ZJUI学院2023年暑期科研成果展
23号队伍：龙奇、郭柄君、李宇轩
指导老师：苗晓晔（浙江大学数据科学研究中心）

Large Language Model in Representative Learning of Tabular Data

ABSTRACT

To enhance the data cleaning process on huge data nowadays, our team proposed an innovative representation learning method to improve data utility efficiency and the performance on the wide-range down stream tasks. Our proposal aimed at dealing with missing data, dirty data and normal data representation at the same time, and taking missing data and dirty data into consideration along with normal data to even enhance the model performance on down stream tasks. Specifically, our model is designed based on **Raha error detection model** and **SAINT tabular data deep learning model.** Besides, applying **Large Language Model (LLM)** into this whole industrial pipeline is another outstanding attempt in tabular data area. The experiment results showed that our model can fully use missing data, dirty data and normal data in a whole to reach a state-of-art performance on downstream tasks.

EXPERIMENT

Comparing SAINT and our model on Hospital Dataset, our model had a better accuracy in 100 epochs and lower pretrain loss as well.

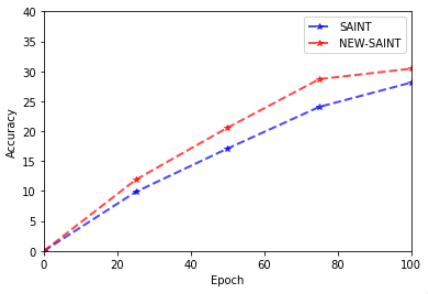


Table 1: Experiment on SAINT and NEW-SAINT

Flight Model	Raha + SAINT Dirty	Raha + New-SAINT Dirty
Pretrain Loss (init)	251.7421341	251.4068947
Pretrain Loss (final)	143.5876174	106.1310303
Accuracy (25)	9.855	**11.884**
Accuracy (50)	17.101	**20.580**
Accuracy (75)	24.058	**28.696**
Accuracy (best)	28.116	**30.435**

FRAMEWORK

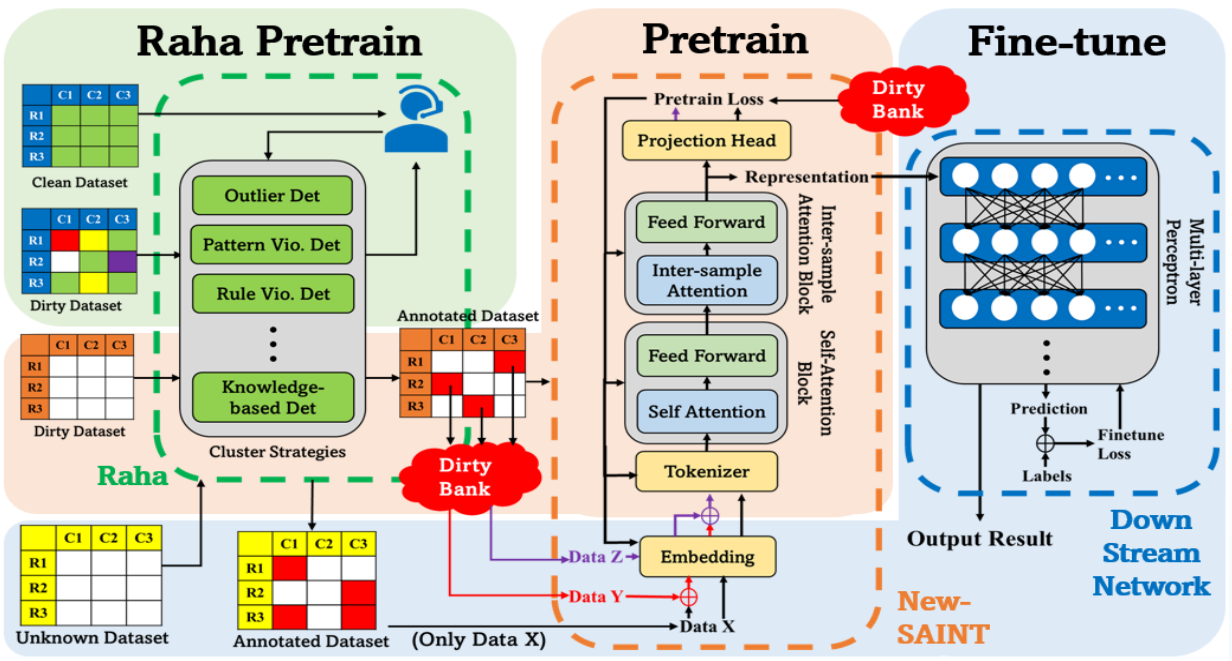


Figure 1. Innovative representation learning method for tabular data.

Raha-guided NEW-SAINT pretrain strategy:

$$TargetC = \begin{cases} Z_i, & else \\ Z_j, & if\ X_i\ dirty; X_j\ clean \\ Z_k, & if\ X_i, X_j\ dirty; X_k\ clean \end{cases}$$

$$WeightC = \begin{cases} 0.5, & if\ X_i, X_j, X_k\ dirty \\ 1, & if\ X_i, X_j, X_k\ clean \\ 2, & if\ X_i \rightarrow one\ of\ X_j, X_k\ diff \\ 4, & if\ X_i \rightarrow X_j, X_k\ diff \end{cases}$$

$$WeightD = \begin{cases} 0, & if\ X_i\ dirty \\ 1, & else \\ 2, & if\ X_i\ clean; X_j\ or\ X_k\ dirty \\ 3, & if\ X_i\ clean; X_j, X_k\ dirty \end{cases}$$

MODEL

➢ Raha

Raha can detect dirty data at a state-of-art performance. It has multiple strategies including outlier, pattern violation, rule violation, knowledge-based violation detection and so on.

➢ SAINT

SAINT makes use of transformer's attention mechanism, extends it to tabular data and innovatively proposes a Self-Attention connected with Inter-Sample Attention Block to encode each row of tabular data along with its surrounding inline and crossline information. Besides, its pretrain-finetune framework can fully utilize huge quantities of dataset and can easily fit into Large Language models.

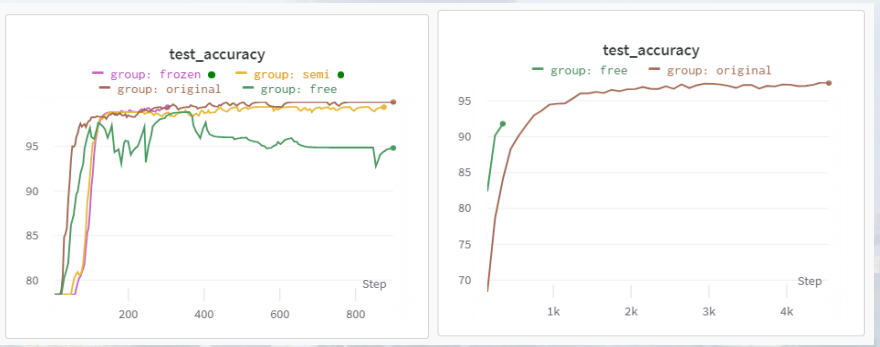
➢ NEW-SAINT

Considering above, we decide to build our model upon SAINT, taking advantage of Raha and LLMs simultaneously. Specifically, the model is divided into three stages:

1) Raha pretrain: trained under user guide to detect potential dirty data in the table and annotates them.
2) Pretrain: Innovative Raha-guided pretrain makes use of known clean or dirty data, mixes data with **noises** on instance level and embedding level. Finally, by maintaining the representation with different **weights** based on mixed data type (clean or dirty), it can reach a high robustness and state-of-art representation strategy.
3) Finetune: Executed on specific downstream tasks including multi-class classification and prediction.

LARGE LANGUAGE MODEL (LLM)

➢ Due to the typically high-quality and comprehensive knowledge coverage of the training data of pre-trained language models, numerous studies have found that retaining some of their parameters and applying them in the form of transfer learning to domains beyond text data representation can yield very positive results. We utilize the parameters of a well pretrained BERT model and the representative results are as follows:



➢ It turns out that wiht sufficient data for fine tuning to avoid overfitting, our model utilizing the pretrained LM obvious outperformst the model that is not pretrained.

CONTACT

Qi Long: qi.21@intl.zju.edu.cn
Bingjun Guo: bingjun.21@intl.zju.edu.cn
YuXuan Li: yuxuan.21@intl.zju.edu.cn