ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Continual Contrastive Learning on Tabular Data with Out of Distribution

Achmad Ginanjar[a] *, Xue Li[a], Priyanka Singh[a], and Wen Hua[b]

[a] School of Electrical Engineering and Computer Science
The University of Queensland, Queensland, Australia
[b] Department of Computing, The Hong Kong Polytechnic University Hong Kong

Abstract. Out-of-distribution (OOD) prediction remains a significant challenge in machine learning, particularly for tabular data where traditional methods often fail to generalize beyond their training distribution. This paper introduces Tabular Continual Contrastive Learning (TCCL), a novel framework designed to address OOD challenges in tabular data processing. TCCL integrates contrastive learning principles with continual learning mechanisms, featuring a three-component architecture: an Encoder for data transformation, a Decoder for representation learning, and a Learner Head. We evaluate TCCL against 14 baseline models, including state-of-the-art deep learning approaches and gradient-boosted decision trees (GBDT), across eight diverse tabular datasets. Our experimental results demonstrate that TCCL consistently outperforms existing methods in both classification and regression tasks on OOD data, with particular strength in handling distribution shifts. These findings suggest that TCCL represents a significant advancement in handling OOD scenarios for tabular data.

1 Introduction

When a machine learning model encounters new data that is out of distribution (OOD), it may experience a significant decline in performance [1]. OOD refers to data samples that differ from the distribution of the training data, which can often lead to unreliable predictions [2]. Addressing the challenges associated with OOD is crucial for maintaining a model's performance. This highlights the significance of our research in identifying effective strategies for managing OOD scenarios.

Significant progress has been made in OOD detection using algorithms such as MCDD [3], OpenMax [4], and TemperatureScaling [5]. However, challenges remain for prediction tasks involving tabular data. Recent advancements in deep learning for tabular data show promise [6, 7, 8]; however, Gradient Boosted Decision Trees (GBDT) still tend to be the best option for tabular datasets [9]. On the other hand, the tree-based structure of GBDT can make it difficult to extrapolate beyond the training distribution[10].

In this study, we present Tabular Continual Contrastive Learning (TCCL), a method specifically designed to address OOD challenges in tabular data. TCCL adopts a similar contrastive learning framework [11, 12] , which involves an

*Indonesia Endowment Fund for Education Agency (LPDP)

encoder and a header. However, our approach introduces a novel header to handle new data with shifted distributions.

TCCL consists of three main components: an Encoder, a Decoder, and a Learner Head. The Encoder processes input data into an augmented format while the Decoder translates this encoded data into a new representation. The Learner Head is specifically designed to mitigate the issue of catastrophic forgetting by incorporating a mechanism that acts as a 'break,' effectively preventing the model from losing previously learned information. These features allow TCCL to effectively handle data with shifting distributions. Experimental results demonstrate that TCCL outperforms other models in both classification and regression tasks.

2 Related Work

Neural network models such as Multilayer Perceptron (MLP), Self-Normalizing Neural Networks (SNN), Feature Tokenizer Transformer /FT-Transformer, Residual Network /ResNet, Deep & Cross Network /DCN V2, Automatic Feature Interaction /AutoInt, Neural Oblivious Decision Ensembles / NODE, Tabular Network / TabNet and GrowNet are widely recognized and frequently cited for addressing tabular data prediction problems [13]. Another approach from the tabular data area is contrastive learning models such as CFL [12], SCARF [8], and SubTab [6]. However, these models are not specifically designed to handle OOD data. In our experiments, we utilized these methods as base models. Note that CFL is not applicable in non-federated learning networks. We are excluding CFL from our experiments.

3 Problem Formulation

In machine learning, the goal is to build a model $f : x \to y$ that generalizes well to unseen data. A prediction task can be defined as finding a model $f : (.)$ that minimizes the expected error over a dataset D_{in} with distribution p_{in}. This can be expressed in terms of a loss function $\textbf{min}\ \text{Error}(x, y)_D = [L(f(x), y)]$, which L measures the discrepancy between the model's predictions $f(x)$ and the true labels. A higher loss value indicates poorer model performance $E \uparrow = P \downarrow$. When a different distribution $D_{ood} \leftarrow p_{ood}$ is introduced to a model, its performance typically decreases $P(f((x)_{D_{in}})) > P(f((x)_{D_{ood}}))$ [1].

4 Proposed Method

We introduce Tabular Contrastive Learning (TCL), an improved approach designed to enhance prediction tasks on tabular data with OOD.

4.1 TCCL Main Architecture

TCCL consist of a tabular contrastive learning model M^a, a fisher matrix, a learner header and an updated continual model M^b , see figure1. TCCL starts

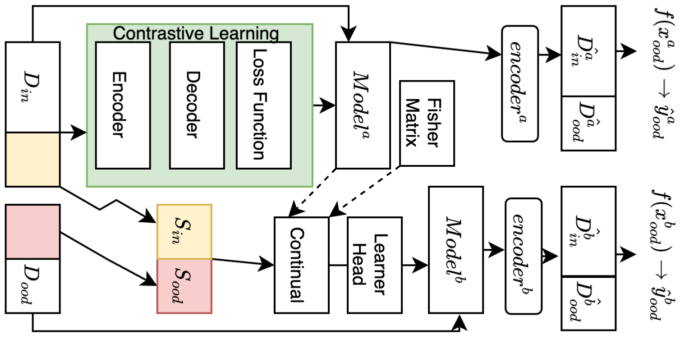

Fig. 1: TCCL architecture. TCCL consist of Encoder, Decoder and a Learner Header

with data D_{in} is trained within tabular contrastive learning M^a to produce encoder E^a . E^a then used to generate new $\hat{D}^a_{in}$ and $\hat{D}^a_{ood} = X^a_{ood}$. Based on the new data, a prediction model $f(x^a_{ood}) \rightarrow \hat{y}^a_{ood}$ is calculated, where $\hat{y}^a_{ood}$ the predicted label from OOD data is based on the model trained from $\hat{D}^a_{in}$. To maintain performance, continual learning is done. M^a is retrained with weight from fisher matrix F. M^a is retrained with $S_{in} \cap D_{in} + S_{ood} \cap D_{ood}$. T stop catastrophic forgetting a learner header is used to stop model learning from new data. This is done by setting weight in M near zero during training. The final result is $f(x^b_{ood}) \rightarrow \hat{y}^b_{ood}$.

5 Experiment

5.1 Datasets

We utilize 8 diverse tabular datasets and 14 models (including TCL) in our experiment. The datasets are Adult [14], Helena [15], Jannis [15], Higgs Small [16], Aloi [17], Epsilon [18], Cover Type [19], California Housing [20], Year [21], Yahoo [22], and Microsoft [23].

5.2 OOD Detection

We have implemented two OOD detection methods, namely OpenMax [4] and TemperatureScaling [5]. We separated the OOD data using these detectors to generate two sets D_{in} and D_{ood} where $(D_{in} + D_{ood} = D)$. While D_{in} it is used for training datasets, D_{ood} is used for test datasets.

5.3 Prediction ON OOD Dataset

We experiment with 7 recent models for tabular data e.g. FT-T, DCN2, GrowNet, ResNet, MLP, AutoInt, and TabR-MLP. In addition, we experimented with the recent implementation of contrastive learning for tabular data SubTab [6] and

Table 1: OOD experiment settings. Train data is the D_{in}, and test data is the D_{ood}.

Dataset	OOD Detector	Norm	Train (In)	Test (OOD)
Adult	OpenMax	l2	31820	9067
Helena	OpenMax	l1	41724	13040
Aloi	OpenMax	l1	85982	522
Covtype	TemperatureScaling	l1	464304	631
California	OpenMax	l1	15665	1058
Year	OpenMax	l2	370972	51630
Yahoo	TemperatureScaling	l1	473134	165660
Microsoft	TemperatureScaling	l1	957079	3843

Table 2: Experiment result. F1 score for classification and RMSE for regression. Datasets with (*) mean a regression problem. Models (^{c}) are contrastive learning based model, and models (^{x}) are GBDT based model.)

	AD↑	HE↑	AL↑	CO↑	CA*↓	YE*↓	YA*↓	MI*↓
FT-T	0.782	0.153	0.407	-	0.867	6.461	-	-
DCN2	0.744	0.129	0.414	0.58	2.602	7.054	0.645	0.746
GrowNet	0.465	-	-	-	0.969	7.605	1.01	0.769
ResNet	0.652	0.10	0.437	0.694	0.892	6.496	0.639	0.736
MLP	0.508	0.146	0.326	0.617	0.894	6.488	0.657	0.741
AutoInt	0.78	0.133	0.401	0.608	0.89	6.673	-	0.739
TabR-MLP	0.688	0.165	0.429	0.688	2.677	2e5	1.285	0.79
TCCL^c	0.861	**0.236**	**0.510**	**0.972**	**0.745**	**6.329**	**0.636**	**0.734**
Scarf^c	0.720	0.00	0.00	0.091	-	-	-	-
SubTab^c	0.714	0.146	0.322	0.59	1.012	6.668	0.656	0.744
Lightgbm^x	0.591	0.080	0.177	0.219	0.848	6.565	0.661	0.740
CatBoost^x	**0.927**	0.152	-	0.753	0.827	6.622	0.655	0.733
XGB^x	0.925	0.127	0.328	0.700	0.845	6.867	0.654	0.739

SCARF [8], which comes from a similar domain to our TCL. In addition, we add Lightgbm, CatBoost, and XGB from GBDT models to our base models.

6 Result and Evaluation

6.1 OOD Detection

Table 1 shows the experiment results on splitting the dataset to in distribution D_{in} and OOD data D_{ood}. In the table, we can evaluate the OOD detector and normalisation used during the experiment. Train data is the D_{in} and test data is the D_{ood} . Our experiments are evaluated based on these settings.

6.2 Model Performance

Table 2 shows the results of our experiments. Deep learning models like FT-T, DCN2, ResNet, MLP, AutoInt, and TabR-MLP showed varied performance across datasets, with notable struggles such as GrowNet's poor performance on the Adult dataset (0.465) and DCN2's high RMSE (2.602) on California Housing. TCCL contrastive learning models outperformed competitors SCARF and SubTab, achieving impressive results, including a 0.972 F1 score on Covtype and low RMSE scores on various regression tasks. Among GBDT models, CatBoost and XGBoost performed exceptionally well, particularly on the Adult dataset (F1 scores: 0.927 and 0.925), while LightGBM generally underperformed compared to its GBDT counterparts.

TCCL exhibits superior generalization capabilities in both classification and regression tasks, effectively handling out-of-distribution scenarios in various tabular data contexts. However, traditional GBDT models remain competitive, particularly excelling in specific datasets such as the Adult dataset, where they outperform TCCL. Despite this exception, TCCL consistently demonstrates strong performance across diverse tasks, indicating that its architecture is well-suited to addressing the challenges posed by distribution shifts in tabular data.

7 Conclusion

While out-of-distribution presents significant challenges to machine learning, TCCL demonstrates promising results in handling this type of data. TCCL performs well on most datasets, except for the Adult dataset, where the GBDT model outperforms it. However, in other datasets, TCCL surpasses both deep learning models and GBDT. GBDT ranks as the second-best model when it comes to out-of-distribution scenarios. Although TCCL shows strong results, future studies should assess its robustness by testing it on a wider range of datasets.

References

[1] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data, 2020.

[2] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 10 2016.

[3] Dongha Lee, Sehun Yu, and Hwanjo Yu. Multi-class data description for out-of-distribution detection, 2020.

[4] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:1563–1572, 12 2016.

[5] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[6] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. volume 23, 2021.

[7] Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. Tabr: Tabular deep learning meets nearest neighbors in 2023, 2023.

[8] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. SCARF: SELF-SUPERVISED CONTRASTIVE LEARNING USING RANDOM FEATURE CORRUPTION. In *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

[9] Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35:507–520, dec 2022.

[10] Hassan Mesghali, Behnam Akhlaghi, Nima Gozalpour, Javad Mohammadpour, Fatemeh Salehi, and Rouzbeh Abbassi. Predicting maximum pitting corrosion depth in buried transmission pipelines: Insights from tree-based machine learning and identification of influential factors. *Process Safety and Environmental Protection*, 187:1269–1285, 2024.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. pages 1597–1607, 11 2020.

[12] Achmad Ginanjar, Xue Li, and Wen Hua. Contrastive federated learning with tabular data silos, 2024.

[13] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 23:18932–18943, 6 2021.

[14] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[15] Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle Sebag, Alexander Statnikov, Wei-Wei Tu, and Evelyne Viegas. Analysis of the automl challenge series 2015-2018. In *AutoML*, Challenges in Machine Learning. Springer, 2019.

[16] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5:4308, 2014.

[17] Jan-Mark Geusebroek, Gertjan J. Burghouts, and Arnold W. M. Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

[18] PASCAL Challenge on Large Scale Learning. Epsilon Dataset: Simulated Physics Experiments. `http://largescale.ml.tu-berlin.de/instructions/`, 2008. Accessed: [Insert Access Date].

[19] Jock A. Blackard and Denis J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 2000.

[20] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

[21] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, pages 591–596, Miami, Florida, USA, October 2011.

[22] Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In *Proceedings of the Learning to Rank Challenge*, volume 14 of *Proceedings of Machine Learning Research*, pages 1–24. PMLR, 2011.

[23] Tao Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013.