浙江大学

ZJUI 学院 2023 年暑期科研项目总结材料

中文论文题目 ： 基于大型语言模型的数据清洗增强框架

英文论文题目： The enhanced data cleaning framework based on large language models

成员姓名：龙奇（3210115418）

郭柄君（3210115445）

李宇轩（3210116308）

指导教师：苗晓晔

指导教师所在学院：浙江大学数据科学研究中心

暑期研究起止日期：2023 年 6 月 15 日-2023 年 7 月 15 日

摘要

为了提升大规模表格数据清洗与表征过程，我们的团队提出了一种创新的表示学习方法，以提高数据利用的效率及其在众多的下游任务上的性能。我们提出的方法旨在同时处理缺失数据、脏数据和正常数据的表示，并考虑缺失数据和脏数据以及正常数据，以进一步提高模型在下游任务上的性能；具体而言，我们的模型基于RAHA错误检测模型和SAINT数据表征模型设计的。此外，将大型语言模型（LLM）应用于整个过程是对于表格数据建模表征的的另一杰出尝试。已有的实验结果表明，我们的模型可以充分利用缺失数据、脏数据和正常数据、在已测试下游任务上达到了先进的性能水平并具有完全超越原有模型的潜力；同时，由于有限的时间与计算资源的限制，部分的判断与更深入的设计仍需进一步的实验证实。

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 as it reaches a state-of-art performance on tested downstream tasks, and thus retains the potential to outperform the original method in all means. Meanwhile, due to the limit time and computing resources, some judgments and deeper designs still requires substantiated with further experiments.

1 背景及论文调研

一方面，随着大数据时代的到来，数据来源拓宽、数据库规模增大、数据信息挖掘价值极高。数据清洗（Data Cleaning）作为数据推理应用的关键步骤，成为本次项目的问题核心。另一方面，随着人工智能、深度学习（Deep Learning）的快速发展，不断增加的计算资源和大型标记数据集的可用性加速了深度神经网络的成功。尽管深度学习方法在同质数据（Homogeneous Data） （如图像、音频和文本数据）上的分类或数据生成任务中表现出色，但表格数据（Tabular Data）仍然对深度学习模型构成挑战[1]。同时，注意到近年来大型语言模型（Large Language Model）的迅速发展，如何搭建能够应用于表格型数据的大模型成为新的研究领域。基于上述背景，我们团队探索将深度学习的方法应用于表格数据处理的可行方案，基于数据表征学习（Representative Learning）的方法，从缺失数据表征及脏数据表征两大方向进行创新，旨在搭建能灵活处理包含缺失值、异常值的表格型数据库，最大程度挖掘有用信息，在完成下游任务的效果上达到提升。同时，对大型语言模型应用于表格型数据展开研究，实现大模型赋能表格数据的处理。（仿宋小四号或12磅,1.5倍行距）

1.1 表格型数据的特征

表格型数据是高度结构化的数据类型，具有异质性（heterogeneous），数值密集（dense numerical），类别稀疏（sparse categorical），弱相关性（weak correlation）等特征[1]。表格中不同实例以行表示，不同特征以列表示。值间可能独立也可能相关。没有固有位置信息，列的顺序随机的特征使得表格型数据的值间关系挖掘不依赖于值的位置信息。

由于表格数据来源广泛、人为输入，脏数据出现频繁。数据错误类型包括语法（syntactic）和语义（semantic）。数据清洗则包括错误检测（Error Detection）和错误修正（Error Correction）两个步骤。[2][3]

1.2 表格型数据与机器学习

传统的机器学习方法以非参数树模型（non-parametric tree-based models）为主，包括XGBoost, CatBoost, LightGBM等。基于深度学习的新的模型包括TabNet, VIME, TABERT等。到目前为止，传统的机器学习方法在处理表格型数据上仍占主导地位，效果普遍高于深度学习方法。其中决策树（Decision Tree）具有对有近似超平面边界的决策流形来说非常有效、通过跟踪决策节点可实现高度可解释性、训练周期短等优点。[1]

1.3 深度学习中的迁移学习与表征学习

1.3.1 迁移学习（Transfer Learning）指当训练数据在一个领域中稀缺且昂贵，而在类似领域中广泛可用时，学习模型可以在相关领域上进行预训练，然后在当前数据集上进行微调[1]。可以解决工业应用场景下表格型数据隐私权限、脏数据比例高、大面积数值缺失等问题。

1.3.2 表征学习（Representative Learning）即为通过无监督或半监督的方式学习输入的“表征”，即学习保留特定数据特征地将输入数据映射至向量空间中，以方便对数据的后续处理与应用。迁移学习即为表征学习的一种重要形式——当在一组数据集上提取特征过于困难或者其特征不够丰富，可以将在其他大规模、易提取数据集上训练出的表征“迁移”至此场景中以便利下游任务。

1.4 前沿的表格型数据清洗模型

1.4.1 RAHA

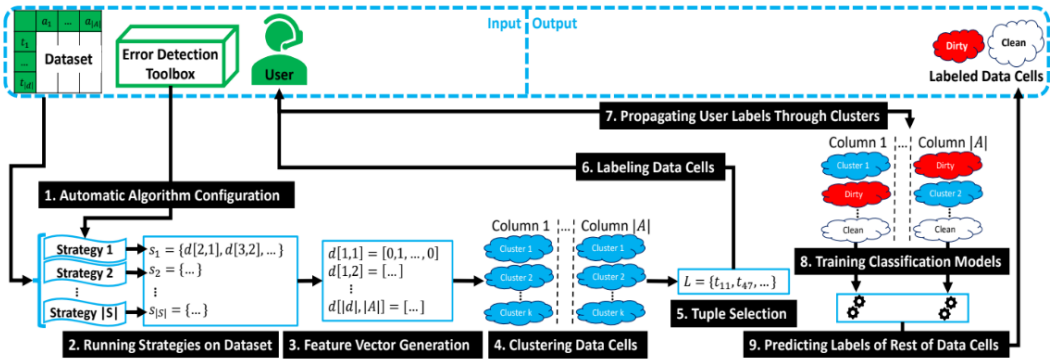


Figure 1: The workflow of Raha.

RAHA模型是最前沿的针对表格数据的错误探查模型之一，其创新之处在于统合多种不同算法来选择的错误探查策略以及利用该策略生成的特征向量将表格数据切分成多个不同模块分别进行人为打标的预训练方式。在获取输入的表格数据（常为.csv文件）后，RAHA会利用内置的算法来生成适合该组数据的多种错误探查策略（detector）；每一组策略都会生成一组探查结果，而RAHA模型随后会利用探查结果生成一组由1和0组成的特征向量；程序会根据特征向量的相似度来切分数据并分组，从每组数据中抽取元组来进行人为打标，其后再将人为标注正误的数据输入分类模型进行训练。

该模型的探测精度和处理速度都相当突出，非常适合本次研究有关表格数据清洗的主题，同时程序会自动储存对于每一组数据探查过程中生成的错误探查策略，并保存在对应数据的文件夹供下次调用，能够显著减少对同一组数据多次操作时所耗费的时间。但是其

缺点在于泛用性不够强，当更换来自其整合包外部的新数据组时，若新数据与模型预训练中使用的数据相似度不够高则需要通过改写程序结构来对新数据进行手动的标注，面对较大的外部数据组时会消耗大量时间在重复性标注工作上。

1.4.2 BARAN

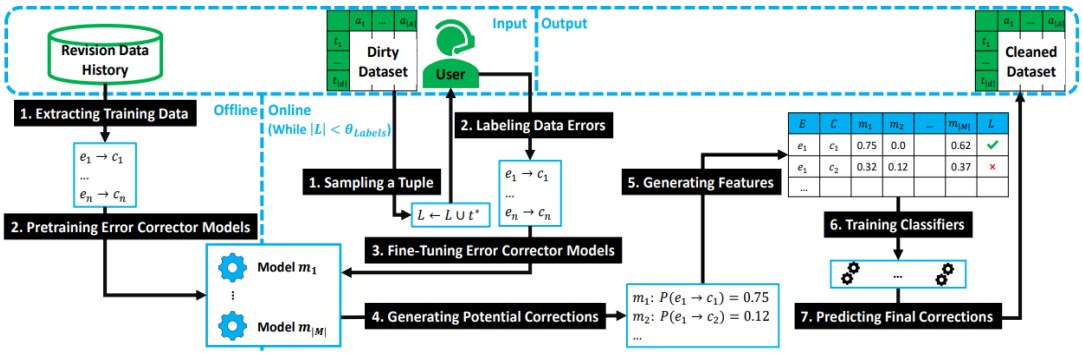


Figure 1: The workflow of Baran.

BARAN模型是与RAHA模型配套的错误修正模型，其突出优势在于可以以较高的精度修正多数自然语言模型较难处理的表格类数据。BARAN模型的作用原理和RAHA模型很相似，首先对于输入数据的元组进行抽取与人工标注，并且利用人工标注正误的元组对模型进行训练；该模型内部也整合了多种不同的修正策略（corrector），每个corrector都会针对元组数据生成候选的修正结果，程序随后会利用对于元组的人工标注来计算每个修正结果的适应性（fitness）来决定最佳的修正方式，且同样生成一套标注修改点的特征向量；最后将修正后的数据和人为标注修正的数据混合，输入到程序的二分类器（classifier），训练模型区分人为标注答案与机器生成的答案。

值得一提的是，开发者借鉴了Bert，GPT等大语言模型等训练步骤，在正式开始针对目标任务的训练之前，该模型先利用维基百科等开源大型数据库进行了预训练，该步骤有效提高了模型在面对具体任务时的学习效率。

BARAN模型的不足之处在于，其相当依赖利用人为标注的数据来训练二分类器这一步骤，而现实中获取的大量数据样本很可能是不存在正确数据的样本备份，或者难以进行人为标注的，在这种条件下BARAN模型的修正效果可能较为不理想。

1.5 前沿的深度学习模型

1.5.1 Transformer

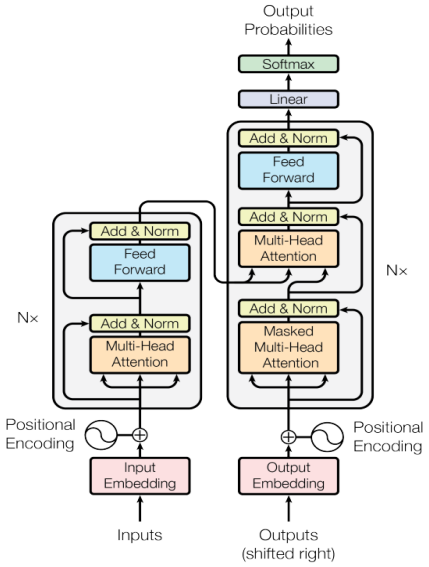


Figure 1: The Transformer - model architecture.

作为现如今自然语言处理领域最受欢迎的模型之一，Transformer模型分为encoder与decoder两个结构，创新地引入了self-attention机制和multi-head attention机制来处理针对翻译任务的三种关系（relation），同时使用了positional encoding机制来保存词语在句子中的相对位置。

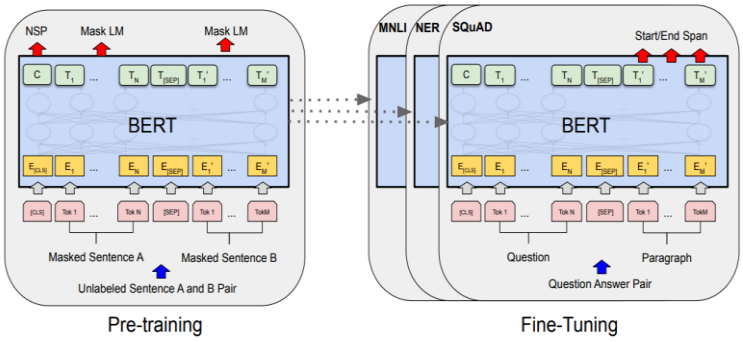
对于自然语言处理任务而言，元素被分为翻译目标（Target）与翻译对象（Source）两组，与大部分注重于两者间数学关系的机制不同，Self-attention机制注重于捕获单个句子中单词之间的语义特征，例如代词的指代对象等；为了学习内容词组之间的关系，Self-attention模块通过embedding将任务和目标词句分成问题（Queries），被查询对象（key）与实际特征信息（value）三组加权矩阵，其中Q，K在每两层嵌入层之间都是相关联的；经过矩阵计算并对每一行进行softmax操作后，输出一个包含了输入语句特征信息的矩阵Z。Multi-headed Attention是指通过并联多组不同的self-attention层，利用多组处理得到的QKV矩阵进行计算得出不同的特征矩阵Z，最后拼接。

此外，由于Transformer是用的是position embedding来表现单词在句中的相对位置，该模型同时配套了掩码（mask）来确保让decoder的预测只基于此前的输出，而不会错误地使用在目前单词之后才被翻译的内容作为预测的参考信息。

在本次课题研究中，我们选择 Transformer 而不是决策树作为基本深度网络，因为它在表征学习方面非常有效，而且拥有包括 BERT 在内的先进架构；同时，注意机制对表

格数据也很有用，类似于处理翻译任务中的输入句，我们也主要关注输入数据行内和行间的关系。

1.5.2 BERT



在Transformer的基础上，BERT提出预训练加微调的模型架构。

预训练（pretrain）的任务固定，在大规模无需标签的数据集上采用无监督的训练方式。具体包括MASK任务和判断句子连接任务。MASK任务将句子中的词随机替换为特殊嵌入（special tokens），训练BERT根据句子中剩余信息预测替换的词；判断句子连接任务将任意两个句子相连，训练BERT判断两个句子是否满足先后相连关系。[5]

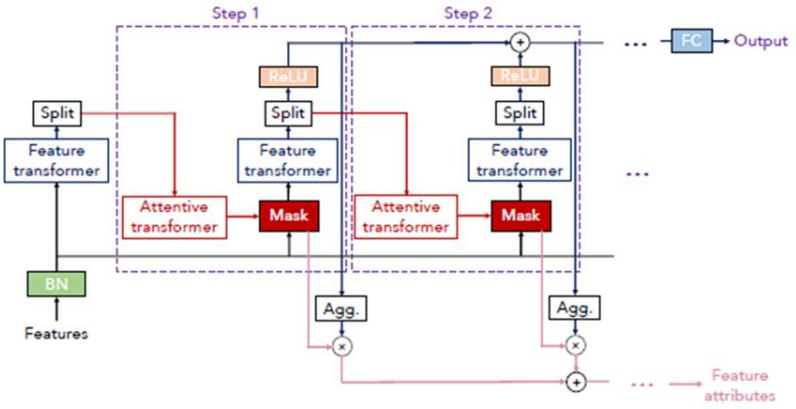
微调（fine tune）以满足下游任务，在预训练参数权重的基础上，监督学习以达到预测结果靠近标签。[5]

此外，BERT采用的双向嵌入不同于传统的Transformer，单词的嵌入不再只取决于之前的子句，而取决于整个剩余句子，使单词嵌入可包含更多信息。[5]

BERT的预训练阶段将编码器（encoder）的初始化从下游任务中分离出来，适应了大量数据无法完全人工打标的表格型数据场景。同时，预训练-微调的框架成为后续大模型诞生、发展的重要基石，对表格型数据清洗、推理模型的训练也提供了新的思路。但是，BERT位置嵌入、句子输出等设计不能很好地处理表格型数据。BERT的预训练任务也没有与表格型数据完美契合，对表格型数据的定制化预训练方案仍待研究。

1.6 前沿的表格型数据深度学习表征模型

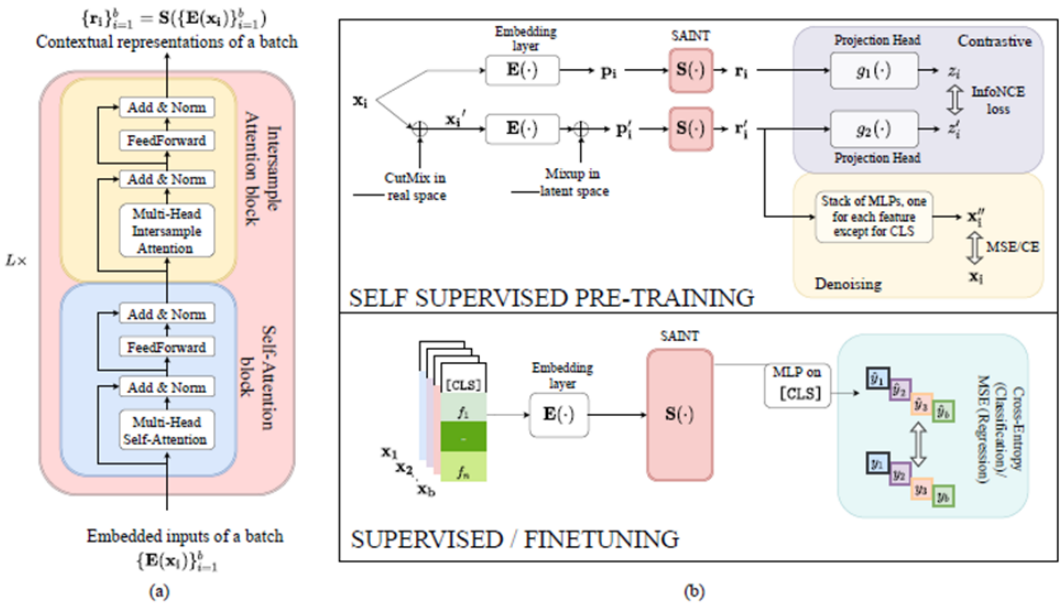
1.6.1 TabNet



基于Transformer主要使用的注意力机制（attention），TabNet提出多层特征选择的句子注意力机制（sequential attention）。具体而言，句子经过嵌入后每一步都将只对句子中部分特征进行注意力机制计算，所有的步计算结果通过一定参数合并得出最后表征。[6]

TabNet针对表格型数据提出连续值和离散值区别嵌入的方法，保留了列与列之间的区别。同时，特征选择的注意力机制也很好配适了表格型数据值与值间关系不固定的性质，是注意力机制在处理表格数据的成功探索。

1.6.2 SAINT



SAINT优化了Transformer注意力机制在表格型数据的应用，采用自注意力（self-attention）和实例间注意力（inter-sample attention）结合的方法表征表格数据。具体而言，自注意力通过多头自注意力层（multi-head self-attention layer），用注意力矩阵计算行内单元格间关系系数，在表征中加入行内其他单元格值信息；实例间注意力层则将单元行q分享给其他行，结合其他行的k，v计算行间关系系数，对单元行的表征进一步加入其他行的有关信息。[7]

SAINT沿用了BERT的预训练-微调框架。对预训练任务进行了优化，包括对比学习（contrastive learning）和噪声消除（denoising）。对比学习在实例层和嵌入层分别将指定单元行内部分数据替换为其他行对应数据，经过训练好的SAINT后，其表征与指定单元行本身的表征接近。噪声消除则在指定单元格混入其他单元格数据表征后进行固定全连接层多分类，训练SAINT达到还原表征为原指定单元格值。[7]

SAINT的两种注意力机制很好配适了表格型数据，使表征既包含所在行内信息，也包含其他单元行信息。SAINT的预训练任务设计对增强表征的鲁棒性，减轻表征可能受到的脏数据影响有巨大作用。在表格数据清洗和相关推理下游任务上达到了目前最好的效果（state of art）。但处理脏数据表征时可能还有改进空间。

2 部分模型评估

在模型调研过程中，我们选出了RAHA, BARAN这两个在表格型数据清洗上效果最好的模型和 TabNet, SAINT这两个在表格型数据表征上效果最好的模型进行测试评估。源码测试的实验细节如下：

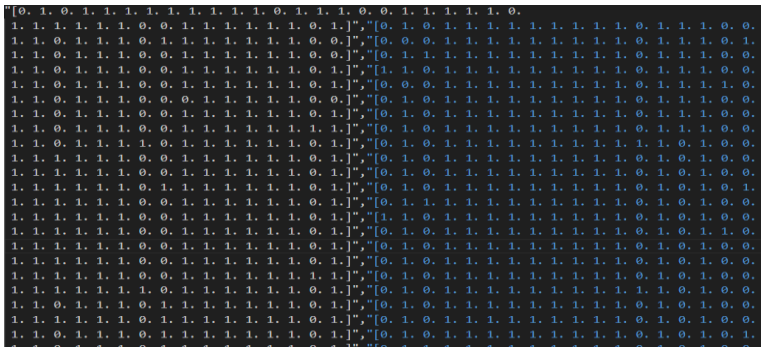
2.1 RAHA-BARAN

在RAHA-BARAN模型内自带了7组大小、复杂程度、数据类型与内容皆有所不同的表格数据集;RAHA模型和BARAN模型既可以分别独立根据这些数据运行计算，也可以通过开发者整合的将二者串联的pipeline将两个模型按顺序在同一组数据上进行训练。RAHA与BARAN模型输出结果均为探测/修改后表格数据与正确的表格数据（clean data）相比较从而计算出的PRF值。在其中的‘flights.csv’数据集中运行RAHA-BARAN的整合pipeline结

```
(raha) summer2023@dell-PowerEdge-R750xa:~/home/summer2023/rahaCode/raha-master/raha$ python detection.py
I just load strategies' results as they have already been run on the dataset!
Raha's performance on flights:
Precision = 0.80
Recall = 0.82
F1 = 0.81
```

果如下：

同时，通过修改代码，RAHA-BARAN模型也可以实现在运行中输出detection和correction过程中生成的特征向量，并存储在‘feature.csv’中：



2.2 TabNet

在Forest-cover-type训练集上，根据树木种类、地形、气候等完成植被类型的多分类任务。Forest-cover-type训练集数据量巨大，没有缺失数据或错误数据。训练共100轮。最后达到0.90的正确率，验证集和测试集结果接近。

```
BEST VALID SCORE FOR forest-cover-type : 0.9014889923000962
FINAL TEST SCORE FOR forest-cover-type : 0.9038320869512835
```

2.3 SAINT

在OpenML数据集上，选取ID1、ID2、ID5训练集完成多分类任务。分有预训练和无预训练两个实验。ID1和ID2数据量小，ID5数据量极大，均为干净数据集。训练共100轮，实验结果如下：

Dataset	Pretrain Loss	Best Accuracy（PretrainedRAHA	Best Accuracy（Not PretrainedRAHA
ID1	3.8757	97.740	99.432
ID2	3.7508	97.740	99.432
ID5	9.7308	66.667	73.267

由于预训练需要大规模数据集，选取下游任务数据集的部分数据先进行预训练，效果不如直接微调。阅读SAINT源码，发现其本身可以处理缺失值。

3 模型提议及实验

3.1 Model 1: RAHA + SAINT

我们构筑的第一个模型是RAHA模型与**SAINT**模型的串联。从之前的源码测试中，我们总结到：RAHA模型的错误检测能力相当可观，而SAINT模型虽然不能分辨脏数据却可以检测出空白数据并利用一些策略来替代缺失数据从而将其加入学习流程。因此我们创造了一层简单的修改程序，将RAHA检测的所有错误结果都替换成空白值（NULL），并结合SAINT的不同修正策略进行试验。分别利用以下的修改策略对其进行修正：

$$Categ\ Dirty_{i,j} = \begin{cases} "MissingValue", & (1) \\ X_{\max freq\ in\ j}, & (2) \end{cases}$$

其中“*Conti*”指连续值类型的数据，“*Categ*”指离散型的数据；“*Avg of j*”指该数据所在列的均值，“*max freq in j*”指该数据所在列中的众数。利用不同的修改策略替代RAHA识别的脏数据后将表格输入SAINT模型进行分类，运行结果如下：

Accuracy	Model (1 & 3)	Model (1 & 4)	Model (2 & 3)	Model (2 & 4)
SAINT & dirty	19.958	**20.166**	18.503	19.958
Raha-SAINT & dirty	20.582	20.582	22.245	**22.453**
SAINT & clean	20.582	21.206	22.037	**22.453**

可以从表格中看到，采用处理策略2和4的组合时，Model 1的效果要强于SAINT模型，但是效果并不算是很明显，且在其他的条件下学习效率仍然不如SAINT；另外，总体而言Model 1所实现的分类正确率仍然偏低，整体价值有限。

3.2 Model 2: RAHA + BARAN+ SAINT

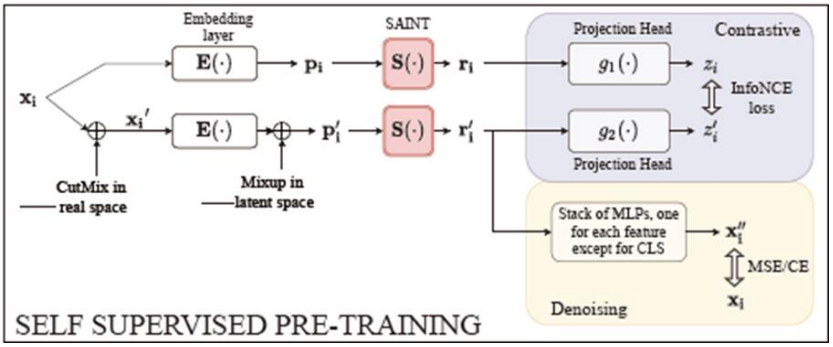
我们构筑的第二个模型是利用RAHA-BARAN的串联pipeline实现对于含有脏数据的输入表格的整体错误检测以及修正，在此之后将修正后的表格输入SAINT模型，由其来进行分类的学习任务。从Model 1的实验结果中我们可以发现，SAINT模型处理干净数据时相比处理脏数据，效率和正确率都有所提升；因此我们在Model 2中尝试将RAHA-BARAN模型检测并修正后的正确率更高的数据输入SAINT模型，运行结果如下：

Dataset	Model	Repair's Precision	Repair's Recall	Final Accur
Flights	SAINT only	N/A	N/A	86.287
Flights	Raha+SAINT	0.88	0.84	76.160
Flights	Baran+SAINT	1.00	1.00	87.975
Flights	Raha+Baran+SAINT	0.84	0.50	87.975
Hospital	SAINT only	N/A	N/A	90.576
Hospital	Raha+SAINT	0.73	0.44	88.968
Hospital	Baran+SAINT	0.87	0.85	90.576
Hospital	Raha+Baran+SAINT	0.84	0.50	90.576
Rayyan	SAINT only	N/A	N/A	94.562
Rayyan	Raha+SAINT	0.86	0.77	90.812
Rayyan	Baran+SAINT	0.74	0.48	93.194
Rayyan	Raha+Baran+SAINT	0.98	0.02	93.194

在上文的表格中，我们针对三组数据做了实验。其中Flights的错误率最高，Hospital的缺失数据最多，Rayyan的数据规模最大。我们对比了四种处理方式，直接将脏数据输入SAINT模型（SAINT only），Model 1中的RAHA与BARAN串接（RAHA+SAINT），直接利用BARAN模型处理脏数据后输入SAINT模型（BARAN+SAINT），以及Model 2中探讨的RAHA-BARAN串接处理后输入SAINT（RAHA+BARAN+SAINT）。

根据实验结果我们总结出，Model 2的模型架构相对于Model 1有所提升，该模型在面对较小的数据集时可以发挥更好的效果，且数据集错误率偏高时该模型的效果相对其他模型更显著，也更稳定。Model 2或许是一个在未来研究中可以利用的有效架构。

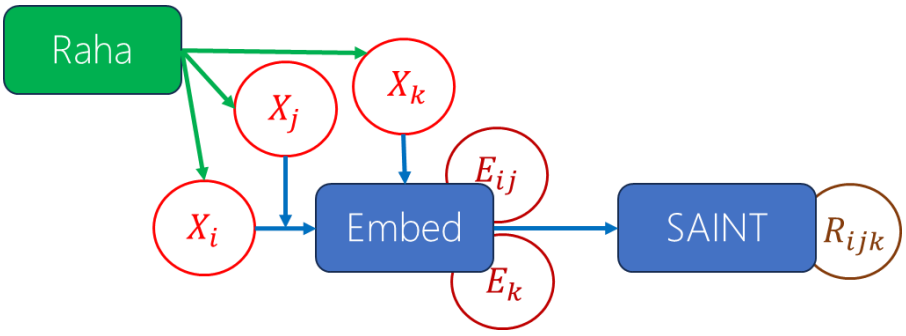
3.3 Model 3: RAHA + New-SAINT Pretrain



SAINT预训练任务存在不足。第一，没有区分原本的干净数据和脏数据，导致一些错误无法用表征消除其影响。第二，脏数据表征的鲁棒性相较于干净数据对下游任务意义不大，但在预训练中二者的鲁棒性对损失函数影响相同，可能导致预训练收敛方向不理想。第三，对脏数据的表征理想情况下是机器可以学习到如何提取其中的有用信息，对应SAINT预训练中混合数据表征接近原始数据。但若原始数据为脏，表征后接近原始数据则相当于覆盖了有用信息。

RAHA具有能够标记出干净数据和脏数据，引导SAINT预训练的作用。相较于BARAN，RAHA准确率更高，能不改变原来的数据输入，避免丢失信息。

基于上述分析，我们提出RAHA + New-SAINT Pretrain模型。改进的目标是使干净数据表征不受脏数据影响，脏数据的表征尽量保留有用信息。模型的核心是RAHA引导SAINT有意学习有价值的表征能力。



预训练任务沿用SAINT提议，分为对比学习和噪声消除。

对比学习（Contrastive Learning）：

Original	Mix	Label	Weight
Clean	Clean	Clean	Medium
Clean	Dirty	Clean	**High**
Dirty	Clean	Clean	**High**
Dirty	Dirty	Dirty	Low

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

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

噪声消除（Denoising）：

Original	Noise1	Noise2	Weight
Dirty	Clean / Dirty	Clean / Dirty	0
Clean	Clean	Clean	1
Clean	Clean	Dirty	**2**
Clean	Dirty	Clean	**2**
Clean	Dirty	Dirty	**3**

理论分析：增大对比学习中干净数据与脏数据混合在损失函数中的权重，能够训练机器从同时带有干净信息和脏信息的混合数据中提取干净信息；降低对比学习中完全由脏数据混合而成的样本在损失函数中的权重，能够使训练减少关注这部分错误信息。增大噪声消除中增大干净数据掺入脏数据的样本在损失函数中的比重能够训练模型更好地从干扰数据中提取原本干净的数据，增强鲁棒性；若原始数据为脏，则无视其噪声消除对训练的影响。

实验结果：在RAHA数据集Flight和Hospital上分别实验，首先由RAHA标记脏数据，再由SAINT预训练、微调完成多分类任务。Flight数据集相较于Hospital，脏数据比例更高。预训练CutMix参数为0.5，Mixup参数为0.3。预训练共300轮，微调共100轮。对比SAINT on clean data, SAINT on dirty data, RAHA-SAINT on dirty data。结果可见，在含有脏数据的数据集上，SAINT本身效果受损严重；在脏数据比例高的数据集上，Model 3相较于SAINT本身效果有所提升。

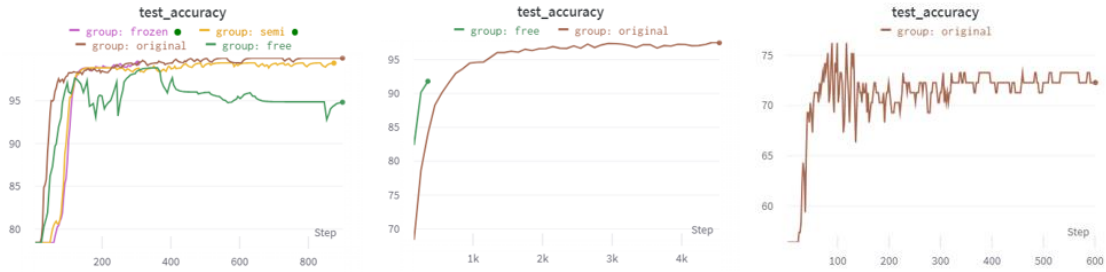
Flight Model	Raha + SAINT Clean	Raha + SAINT Dirty	Raha + New-SAINT Dirty
Pretrain Loss (init)	223.7497139	251.7421341	251.4068947
Pretrain Loss (final)	135.5933528	143.5876174	106.1310303
Accuracy (25)	21.739	9.855	11.884
Accuracy (50)	66.087	17.101	20.580
Accuracy (75)	84.348	24.058	28.696
Accuracy (best)	87.246	28.116	30.435

Hospital Model	Raha + SAINT Clean	Raha + SAINT Dirty	Raha + New-SAINT Dirty
Pretrain Loss (init)	167.2853813	200.1712112	201.1499634
Pretrain Loss (final)	85.5215111	92.7199440	94.9930801
Accuracy (25)	50.658	43.421	37.500
Accuracy (50)	82.237	63.158	48.684
Accuracy (75)	87.500	73.026	51.316
Accuracy (best)	89.474	75.658	56.579

此模型提议创新优化了表格数据表征学习的预训练任务，为SAINT预训练引入RAHA带来的数据干净与否的信息，引导SAINT着重学习干净与脏数据混杂的数据表征，既避免了脏数据对表格表征的消极影响，也避免了脏数据中有用信息被一概抹去。

3.4 结合预训练语言大模型的表征方法

由于预训练语言模型训练使用的文本数据通常具有较高的质量与较全面的知识范围，已有许多研究发现将其部分参数被保留下来并以迁移学习的形式运用到文本数据表征以外的领域时能起到非常好的效果，甚至一度优于若干领域原本的最优方法的表现 [8]。我们此次项目最终的想法便是由此出发，旨在运用具有较大量参数的优质预训练语言模型来进行表格数据的表征。结合计算资源与模型性能进行综合考量后，我们最终选择了标准大小的BERT模型的参数，选取了OpenML中特征迥异的三个数据集进行针对表格数据分类任务的微调，取得了如下实验结果：



上图中从左往右数据集的特征依次为小样本量小特征量、大样本量小特征量及大样本量大特征量；各图中褐色的曲线均代表未经预训练的SAINT模型的表现。可以看出在样本量较小时，采用预训练语言模型的完全微调会导致明显的过拟合，即绿色曲线最终随着训练波动下降；而第二张图显示在样本量较大时，采用微调预训练语言模型参数的方式则明显地提升了模型的拟合速度与精度。由于第三份数据集的 特征过多及计算资源的限制，我们的方法暂未能实现在第三份数据集上的实验；可以看出由于特征空间过于复杂，未经预训练的SAINT的表现并不理想。我们小组认为结合预训练语言模型的表征方法在对此类较复杂表格数据的分析表现上应能体现出绝对的优势，这一点仍有待在计算资源更加充裕时验证。

4 结论

通过比较以上几种我们搭建的模型与初始源模型的作用效果，分别比对其产出结果后，我们总结发现以上的模型中，除了RAHA + New-SAINT Pretrain模型之外的其他经典模型相比于原始的SAINT模型的效果提升十分有限，有时甚至会出现过拟合导致的负优化。我们总结的结论是，针对“利用类语言模型架构实现表格数据的清洗”这一任务有相当的实现难度：在探查阶段虽然往往能有较好的效果，但是在修正阶段的准确率与稳定性很难得到提升，尤其在GPT-4等生成式大语言模型增加了对于表格数据输入的处理方式的当下，利用其它规模较小的类语言模型相对更缺少竞争力。

5 项目展望

5.1 Model 4: Error-guided SAINT Pretrain (in progress)

基于这一结论，在学长的指导下我们提出了关于Model 4的设想：相比于探查并修正所有的脏数据，我们将目标先定位在利用对于脏数据的表征来实现“1+1>2”的效果，也就是通过表征来定向创造更适合在下游任务中使用的“脏数据”;这些数据正确率显然不能达到“数据清洗”的标准，但是经过类语言模型对于其中数据的学习以及分析，这些定向表征的脏数据也许相比于完全正确的干净数据更适合特定下游任务模型的学习。同时，根据下游模型学习结果的反馈，Model 4应该做到可以随时人工地对模型参数进行调整，同时支持在下游接入微调（fine tune）模型来进一步调整输出数据。

目前的构想中，我们决定学习SAINT模型的结构来实现对于数据的表征，利用RAHA模型对于脏数据的识别以及对于脏数据的信息学习能力来实现对于目标表征方向的学习。此外，我们也在考虑通过接入LLaMa等具有更多参数、表现更为优异的大语言模型来提升该架构的学习能力。

5.2 项目进程

总的来说，本次暑研，我们小组在苗晓晔老师的带领下、吴洋洋学长、倪炜学长及浙江大学数据科学研究中心的帮助下顺利完成“基于大型语言模型的数据清洗增强框架”暑研项目。在项目中，我们进行论文调研，学习数据清洗、数据挖掘、深度学习、大模型相关的前沿技术，测试模型源码，讨论、设计、改进模型方案，提出RAHA+SAINT, RAHA+BARAN+SAINT, RAHA+New-SAINT Pretrain, Error-guided SAINT Pretrain等表格

型数据表征学习模型。经过实验评估，验证了模型提议的可行性。在老师、学长的指点下发现模型存在的问题，明确了下一步模型改进的方向；更具体地，模型提议3仍然存在不足，改进方案的讨论与更大规模的实验需要时间进一步完成；模型提议4仍在代码实现阶段，Error-guided SAINT Pretrain具备较高的理论可行性，值得我们进一步深入研究。

参考文献

[1] V. Borisov et al, Deep Neural Networks and Tabular Data: A Survey, 2022.

[2] M. Mahdavi et al, Raha: A Configuration-Free Error Detection System, 2019, International Conference on Management of Data (SIGMOD'19) https://doi.org/10.1145/3299869.3324956.

[3] M. Mahdavi, Z. Abedjan, Baran: Effective Error Correction via a Unified Context Representation and Transfer Learning, 2020, PVLDB, 13(11): 1948-1961, DOI: https://doi.org/10.14778/3407790.3407801.

[4] A. Vaswani et al, Attention Is All You Need, 2017, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA

[5] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, arXiv:1810.04805v2 [cs.CL] 24 May 2019.

[6] S. Arik, T. Pfister, TabNet: Attentive Interpretable Tabular Learning, 2020, arXiv: 1908.07442v5 [cs.LG] 9 Dec 2020.

[7] G. Somepalli et al, SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training, 2021, arXiv:2106.01342v1 [cs.LG] 2 Jun 2021.

[8] T Zhou et al, One Fits All: Power General Time Series Analysis by Pretrained LM, 2023, arXiv:2302.11939v4 [cs.LG] 25 May 2023