TabContrast: A Local-Global Level Method for Tabular Contrastive Learning

Hao Liu[1] **Yixin Chen**[1] **Bradley Fritz**[1] **Christopher King**[3]
{liuhao, ychen25, bafritz, christopherking}@wustl.edu,
[1]Washington University in St. Louis

Abstract

Representation learning is a cornerstone of contemporary artificial intelligence, significantly boosting performance across diverse downstream tasks. Notably, domains like computer vision and NLP have witnessed transformative advancements owing to self-supervised contrastive learning techniques. Yet, the translation of these techniques to tabular data remains an intricate challenge. Traditional approaches, especially within the tabular arena, tend to explore model architecture and loss function design, often overlooking the nuanced creation of positive and negative sample pairs. These pairs are vital, shaping the quality of the learned representations and the overall model efficacy. Recognizing this imperative, our paper probes the specificities of tabular data and the unique challenges it presents. As a solution, we introduce "TabContrast". This method adopts a local-global contrast approach, segmenting features into subsets and subsequently performing tailored clustering to unveil inherent data patterns. By aligning samples with cluster centroids and emphasizing clear semantic distinctions, TabContrast promises enhanced representation efficacy. Preliminary evaluations highlight its potential, particularly in tabular datasets with more features available.

1 Introduction

Representation learning has become the backbone of most modern AI systems. Effective pretrained representations are crucial for enhancing downstream task performance. With the challenges of obtaining labeled data, self-supervised learning, particularly contrastive learning, has gained traction. This method has often demonstrated performance on par with or superior to supervised methods in domains like computer vision [1–3] and natural language processing (NLP)[4].

The essence of contrastive learning lies in crafting a discriminative embedding space. This is predominantly achieved by amplifying the similarity of positive pairs and emphasizing the dissimilarity of negative pairs. Positive pairs are usually different augmented views of the same sample, while negative pairs consist of distinct samples. However, the construction of different views involves domain-specific augmentation techniques, for example, rotating, jigsaw puzzle in image domain [3], and token masking in NLP [4], which contribute to learning perturbation-invariant representations.

While there is significant research on contrastive learning in the tabular domain, the focus has largely been on model architecture and loss function design. The generation of positive and negative samples has received less attention. Current methods typically take a local-local level contrast [5, 6]: corrupting some features of a sample to create a positive pair and using distinct samples as negatives, often neglecting the interplay between different samples.

In this paper, we introduce TabContrast, a method that employs a local-global level contrast approach for tabular contrastive learning. We segment features into random subsets and then cluster within each subset to identify feature-aligned groupings. Positive pairs are obtained by aligning a sample

Table Representation Learning Workshop at NeurIPS 2023.

with the centroids of its groups across all partitions, while negative pairs differentiate the sample from those outside its affiliated clusters. By harnessing inter-sample relationships based on shared feature subsets, TabContrast not only enhances the robustness of representations but effectively reduces false negatives. Our evaluations on three tabular datasets highlight its advantages, especially when handling a larger number of features.

2 Related Work

Self-supervised Tabular Representation Learning. Self-supervised learning aims at utilizing unlabelled data and learning the invariant information to generate discriminative embeddings. Recently, SSL has been adopted in the tabular domain. VIME [7] uses another sample to perturb the original sample and proposes to use reconstruction loss for recovering perturbed features and estimating the mask vector. SubTab [8] reconstructs input in an autoencoder using a subset of its features, and it corrupts the input by adding noise, randomly masking, or randomly replacing the feature with another sample. SCARF [5] is a contrastive learning method that takes the InfoNCE loss, where the positive sample is obtained by the same method as VIME. STab [9] proposes not to implement data augmentation on the features, and apply augmentation to each layer of the encoders. SAINT [6] not only mixes the original feature but also implements the mix-up operation in the latent space to get the augmented embeddings. TransTab [10] is a method that can fit tables with different features, and it regards different feature subsets of a single sample as positive pairs and others as negatives. Most works focus more on designing the loss function or the model framework, and the positive pairs are always constructed by the local-local method. None of the current methods consider the interplay between samples for the construction of positive and negative pairs, which motivates us to propose TabContrast.

3 Local-Global Method for Positive and Negative Pair Construction

In this section, we lay down the foundational notations and preliminaries and then delve deep into the specifics of our method: TabContrast.

Given a tabular dataset represented as $D = \{x_i, y_i\}_{i=1}^N$, each x_i stands as a d-dimensional feature vector with y_i as its associated label, and N as the total sample count. The feature set is expressed as $F = \{f_1, \cdots, f_d\}$.

In this work, our goal is to generate positive and negative pairs taking advantage of the interaction between samples. Drawing inspiration from the image domain, where positive pairs should be semantically similar and invariant to perturbations like rotations or cropping, our method translates these principles to the tabular domain. In the context of tabular data, this invariance translates to ensuring that positive pairs, despite perturbations or modifications in features, still fundamentally convey the same semantic information.

Our method, TabContrast, employs a strategy of segmenting features into random subsets and subsequently clustering within these subsets. This process is aimed at unveiling inherent groupings or patterns in the data, akin to creating perturbation-invariant views of the data in the image domain. Positive pairs are devised by aligning individual samples with the centroids of their respective clusters, preserving the core semantic relationship even amidst variations in specific feature values. For negatives, rather than random pairings, we specifically pair samples with those from different clusters, ensuring true semantic distinctions.

We now describe our method, which is illustrated in Figure 1. For each batch of samples, we first randomly partition the feature set F into k non-overlapped subsets $F_1, \cdots, F_k$, such that $F = \cup_{j=1}^k F_j$ and $F_{j1} \cap F_{j2} = \emptyset$ for each pair of (j_1, j_2). Then, we perform clustering according to each feature subset F_j, and will obtain clustering groups $G = \{G_1, \cdots, G_k\}$, where G_j is the set clustering results from j-th feature subset. We denote the i-th group in G_j as g_j^i, and denote the clustering center of g_j^i as $c_j^i = MEAN(x)$ for all $x \in g_j^i$.

Then we introduce the construction of positive and negative pairs. For a given sample x_i, suppose the groups it belongs to are denoted as $\{g_j^p\}_{j=1}^k$. Formulating positive pairs for a given sample x_i is achieved by **local-global matching**: maximizing its alignment with the clustering centers of its

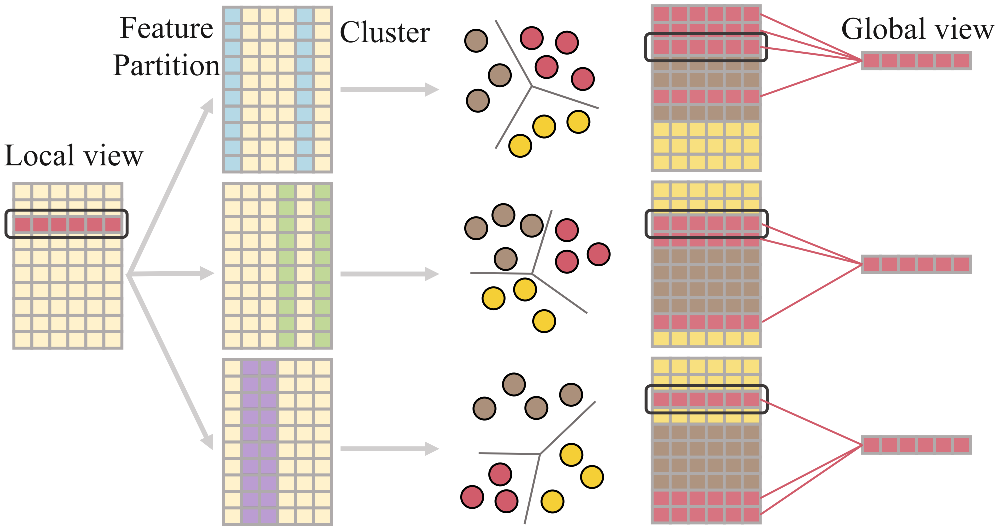

Figure 1: An overview of the TabContrast Framework. Given a table with rows representing samples and columns representing features, randomly implement feature partition. Then clustering on the selected features. Using the mean value to represent the embedding of the cluster center the current sample belongs to. Finally, construct positive pairs using local and global views.

affiliated clusters, thus deriving k positive pairs per sample, denoted as:

$$P = \{(x_i, c_j^p) | j \in (1, k), x_i \in g_j^p\}. \tag{1}$$

To construct negative pairs, in order to avoid including false negatives, we propose to exclude the samples that are highly correlated with the sample x_i instead of using all samples in the batch. Given the possible variation in clustering outcomes across feature subsets, samples cohabitating in at least $\lfloor \frac{k}{2} \rfloor$ groups with x_i are considered highly correlated. Denote the set of these highly correlated samples as X_p and the entire samples in the current batch as X_b, the negative pairs are represented as:

$$N = \{(x_i, x_n) | x_n \in X_b \backslash X_p\}. \tag{2}$$

Our method of constructing positive and negative pairs is flexible to any contrastive learning method in the tabular domain. Currently, we follow the framework of SCARF [5] and use the same InfoNCE loss.

4 Experiment

In this section, we present experiments on two public datasets as well as one local dataset to demonstrate the effectiveness of TabContrast.

4.1 Datasets, Setup, and Baselines

Datasets. We conducted our experiments on three tabular benchmark datasets: UCI adult income (Income) [11], MNIST, and an electronic health record (EHR) dataset from a local hospital. The Income dataset includes 32,561 samples with 14 features; the MNIST dataset includes 60,000 samples with 784 features. For the EHR dataset, we regard mortality in thirty days as a binary label, and this dataset has 89,246 samples with 190 columns.

Baselines. To evaluate the performance of using TabContrast to generate positive and negative pairs, we substitute the corresponding part in SCARF with our method and compare it with several baselines. We include some traditional machine learning baselines, such as Logistic Regression (LR), XGBoost, and Random Forest. We also include self-supervised baselines including VIME-self [7], SCARF [5], and SubTab [8].

Table 1: Results on Income, MNIST, EHR datasets. (Top rows) Traditional Machine Learning. (Middle rows) Self-supervised Learning. (Bottom row) TabContrast (our method) and the improvement ratio over SCARF. Evaluation metrics were scaled to 100 for readability purposes. In bold are methods with the best results for each task. In blue are methods with the best results in each group.

Dataset	Income (Acc↑)	MNIST (Acc↑)	EHR (AUC↑)
Maching Learning			
Logistic Regression	83.57 ± 0.06	92.45 ± 0.06	93.13 ± 0.09
Random Forest	83.51 ± 0.04	93.18 ± 0.11	92.45 ± 0.13
XGBoost	84.67 ± 0.17	**96.07 ± 0.09**	93.58 ± 0.18
Self-supervised Learning			
VIME [7]	84.83 ± 0.18	93.69 ± 0.25	92.47 ± 0.21
SubTab [8]	85.06 ± 0.12	95.74 ± 0.18	**94.56 ± 0.23**
SCARF [5]	85.24 ± 0.14	95.17 ± 0.22	92.30 ± 0.16
TabContrast (ours)	**85.33 ± 0.19**	95.91 ± 0.27	94.25 ± 0.08
improv. over SCARF	0.11%	0.78%	2.11%

Implementation Details. We utilized a four-layer multilayer perception (MLP) encoder, and a two-layer MLP as the projection head. We report mean accuracy over ten experiments for Income and MNIST datasets and report AUC for the EHR dataset due to imbalanced labels. We use k-means as the clustering method and set the number of clusters to be 3, 10, and 5 for each dataset. For the number of feature subsets k, we set k as 3, 10, and 5 for these three datasets respectively.

4.2 Main Results

We present the results in Table 1. From the results, we can conclude that XGBoost always achieves the best performance in the traditional machine learning group. Methods from self-supervised learning group can achieve comparable or better performance than XGBoost. Our method TabContrast, compared with SCARF, has improvement in all three tasks, indicating the proposed method of generating positive and negative samples is effective. Segmenting features into random subsets and clustering within the subsets might help the model understand intrinsic relationships between samples. The improvement is more significant in MNIST and EHR datasets. One reason might be the dimensions of features on these two datasets are larger than those of the Income dataset, thus the clustering results will be more diverse and the inherent representation of each different feature subset can be more robust.

5 Conclusion

In this study, we delved into the challenges of representation learning for tabular data, which stands in contrast to more structured domains such as images and language. We pay attention to the underexplored direction of constructing positive and negative samples. Through our global-local level TabContrast method, we have demonstrated that by segmenting features and implementing strategic clustering, contrastive learning models can be effectively trained to understand the intricate relationships among samples in a tabular dataset. Notably, our approach surpassed the performance of most baseline models, shedding light on the potential benefits of our positive and negative sample generation techniques.

Looking ahead, several promising avenues emerge for further refining and expanding upon our work: 1) exploring the implementation on many other tabular contrastive learning works (rather than only on SCARF) would provide a more comprehensive evaluation of our method 2) the current approach relies on randomly segmenting features, and investigating feature segmentation potentially based on feature correlation or importance could lead to more nuanced representations.

References

[1] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[5] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv preprint arXiv:2106.15147*, 2021.

[6] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[7] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020.

[8] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021.

[9] Ehsan Hajiramezanali, Nathaniel Lee Diamant, Gabriele Scalia, and Max W Shen. Stab: Self-supervised learning for tabular data. In *NeurIPS 2022 First Table Representation Workshop*, 2022.

[10] Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. *Advances in Neural Information Processing Systems*, 35:2902–2915, 2022.

[11] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.