A Survey of Task-driven Heterogeneous Feature Embedding and Selection

Hedong YAN,
Computer Science,
Hong Kong Baptist University
Supervisor: Yiu-ming Cheung

- Background
- Related Works
  - Feature Embedding
  - Feature Selection
- Methodology
- Futural Plan

Background

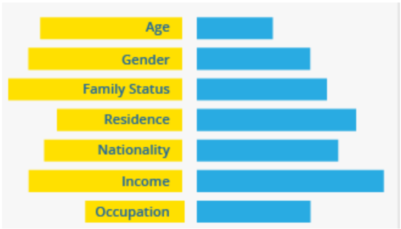
- Heterogeneous data widely exists in reality, such as user information, EHR, and surveys.
- Heterogeneous data is critical for many tasks in the real world.

CTR

vs.

Conversion

CTR and Conversion Rate Prediction

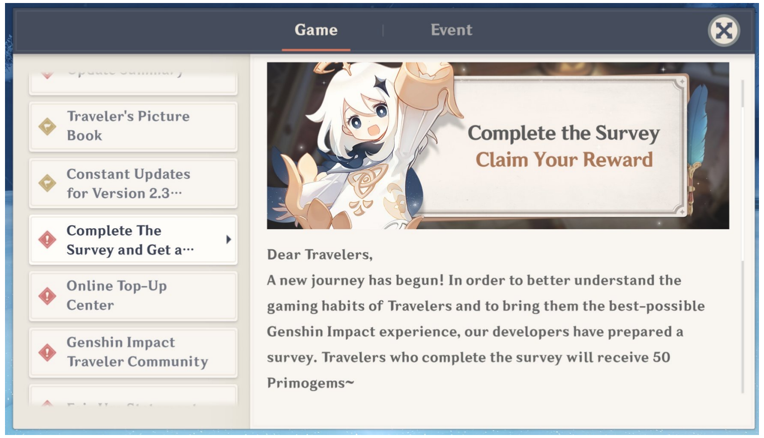


Demographic

Laboratory

Vital Sign

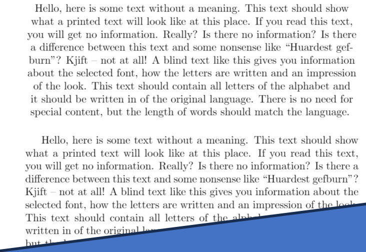
Disease Progression Prediction



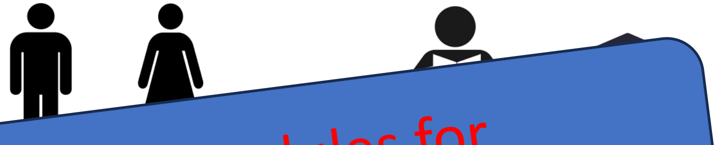
Survey Analysis

Problem

- Traditional deep learning models for homogeneous features can not be directly applied to heterogeneous data. Not much attention has been paid to describing how DNN can be designed for heterogeneous datasets.


Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original



Research Question 1: How to design embedding modules for heterogeneous dataset?

Video

Voice

Homogeneous Features

Interval

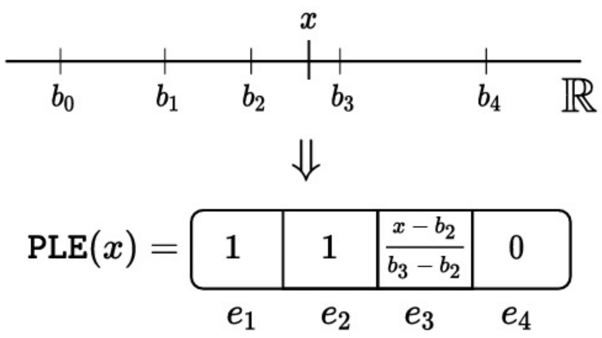
Ratio

Heterogeneous Features

Related works

Feature Scale	Encoder	Input	Output
Nominal	One-hot	[1,2,3]	[[1,0,0],[0,1,0],[0,0,1]]
	Binary	[1,2,3]	[[0,0],[0,1],[1,0]]
	Dumpy	[1,2,3]	[[1,0],[0,1],[0,0]]
	Count	[1,1,3]	[[2],[2],[1]]
	Simple	[1,2,3]	$[[\frac{2}{3},-\frac{1}{3},-\frac{1}{3}],[-\frac{1}{3},\frac{2}{3},-\frac{1}{3}],[-\frac{1}{3},-\frac{1}{3},\frac{2}{3}]]$
Ordinal	Ordinal	[1,2,3]	[1,2,3]
	Rank-hot	[1,2,3]	[[1,0,0],[1,1,0],[1,1,1]]
	Gray	[1,2,3]	[[0,0],[0,1],[1,1]]
Continuous	Bins + One-hot	[0.11,0.27,0.34]	[[1,0,0],[0,1,0],[0,0,1]]
	Piece-wise linear [1]	[0.11,0.27,0.34]	[[0.1,0,0],[1,**0.2**,0],[1,1,0.1]]

- Can we use existing encoders to transform the heterogeneous feature into homogeneous features?



[1] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, *34*, 18932-18943.

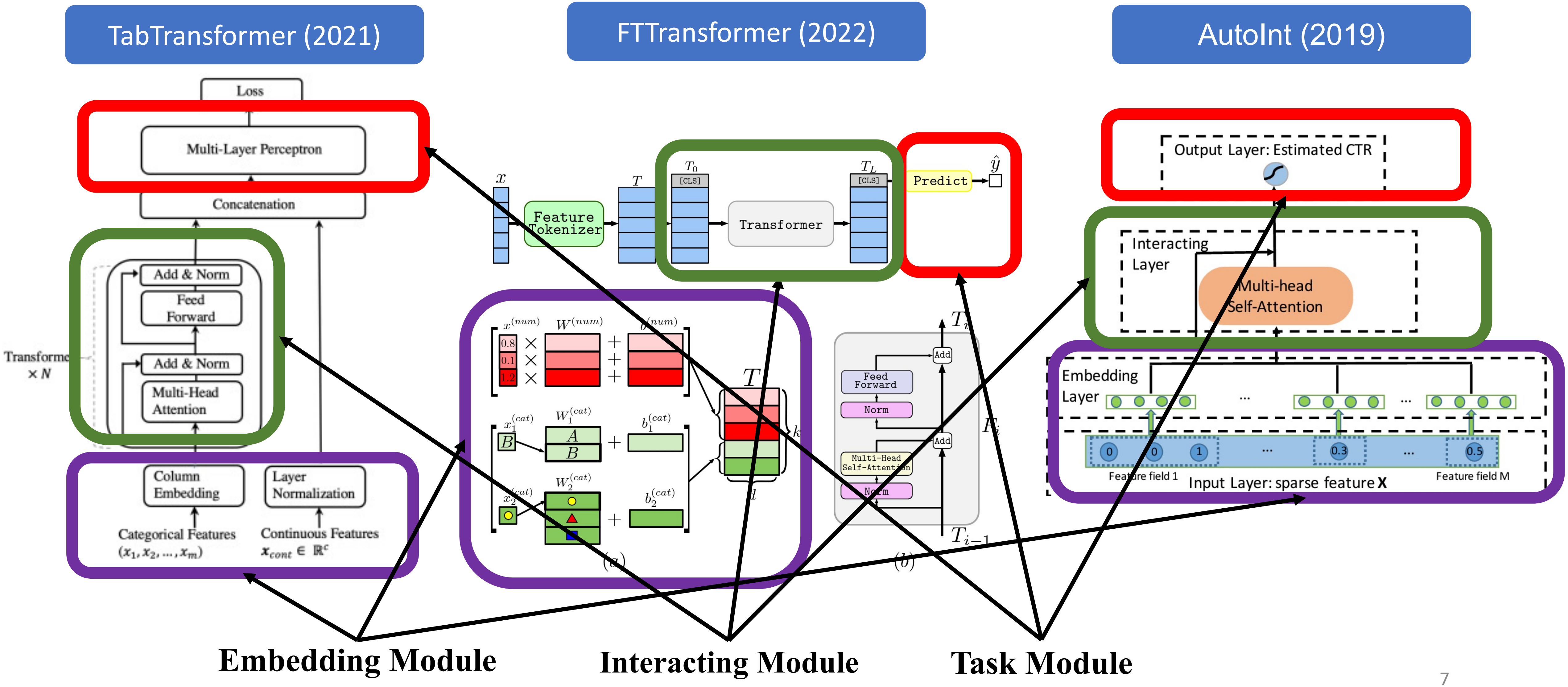
Related works

Heterogeneous embedding for different models

- Transformer-based model
  - TabTransformer, FTTransformer, AutoInt, ILEAHE
- MLP-based model
  - DeepFM, DANETs, DVN v2
- Diffusion-based model
  - TabDDPM
- Graph-based model
  - T2G-Former

Related works

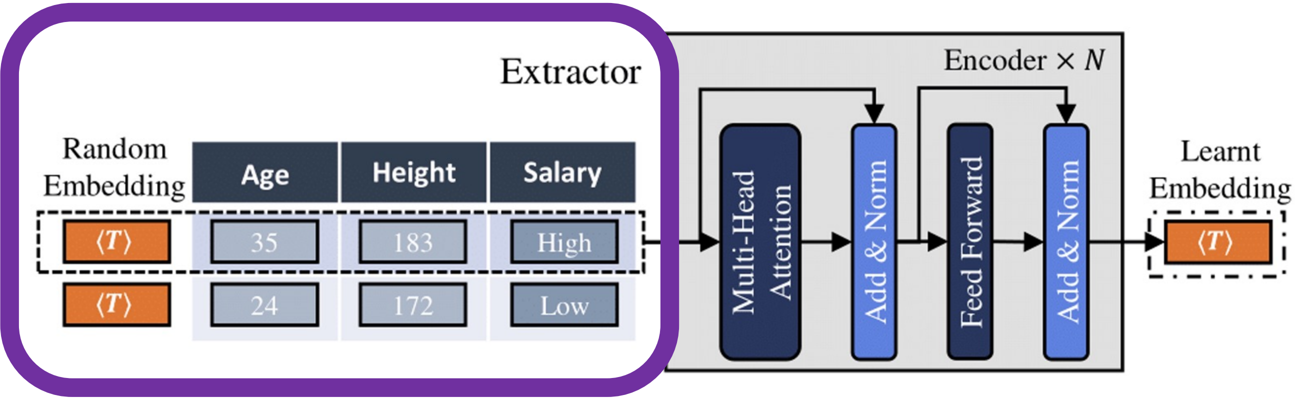
Transformer-based model



Related works

Transformer-based model

ILEAHE (2023)



Categorical: Dictionary embedding

Numerical: 2-layer perceptron

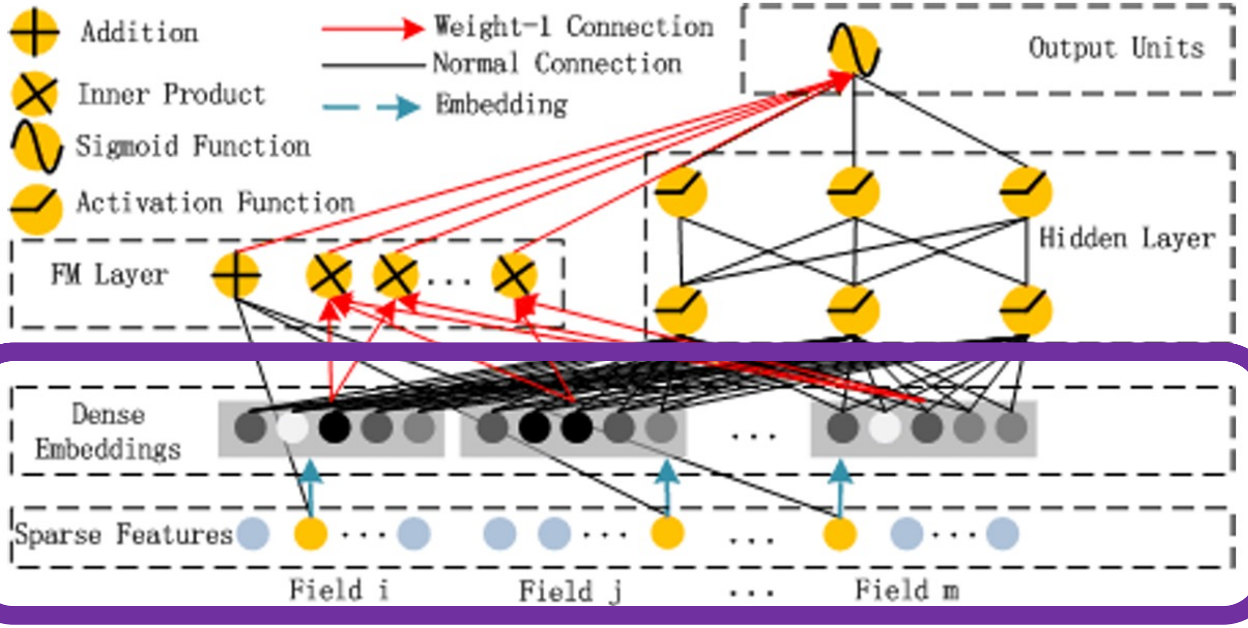
Heterogeneous Embedding Modules

- TabTransformer
  - Categorical: Dictionary embedding
  - Continuous: None
- FTTransformer
  - Categorical: Dictionary embedding
  - Continuous: Linear
- AutoInt
  - Categorical: One-hot + linear
  - Continuous: Linear
- ILEAHE
  - Categorical: Dictionary embedding
  - Continuous: 2-layer perceptron

Related works

MLP-based model

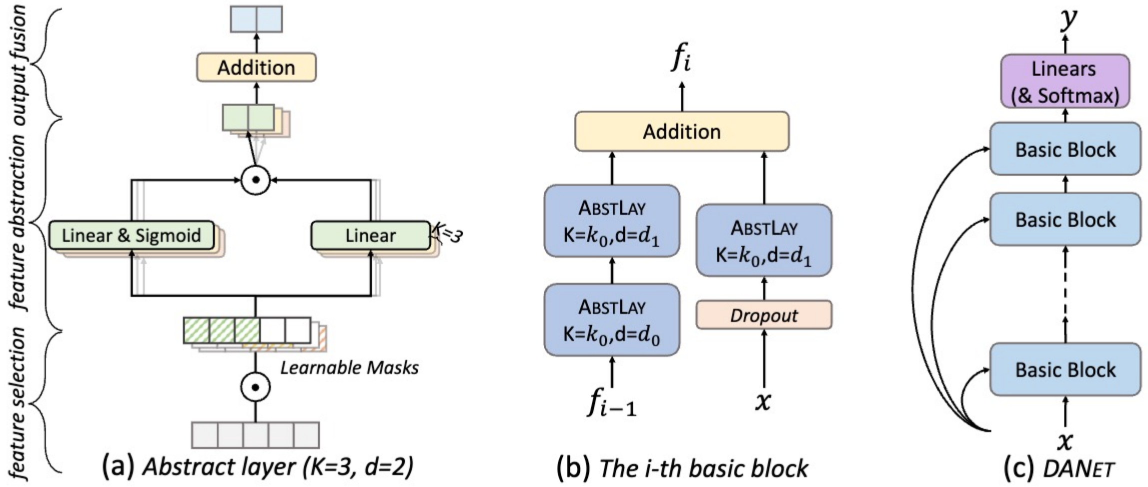
DeepFM (2017)



Categorical: One-hot encoder + linear

Numerical: Linear

DANETs (2017)



$$x_k^i = \frac{\sum_{j=1}^{n} \mathbb{I}_{x_j^i = x_k^i} * y_j + ap}{\sum_{j=1}^{n} \mathbb{I}_{x_j^i = x_k^i} + a}$$

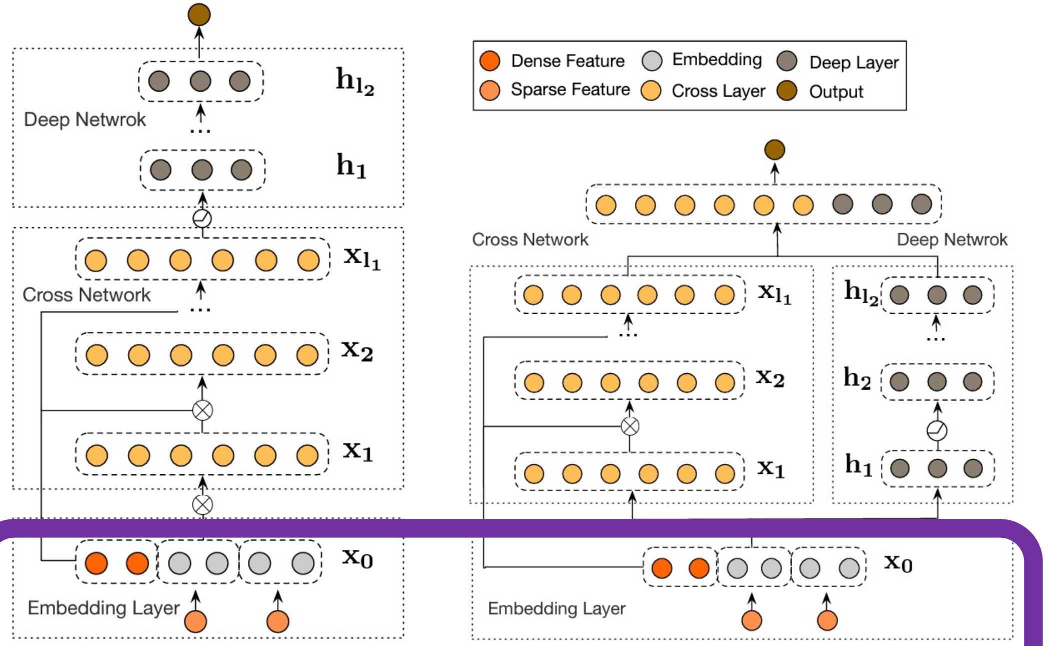
Categorical: Target Statistic

Numerical: None

Related works

MLP-based model

DVN v2 (2020)

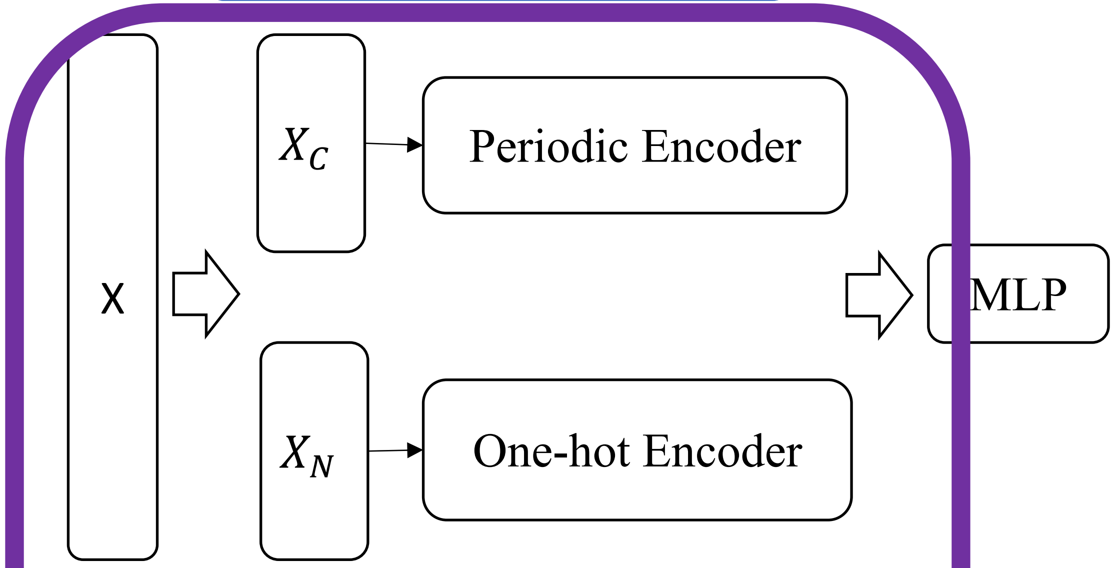


(a) Stacked **(b) Parallel**

Categorical: Dictionary embedding

Numerical: None

Gorishniy (2022)



$$f_i(x) = \texttt{Periodic}(x) = \texttt{concat}[\sin(v),\ \cos(v)], \quad v = [2\pi c_1 x,\ \ldots,\ 2\pi c_k x] \tag{8}$$

where c_i are trainable parameters initialized from $\mathcal{N}(0, \sigma)$. σ is an important hyperparameter that is tuned using validation sets.

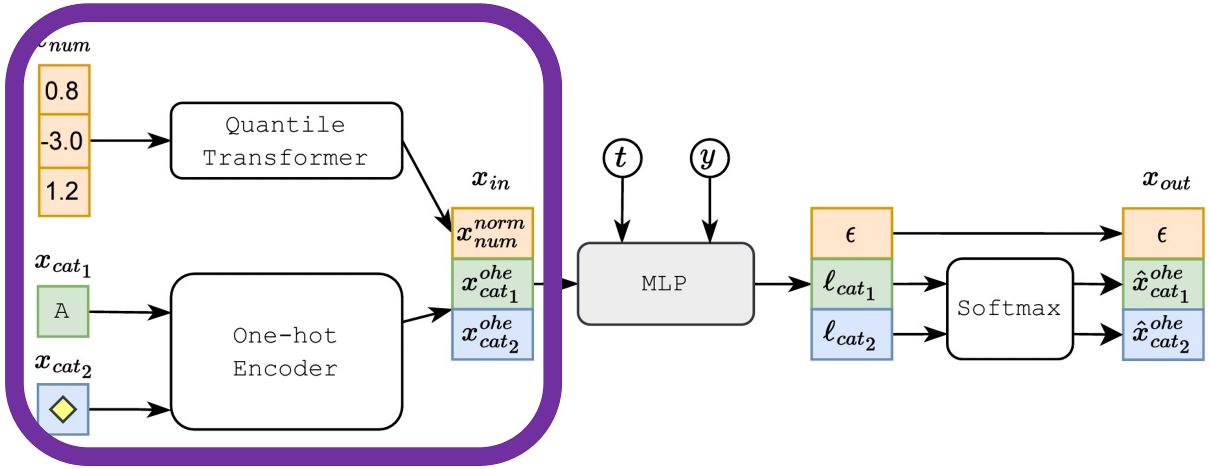
Categorical: One-hot

Numerical: Periodic encoder

Related works

Diffusion-based model

TabDDPM (2023)



- Use diffusion procedure to optimize the parameters
- Categorical: One-hot
- Numerical: Quantile Gaussian Normalization

Graph-based

T2G-Former (2023)

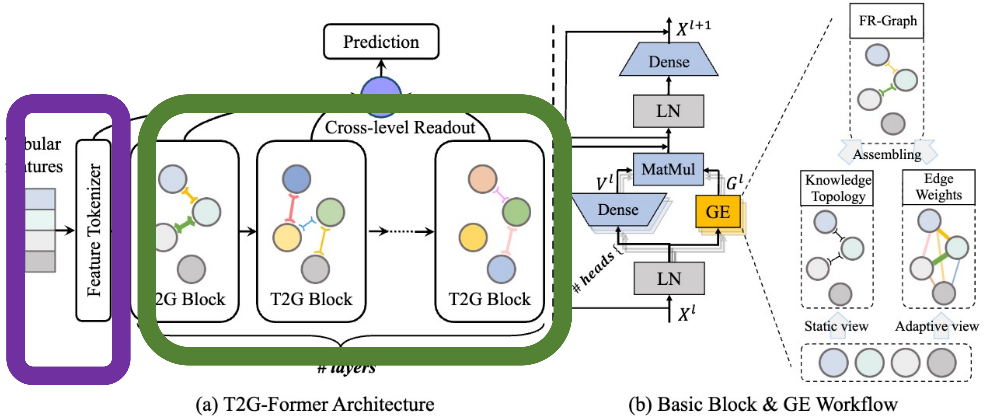


Figure 2: (a) The architecture of T2G-FORMER for tabular learning. Each T2G block builds an FR-Graph for a feature level and performs selective interaction. A global readout node collects salient features from each layer to form tabular semantics. (b) Illustrating a basic block in Sec. and GE in Sec. .

- Add graph blocks to model the features' interaction
- Categorical: Dictionary embedding
- Numerical: Linear

Related works

Section Conclusion

	GE ↑	CH ↑	CA ↓	HO ↓	AD ↑	OT ↑	HI ↑	FB ↓	SA ↑	CO ↑	MI ↓	Avg. Rank
CatBoost	0.692	0.861	0.430	3.093	0.873	0.825	0.727	5.226	0.924	0.967	**0.741**	3.6 ± 2.9
XGBoost	0.683	0.859	0.434	3.152	**0.875**	0.827	0.726	5.338	0.919	0.969	0.742	4.6 ± 2.7
MLP	0.665	0.856	0.486	3.109	0.856	0.822	0.727	5.616	0.913	0.968	0.746	8.5 ± 2.6
MLP-LR	0.679	0.861	0.463	3.012	0.859	0.826	0.731	5.477	0.924	0.972	0.744	5.5 ± 2.7
MLP-Q-LR	0.682	0.859	0.433	3.080	0.867	0.818	0.724	**5.144**	0.924	0.974	0.745	5.1 ± 1.9
MLP-T-LR	0.673	0.861	0.435	3.099	0.870	0.821	0.727	5.409	0.924	0.973	0.746	5.1 ± 1.7
MLP-PLR	**0.700**	0.858	0.453	**2.975**	0.874	**0.830**	**0.734**	5.388	**0.924**	0.975	0.743	3.0 ± 2.4
ResNet	0.690	0.861	0.483	3.081	0.856	0.821	0.734	5.482	0.918	0.968	0.745	6.7 ± 3.3
ResNet-LR	0.672	0.862	0.450	2.992	0.859	0.822	0.733	5.415	0.923	0.971	0.743	5.6 ± 2.7
ResNet-Q-LR	0.674	0.859	0.427	3.066	0.868	0.815	0.729	5.309	0.923	0.976	0.746	4.7 ± 2.0
ResNet-T-LR	0.683	0.862	**0.425**	3.030	0.872	0.822	0.731	5.471	0.923	0.975	0.744	4.1 ± 1.9
ResNet-PLR	0.691	0.861	0.443	3.040	**0.874**	0.825	0.734	5.400	0.924	0.975	0.743	3.2 ± 1.3
Transformer-L	0.668	0.861	0.455	3.188	0.860	0.824	0.727	5.434	0.924	0.973	0.743	5.9 ± 2.2
Transformer-LR	0.666	0.861	0.446	3.193	0.861	0.824	0.733	5.430	0.924	0.973	0.743	5.2 ± 2.2
Transformer-Q-LR	0.690	0.857	**0.425**	3.143	0.868	0.818	0.726	5.471	**0.924**	0.975	0.744	4.4 ± 2.2
Transformer-T-LR	0.686	0.862	**0.423**	3.149	0.871	0.823	0.733	5.515	0.924	**0.976**	0.744	3.7 ± 2.2
Transformer-PLR	0.686	**0.864**	0.449	3.091	0.873	0.823	0.734	5.581	**0.924**	0.975	0.743	3.9 ± 2.5

- The key to handling the heterogeneous features is the **embedding layer**
- Resnet and Transformer is not better than **MLP** with suitable heterogeneous embedding

Gorishniy, Y., Rubachev, I., & Babenko, A. (2022). On embeddings for numerical features in tabular deep learning. *Advances in Neural Information Processing Systems*, *35*, 24991-25004.

Problem

- Feature selection is proved critical for heterogeneous datasets.

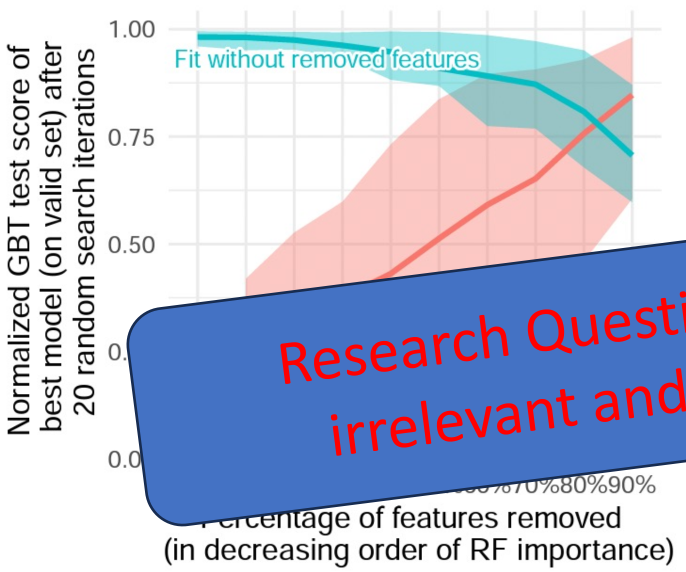


Figure 3: **Test accuracy of a GBT for varying proportions of removed features**, on our classification benchmark on numerical features. Features are removed in increasing order of feature importance (computed with a Random Forest), and ... lines corre... scores are averaged across 30 random search orders, and the ribbons correspond to the 80% interval among the different datasets.

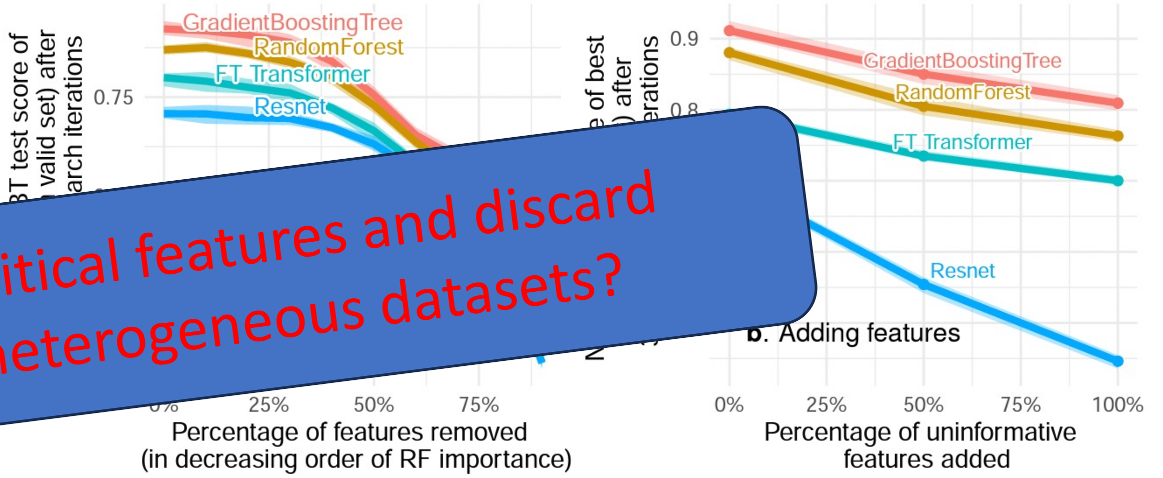


Figure 4: **Test accuracy changes when removing (a) or adding (b) uninformative features.** Features are removed in increasing order of feature importance (computed with a Random Forest). Added features are sampled from standard Gaussians uncorrelated with the target and with other features. Scores are averaged across datasets, and the ribbons correspond to the minimum and maximum score among the 30 different random search reorders (starting with the default models).

Research Question 2: How to preserve critical features and discard irrelevant and redundant features for heterogeneous datasets?

Heterogeneous datasets contain many uninformative features.

MLP-like architectures are not robust to uninformative features.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. *Advances in Neural Information Processing Systems*, *35*, 507-520.

Related works

Name	Loss
LASSO	$\min_{w} loss(w; X, y) + \alpha\|w\|_1$
Group LASSO	$\min_{w} loss(w; X, y) + \alpha \sum_{i=1}^{g} h_i \|w_{G_i}\|_2$
Sparse Group LASSO	$\min_{w} loss(w; X, y) + \alpha\|w\|_1 + (1-\alpha) \sum_{i=1}^{g} h_i \|w_{G_i}\|_2$
Tree-guided Group LASSO	$\min_{w} loss(w; X, y) + \alpha \sum_{i=0}^{d} \sum_{j=1}^{n_i} h_j^i \|w_{G_i}\|_2$
Graph LASSO	$\min_{w} loss(w; X, y) + \alpha\|w\|_1 + (1-\alpha) \sum_{i,j} M(i,j)(w_i - w_j)^2$
GFLASSO	$\min_{w} loss(w; X, y) + \alpha\|w\|_1 + (1-\alpha) \sum_{i,j} A(i,j)(w_i - sign(i,j)w_j)^2$

- Can we combine the existing feature selection approaches with state-of-the-art models for heterogeneous datasets?

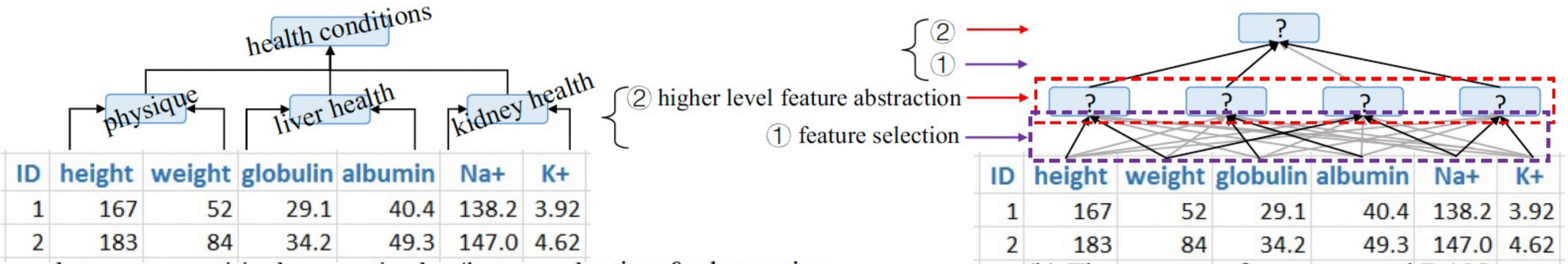
Related works

Heterogeneous feature selection approach

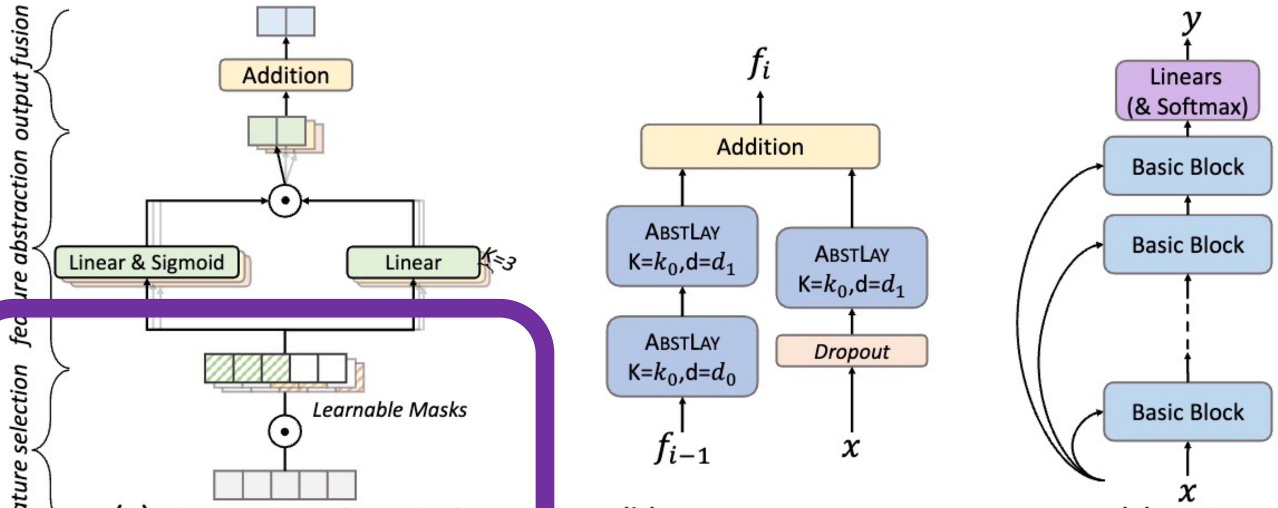
- Mask-based
  - DANETs
- Fuzzy rough set-based FS
  - Fuzzy relation: Hu et al. (2006)
  - Categorical: Wang et al. (2019)
  - Supervised: Yuan et al. (2018), Yuan et al. (2021a)
  - Unsupervised: Yuan et al. (2021b), Zhang et al. (2022)

Related works

DANETs (2017)



(a) A feasible path to extract critical semantics by feature selection & abstraction. (b) The process of our proposed DANET.



(a) *Abstract layer (K=3, d=2)* (b) *The i-th basic block* (c) *DANet*

- Use multiple learnable masks to discard the uninformative features parallelly
- Feature abstraction is used to abstract high level information

Related works

Yuan 2021

Algorithm 1: FMIUFS algorithm.

```
Input: IS = ⟨U, C⟩, threshold value λ, |C| = m
Output: An ordered feature sequence S
S ← ∅, S_u ← C;
for k ← 1 to m do
    Calculate the fuzzy relation matrix M_{R_{c_k}};
    Calculate the fuzzy entropy FE(c_k);
end
for k ← 1 to m do
    for s ← 1 to m do
        Calculate the fuzzy joint entropy FE(c_k, c_s);
        Calculate the fuzzy mutual information FMI(c_k; c_s);
    end
end
for k ← 1 to m do
    Calculate the fuzzy relevance FRel(c_k);
end
Select feature c_{ℓ_1} so that FRel(c_{ℓ_1}) has the maximum value;
S ← S ∪ {c_{ℓ_1}}, S_u ← S_u − {c_{ℓ_1}};
while |S_u| ≠ 0 do
    for l ← 1 to |S_u| do
        for s ← 1 to |S| do
            Calculate the fuzzy redundancy FRed(c_l, c_{ℓ_s});
        end
    end
    Select feature c_{ℓ_r} so that FRel(c_{ℓ_r}) − (1/|S|) Σ_{s=1}^{|S|} FRed(c_{ℓ_r}, c_{ℓ_s})
    has the maximum value;
    S ← S ∪ {c_{ℓ_r}}, S_u ← S_u − {c_{ℓ_r}};
end
return S.
```

Fuzzy relation

$$
r_{ij}^{k}=\begin{cases}1, \text{if} c_k(x_i)=c_k(x_j) \text{ and } c_k \text{is discrete}\\ 0, \text{if} c_k(x_i)\neq c_k(x_j) \text{ and} c_k \text{is discrete}\\ 1-|c_k(x_i)-c_k(x_j)|, \text{if} |c_k(x_i)-c_k(x_j)|\leq \epsilon_{c_k} \text{ and } c_k \text{ is continuous}\\ 0, \text{if} |c_k(x_i)-c_k(x_j)|>\epsilon_{c_k} \text{ and } c_k \text{ is continuous}\end{cases} \tag{14}
$$

where c_k is the measured value of data point x for feature c_k and ϵ_{c_k} a adaptive fuzzy radius. The ϵ_{c_k} is calculated as following,

$$
\epsilon_{c_k}=\frac{std(c_k)}{\lambda} \tag{15}
$$

where $std(c_k)$ is standard deviation of the feature values c_k and λ is a hyper-parameter that is fine-tuned with step 0.1 in the range [0.1,2.0].

Fuzzy entropy

$$
\mathrm{FE}(B)=-\frac{1}{|U|}\sum_{i=1}^{n}\log_2\frac{|[x_i]_B|}{|U|}.
$$

$$
\mathrm{FE}(B|E)=-\frac{1}{|U|}\sum_{i=1}^{n}\log_2\frac{|[x_i]_B\cap[x_i]_E|}{|[x_i]_E|}
$$

$$
\mathrm{FE}(B,E)=-\frac{1}{|U|}\sum_{i=1}^{n}\log_2\frac{|[x_i]_B\cap[x_i]_E|}{|U|}.
$$

Fuzzy mutual information

$$
\mathrm{FMI}(B;E)=-\frac{1}{|U|}\sum_{i=1}^{n}\log_2\frac{|[x_i]_B|\times|[x_i]_E|}{|U|\times|[x_i]_B\cap[x_i]_E|}.
$$

$$
[x_i]_B=\cap_{l=1}^{h}[x_i]_{c_{k_l}}.
$$

$$
|[x_i]_B|=\sum_{j=1}^{n}r_{ij}^{B}=\sum_{j=1}^{n}\mathcal{R}_B(x_i,x_j).
$$

Fuzzy relevance

$$
\mathrm{F\,Rel}(c_k)=\frac{1}{m}\sum_{s=1}^{m}\mathrm{FMI}(c_k;c_s).
$$

$$
\mathrm{FRel}(c_{\ell_s}|c)=\frac{\mathrm{FE}(c_{\ell_s}|c)}{\mathrm{FE}(c_{\ell_s})}\mathrm{FRel}(c_{\ell_s}).
$$

Fuzzy redundance

$$
\mathrm{FRed}(c,c_{\ell_s})=\mathrm{FRel}(c_{\ell_s})-\mathrm{FRel}(c_{\ell_s}|c)
$$

The selected feature subset can minimize the uncertainty of other unselected features.

Yuan, Z., Chen, H., Zhang, P., Wan, J., & Li, T. (2021). A novel unsupervised approach to heterogeneous feature selection based on fuzzy mutual information. *IEEE Transactions on Fuzzy Systems*, *30*(9), 3395-3409.

Related works

How do they handle the heterogeneous features?

Hu 2016

$$r_{ij}^k = \begin{cases} 1, \text{if} f(x_i, a) = f(x_j, a) \text{ and A is discrete}, \forall a \in A \\ 0, \text{if} f(x_i, a) \neq f(x_j, a) \text{ and A is discrete}, \forall a \in A \\ f(||x_i - x_j||), \text{if A is continuous} \end{cases}$$

Zhang 2022

$$d_{ij}^k = \begin{cases} 0, \text{if} f(x_i, a) = f(x_j, a) \text{ and} a \text{ is discrete} \\ 1, \text{if} f(x_i, a) \neq f(x_j, a) \text{ and } a \text{is discrete} \\ |f(x_i, a) - f(x_j, a)|, \text{if } a \text{ is continuous} \end{cases}$$

Yuan 2018, 2021

$$r_{ij}^k = \begin{cases} 1, \text{if} c_k(x_i) = c_k(x_j) \text{ and } c_k \text{is discrete} \\ 0, \text{if} c_k(x_i) \neq c_k(x_j) \text{ and} c_k \text{is discrete} \\ 1 - |c_k(x_i) - c_k(x_j)|, \text{if} |c_k(x_i) - c_k(x_j)| \leq \epsilon_{c_k} \text{ and } c_k \text{ is continuous} \\ 0, \text{if} |c_k(x_i) - c_k(x_j)| > \epsilon_{c_k} \text{ and } c_k \text{ is continuous} \end{cases} \tag{14}$$

where c_k is the measured value of data point x for feature c_k and ϵ_{c_k} a adaptive fuzzy radius. The ϵ_{c_k} is calculated as following,

$$\epsilon_{c_k} = \frac{std(c_k)}{\lambda} \tag{15}$$

where $std(c_k)$ is standard deviation of the feature values c_k and λ is a hyper-parameter that is fine-tuned with step 0.1 in the range [0.1,2.0].

Wang 2019

$$r_{ij}^B = \frac{1}{|A|} card(k \in B : c_k(x_i) = c_k(x_j))$$

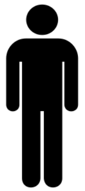
Discretization for continuous features

The goal of those works is to find a feature subset that contains most or all of the information in the original feature set based on **the entropy they defined from the relation function or distance function**.

Methodology for Embedding

Motivation

- The existing embedding module did not utilize the information on **ordinal** features and the **global frequency of assignment**

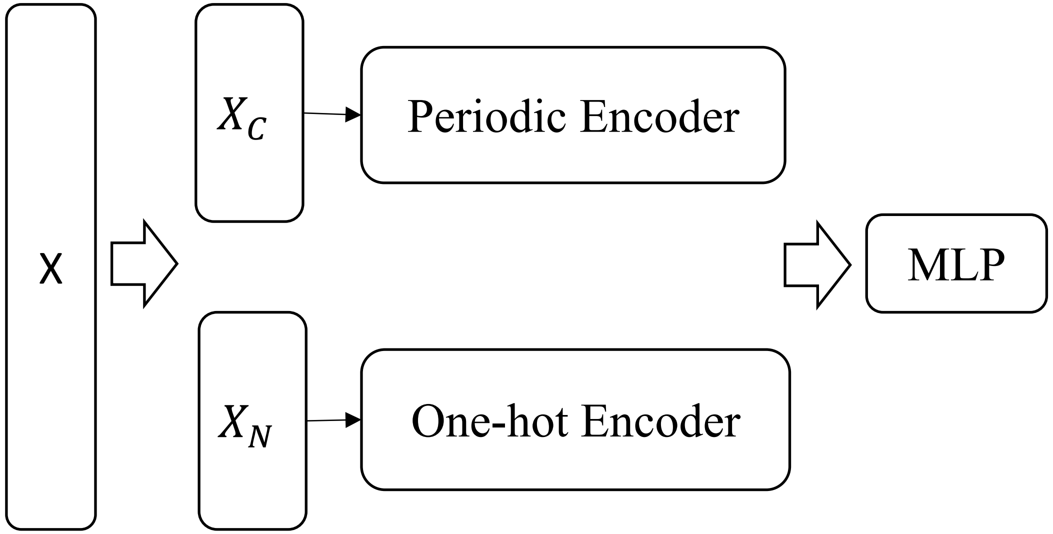


Feature	Value
Sex	Male
Degree	PhD
Income	1.8k

Occurrence	Observation	Probability	Coder
Male	1	0.5	1/0.5
Female	0	0.5	0
High School	1	0.6	1/0.1
Bachelor	1	0.3	1/0.1
PhD	1	0.1	1/0.1
0-10K	1	0.1	1/0.4
10k-20k	0.8	0.4	1/0.32
20-30k	0	0.4	0
>30k	0	0.1	0

Methodology for Embedding

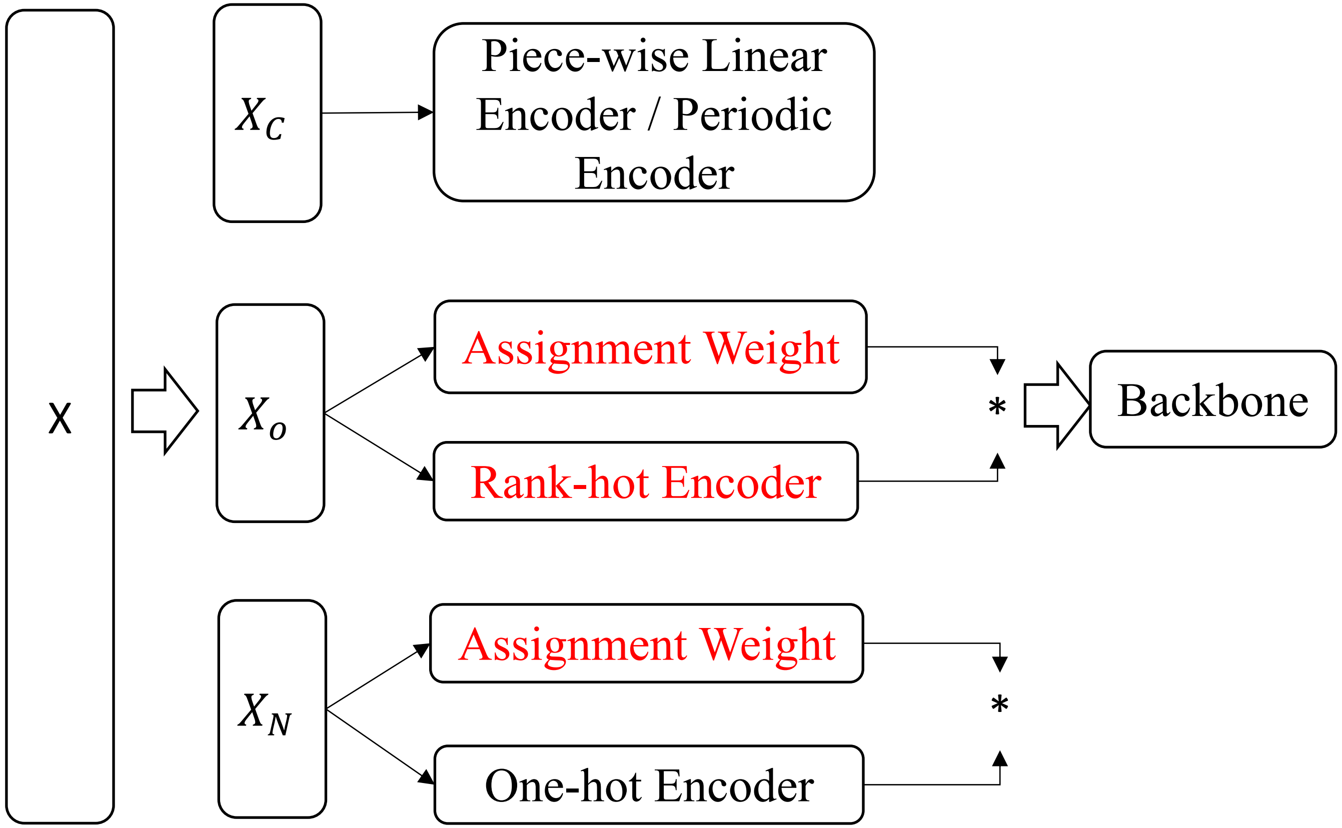
SOTA (2022)



$$f_i(x) = \texttt{Periodic}(x) = \texttt{concat}[\sin(v), \cos(v)], \quad v = [2\pi c_1 x, \ldots, 2\pi c_k x] \tag{8}$$

where c_i are trainable parameters initialized from $\mathcal{N}(0, \sigma)$. σ is an important hyperparameter that is tuned using validation sets.

Our embedding architecture



Assignment weight $w_{f=a} = \dfrac{n}{n_{f=a}}$

Methodology for Embedding

Alert2AKI Dataset

Intervention	AKI Alert or Not
Main-outcome	AKI Progression in 14 Days
Pre-treatment	EHR Records
Patients Num	6030 in 5 Hospitals (5082/948)

SCALE	**NUM**
Nominal	9
Ordinal	19
Interval	3
Ratio	20

We can compare the predicted outcome difference between different treatments for an individual to decide whether a patient should accept the treatment.

PR-AUC (5 Random Splits)

Random
.1568±.0089

MLP Backbone		
HetMLP	HetMLP_nW	MLP
.2117±.0009	.2087±.0164	.2009±.0329
Resnet Backbone		
HetResNet	HetResnet_nW	Resnet
.2087±.0255	.2033±.0145	.1711±.0186

Our HetMLP got a **1.43%** performance up compared with SOTA on this dataset.

AKI: Acute Kidney Injury

Will the patients benefit from the alert?

Splitting 1

Metrics	Num
Patients Num	3536
Benefited: AKI=1→AKI=0	15
Harmful: AKI=0→AKI=1	14

Splitting 2

Metrics	Num
Patients Num	3552
Benefited: AKI=1→AKI=0	8
Harmful: AKI=0→AKI=1	2

Splitting 3

Metrics	Num
Patients Num	3504
Benefited: AKI=1→AKI=0	26
Harmful: AKI=0→AKI=1	4

Splitting 4

Metrics	Num
Patients Num	3536
Benefited: AKI=1→AKI=0	9
Harmful: AKI=0→AKI=1	9

The model's prediction is consistent with the conclusion that Alerts did **not** reduce rates of our primary outcome among hospitalized patients with AKI.

Futural Plan

- **Heterogeneous Embedding**
  - **Heterogeneous feature structure and instance structure (such as cluster)**
  - **Computation Complexity**
  - **Detailed experiments on more backbone and datasets**
- **Heterogeneous Feature Selection**
  - **Discover more effective heterogeneous feature distance (Wasserstein etc.)**
  - **Combine intra-attribute structures and inter-attribute structures**