.98	62.45	50.22	62.01	60.34	67.63	72.94
VIME	35.90	58.99	51.23	61.82	59.81	69.26	63.28	64.75	70.11
SCARF	35.39	55.64	57.86	57.94	60.96	63.31	63.93	68.90	75.32
RTDL	34.34	58.15	47.99	53.61	58.25	62.78	62.87	68.33	76.38
UMTRA(*)	35.46	57.64	25.13	57.23	49.05	49.87	34.26	-	-
SES(*)	34.59	59.97	39.56	56.39	49.19	56.30	49.19	69.52	74.89
CACTUs(*)	36.10	58.92	65.93	64.02	65.59	71.98	67.61	71.49	75.24
STUNT(*)	37.10	61.08	66.20	63.52	71.20	76.94	79.05	71.54	76.92
D2R2	**42.88**	**63.94**	**68.00**	**75.82**	**72.08**	**81.13**	**81.34**	**79.47**	**77.69**

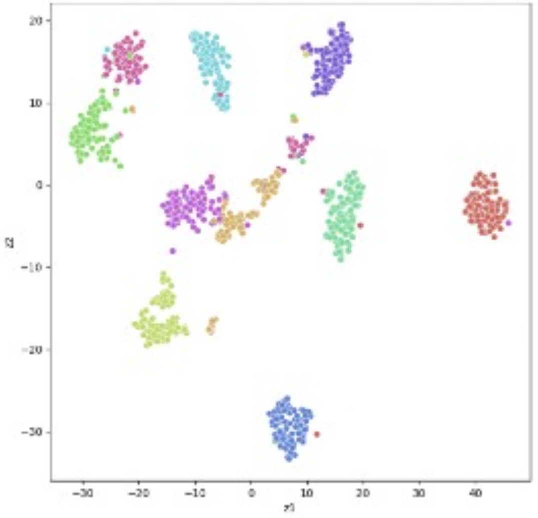
Experiments & Results

Method	cmc	diabetes	dna	income	karkunen	optdigits	pixel	nomao	brest
					#shot=5				
CatBoost	39.89	64.51	60.20	67.99	77.94	83.07	83.38	75.32	77.06
TabPFN	38.31	64.06	-	-	76.59	81.68	-	-	-
KNN	37.65	65.61	61.16	62.19	80.08	84.16	84.75	73.78	79.43
Mean Teacher(*)	37.73	65.45	61.47	67.05	81.08	86.66	85.24	74.78	81.26
ICT(*)	38.09	65.47	63.37	70.13	84.58	87.01	86.12	-	-
Pseudo-Label(*)	37.49	64.46	60.06	66.26	78.60	83.71	82.94	72.87	78.91
MPL(*)	37.47	64.51	59.65	67.61	77.85	83.70	82.39	73.20	79.54
SubTab	39.81	68.26	62.49	72.14	70.88	83.27	80.41	76.15	82.74
VIME	39.83	67.64	71.29	72.19	19.42	83.21	85.24	74.96	85.81
SCARF	37.75	68.66	62.75	66.09	69.96	85.67	81.32	77.65	84.42
RTDL	37.59	64.27	45.49	64.92	60.43	82.58	76.13	73.61	79.66
UMTRA(*)	38.05	64.41	25.08	65.78	67.28	73.29	51.32	-	-
SES(*)	39.04	66.61	52.25	68.27	74.80	78.46	74.80	76.50	84.73
CACTUs(*)	38.81	66.79	81.52	72.03	82.20	85.92	85.25	78.33	86.90
STUNT(*)	40.40	69.88	79.18	72.69	**85.45**	88.42	89.08	81.49	86.82
D2R2	**43.39**	**73.52**	**82.38**	**76.02**	84.96	**90.73**	**91.06**	**82.69**	**88.27**

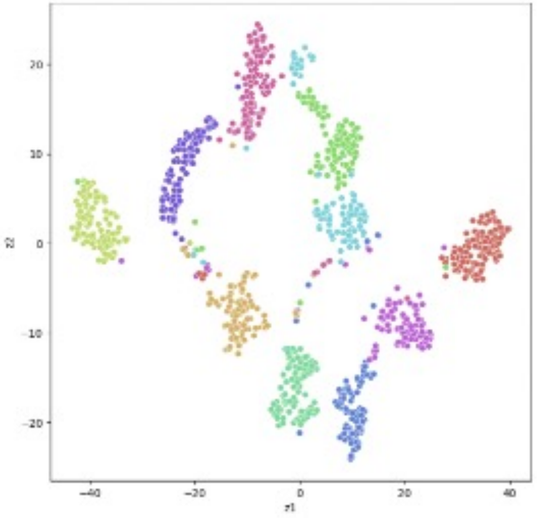
Ablations

Dataset	RDM	DR	DR+RDM	RDM+IP	DR+IP	D2R2-c	D2R2
opt (1-shot)	49.05	72.38	77.41	26.66	76.87	-	**81.13**
dna (1-shot)	45.43	57.14	61.29	26.03	56.17	-	**68.00**
cmc (1-shot)	35.50	35.19	38.69	34.90	34.48	-	**42.88**
opt (5-shot)	70.26	88.64	89.61	51.31	89.32	87.12	**90.73**
dna (5-shot)	47.72	71.84	73.03	32.16	79.24	81.39	**82.38**
cmc (5-shot)	36.04	35.62	40.81	36.27	35.74	43.39	**43.39**

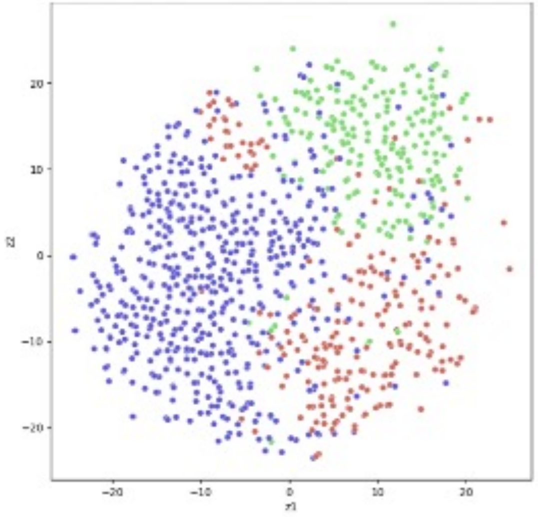
Visualization of learned embeddings



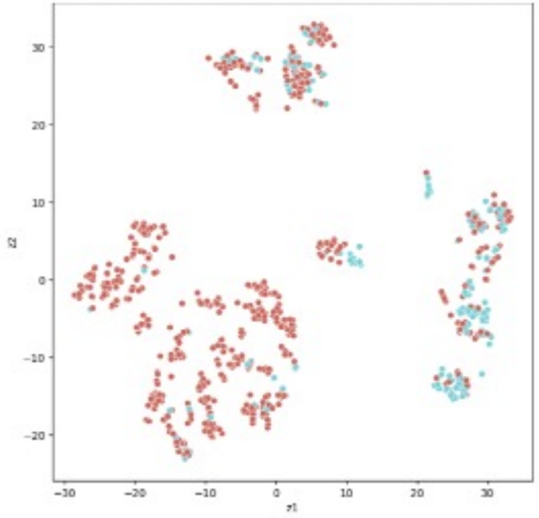
Optdigits



Karkunen



Dna



Income

- The embeddings exhibit multimodal patterns. In the dna dataset, the red class is distributed in both the bottom right corner and the top left corner.

NEURAL INFORMATION PROCESSING SYSTEMS

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Thanks!