Contrastive clustering of tabular data

Piotr Przemielewski [1] **Witold Wydmański** [1] **Marek Śmieja** [1]

Abstract

Contrastive self-supervised learning has significantly improved the performance of deep learning methods, such as representation learning and clustering. However, due to their dependence on data augmentation, these methods are mostly utilized in computer vision. In this paper, we investigate the adaptation of the recent contrastive clustering approach in the case of tabular data. Our experiments show that it outperforms typical clustering methods applicable to tabular data in most cases. Our findings affirm the potential adaptability of successful contrastive clustering techniques from other fields, such as image processing, to the realm of tabular data.

1. Introduction

Contrastive self-supervised models have emerged as one of the most significant topics in the field of deep learning. They have demonstrated remarkable performance in various computer vision tasks, ranging from representation learning to clustering. However, the efficacy of these models relies heavily on data augmentation, i.e., applying transformations that do not alter the class labels in the target task.

The task of data augmentation can be trivial if we have a good understanding of the domain we are traversing - real-world images can utilize a set of operations such as cropping, rotation, adding Gaussian noise, or changing the contrast. Text data makes use of transformations such as the introduction of typos, or artificial mistranslations (Dhole et al., 2021), and metagenomic data can benefit from introducing simulated errors in reads (Sayyari et al., 2019).

However, applying contrastive methods to domain-agnostic tabular data presents unique challenges due to the lack of a regular internal structure and pre-defined relationships such as spatial dependencies. Consequently, defining suitable augmentations for contrastive methods can be difficult in the context of tabular datasets. In prior research, approaches such as SubTab (Ucar et al., 2021) and LoCL (Gharibshah & Zhu, 2022) have attempted to address this issue.

Drawing on new ideas of contrastive learning in computer vision, we present the application of contrastive clustering methods for tabular data. Specifically, we examine three types of augmentation techniques and assess their performance on four different datasets. Our findings contribute to the development of more efficient and versatile contrastive learning methods for a wider variety of data types.

2. Method

Our approach combines the elements of the "Contrastive Clustering" framework (Li et al., 2021) with data augmentation techniques inspired by the "SubTab" method. These modifications are designed to address the unique challenges posed by tabular data.

Contrastive Clustering Base The fundamental structure of our model is built on the Contrastive Clustering technique, utilizing all of its three main components - the backbone network, the instance-level contrastive head, and the cluster-level contrastive head.

A feature extractor $f : \mathbb{R}^D \to \mathbb{R}^N$, represented by backbone component, returns the representation $h = f(x)$ for a given input x. This latent vector, h, is subsequently passed through two projection heads. NTX-ent loss applied to embeddings generated by the instance-level projection head g_I encourages the feature extractor to generate representation invariant to data augmentations (Sohn, 2016). On the other hand, the cluster-level projection head, g_C, yields an output vector y with dimensions equivalent to the number of clusters, as outcome of the softmax activation function. Similar to the instance-level, the NTXent loss function is used to calculate the distances between the clusters.

The instance-level contrastive head distinguishes between feature vectors (row-wise perspective) of different instances of data, while the cluster-level contrastive head identifies

[1]Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, Poland.

The research of M. Śmieja was supported by the National Science Centre (Poland), grant no. 2022/45/B/ST6/01117. Correspondence to: Piotr Przemielewski <piotr.przemielewski@student.uj.edu.pl>.

Proceedings of the Data-centric Machine Learning Research (DMLR) Workshop at ICML 2023.

and differentiates between different clusters (column-wise perspective) in the dataset. This approach enables the model to both recognize individual instances and understand the broader context of clusters within the dataset.

Data Augmentation for Tabular Data To adapt contrastive clustering to the case of tabular data, we propose to use the following augmentations:

- **Gaussian noise**, generated from a standard normal distribution, is incorporated into the original data. This method involves the addition of randomly selected values derived from this distribution. Standard deviation is set through a process of hyperparameter tuning.
- **Swap** introduces permutations within the dataset's features. Specifically, values from one data instance are interchanged with values from another instance, thereby changing their original arrangement in a given feature column
- **Zero** is introduced by applying a dropout operation to the initial layer of the backbone network. This technique involves zeroing out a fraction of the input features.

The degree of data perturbation introduced through the noise is determined through a hyperparameter tuning process. This augmentation strategy allows us to create a "corrupted" version of our dataset for the training phase.

3. Experiments

Datasets For our experimental evaluation, we utilized four datasets:

- **MNIST** dataset includes 70,000 images of handwritten digits, transformed into 784-dimensional vectors.
- **BreastCancer** dataset contains health data from 569 patients, with 30 factors indicating breast cancer diagnoses.
- **Reuters-10k** dataset consists of 10,000 news stories, each described using the 2000 most frequent words, grouped into four distinct classes.
- **Letter** holds data on 20,000 handwritten letters, categorized into 26 classes using 16 features.

Architecture A framework consists of three components: instance-level head, cluster-level head, and backbone. The backbone architectures were individually optimized for each dataset. Each contains up to three fully connected layers with a maximum of 512 neurons. The architecture for both the instance-level and cluster-level heads remained consistent across all experiments.

Hyperparameters including the masking ratio, projection size, batch size, and learning rate were identified through grid search. AdamW optimizer and LeakyReLU activation function were used throughout the entire model.

Table 1. Comparison of our methods (swap, Gauss, zero) with baseline models.

Dataset		kmeans	SubTab	IDEC	Swap	Gauss	Zero
MNIST	ACC	54.6	42.0	**88.1**	80.3±4.4	79.5±7.8	85.5±2.2
	NMI	50.9	45.7	**86.7**	80.1±3.7	78.7±5.2	82.8±2.4
Breast	ACC	90.2	89.5	**92.6**	88.6±1.2	92.3±1.8	92.1±0.6
	NMI	56.2	50.6	63.8	53.6±2.7	**78.7±5.2**	59.1±2.3
R10k	ACC	72.9	71.4	**75.6**	72.8±3.3	66.9±3.9	63.0±4.6
	NMI	48.8	46.9	49.8	**55.9±4.6**	42.6±2.4	45.0±7.2
Letter	ACC	25.9	19.4	-	21.8±1.1	**28.2±1.9**	18.4±1.1
	NMI	35.7	29.1	-	30.1±1.5	**42.1±0.9**	26.6±1.6

Baseline models To evaluate the efficacy of our proposed method, we selected three benchmark models as comparison baselines. The first baseline is the classic k-means clustering algorithm applied directly on the raw data, thus representing a traditional unsupervised learning method. The second baseline utilizes the SubTab framework; however, to maintain consistency with our self-supervised learning, we conducted an evaluation using k-means on the latent representations derived from SubTab. The final baseline employs the IDEC methodology (Guo et al., 2017).

Metrics For evaluating our methods, we used two widely used metrics in unsupervised learning settings: Accuracy (ACC) and Normalized Mutual Information (NMI). Higher values of these metrics indicate better clustering performance.

Results The results obtained by our model are presented in Table 1. The proposed model demonstrates superior performance compared to k-means and SubTab with k-means. The effectiveness of our model parallels that of the IDEC approach and provides better results in terms of NMI in 2 out of 3 datasets. It is noteworthy that the SubTab demonstrated inferior performance compared to the k-means approach, despite its use of contrastive learning. This observation highlights the efficacy of our proposed method, suggesting a potential for constructing meaningful representations. Another important observation is that there is no universally superior data augmentation technique. The performance is significantly influenced by the nature of the dataset.

Conclusion and future work Our research initiates the use of contrastive clustering for tabular data, underlining the necessity of dataset-specific data augmentation strategies and hyperparameter tuning. Drawing inspiration from the achievements of similar techniques in computer vision, we demonstrate that this approach holds promising potential for other domains, including tabular data, and encourages further advancements.

References

Dhole, K. D., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., Mahendiran, A., Mille, S., Shrivastava, A., Tan, S., et al. Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*, 2021.

Gharibshah, Z. and Zhu, X. Local contrastive feature learning for tabular data. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 3963–3967, 2022.

Guo, X., Gao, L., Liu, X., and Yin, J. Improved deep embedded clustering with local structure preservation. In *Ijcai*, pp. 1753–1759, 2017.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J. T., and Peng, X. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8547–8555, 2021.

Sayyari, E., Kawas, B., and Mirarab, S. Tada: Phylogenetic augmentation of microbiome samples enhances phenotype classification, Jul 2019.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Ucar, T., Hajiramezanali, E., and Edwards, L. Subtab: Subsetting features of tabular data for self-supervised representation learning, 2021.