Applications of Optimal Transport Distances in Unsupervised AutoML

Prabhant Singh, Joaquin Vanschoren
Department of Mathematics and Computer Science
Eindhoven University of Technology
p.singh@tue.nl, j.vanschoren@tue.nl

Abstract

In this work, we explore the utility of Optimal Transport-based dataset similarity to find similar *unlabeled tabular* datasets, especially in the context of automated machine learning (AutoML) on unsupervised tasks. Since unsupervised tasks don't have a ground truth that optimization techniques can optimize towards, but often do have historical information on which pipelines work best, we propose to meta-learn over prior tasks to transfer useful pipelines to new tasks. Our intuition behind this work is that pipelines that worked well on datasets with a *similar underlying data distribution* will work well on new datasets. We use Optimal Transport distances to find this similarity between unlabeled tabular datasets and recommend machine learning pipelines on two downstream unsupervised tasks: Outlier Detection and Clustering. We obtain very promising results against existing baselines and state-of-the-art methods.

1 Introduction

Automated Machine Learning (AutoML) aims to automate the design and optimization of machine learning pipelines in a data-driven way, using a variety of optimization techniques to find the best pipeline in a vast search space of possible pipelines consisting of many data preparation steps and modeling techniques. Since optimization techniques usually need ground-truth performance as a signal to guide the search, unsupervised tasks (e.g. clustering, outlier detection, and dimensionality reduction) have long eluded AutoML research, as the lack of labels often makes immediate objective evaluation impossible.

We propose a meta-learning framework for unsupervised machine learning that leverages optimal transport distances [17, 22] to recommend which unsupervised algorithms, preprocessing techniques, and hyperparameters to use based on how well they performed on prior tasks with similar data distributions. Such recommendations can be used as smart defaults, but also more generally to warm-start or reduce the search space of AutoML techniques, also on supervised tasks. In this work, we evaluate this approach for model selection on outlier detection and clustering tasks. The key contributions of this work are:

1. We propose LOTUS, a meta-learning technique based on finding similar datasets using Optimal Transport for unsupervised scenarios.

2. We perform extensive experiments, including strong baselines and existing state-of-the-art methods, demonstrating that our Optimal Transport-based approach yields significantly better results as well as novel insight.

Accepted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023.

2 Background and Related Work

Many AutoML [11] tools leverage meta-learning schemes [26] to find good configurations to warm-start optimization. For instance, AutoSklearn-2.0 [7] learns pipeline portfolios, FLAML [28] uses meta-learned defaults, and MetaBu [20] uses optimal transport (Fused Gromov Wasserstein with proximal gradients) to learn meta-features to find similar prior tasks. Moreover, several approaches have been proposed to perform AutoML on outlier detection and clustering tasks:

Outlier detection: MetaOD [30] uses collaborative filtering [24] to build a recommender system for predicting the best outlier detection techniques and leverage meta-learning based on landmark and model-based meta-features. PyODDS [12] is a related framework but it requires ground truth data to select specific OD techniques. One can argue that the use of internal metrics such as Excess-Mass [9], Mass-Volume [9], and IREOS [15] can be used instead. However, it has been shown that these internal metrics are computationally very expensive and do not scale well to large datasets [14].

Clustering: In clustering, a few studies have considered meta-learning for algorithm selection without hyperparameter optimization [18, 21]. [6, 23] require ground truth labels since they used external metrics during optimization. AutoML4Clust [25] and AutoClust [19] leverage Bayesian Optimization to select the best clustering algorithm and hyperparameters for a given dataset, but they do not support pipelines with data preprocessing steps. AutoCluster [13] uses meta-learning for creating ensembles of multiple clustering algorithms.[1][2].

3 Method

We introduce LOTUS, Learning to learn with Optimal Transport for Unsupervised Scenarios, which is summarized in Algorithms 1 and 2. LOTUS first meta-learns how well different unsupervised algorithms work on prior *labeled* datasets. These can be datasets where the correct labels are known or proxy tasks. More formally, we require:

- A collection of n prior labeled datasets $\mathcal{D}_{meta} = \{D_1, ..., D_n\}$ with train and test splits such that $D_i = (X_i^{train}, y_i^{train}), (X_i^{test}, y_i^{test})$.
- A collection of n optimized pipelines A_i^* with associated hyperparameters λ_i^* for every dataset in $\mathcal{D}_{meta}$; $\mathcal{A} = \{A_{\lambda_1^*}^*, ..., A_{\lambda_n^*}^*\}$

Given a new input dataset $D_{new} = (X_{new})$ without any labels, we aim to select a pipeline $A_{\lambda^*}^* \in \mathcal{A}$ to employ on X_{new}, where $A_{\lambda^*}^*$ is a tuned pipeline for a dataset similar to X_{new}.

Our premise is that, if a prior dataset exists that is very similar to the new dataset, then its optimal pipelines will likely work well on the new dataset. We consider two datasets similar if they have the same underlying data distribution, which we measure using (unsupervised) Optimal Transport.

We first require a transformation function ϕ to map the (tabular) dataset to a metric space. Next, we calculate the dataset similarity $\mathcal{O}$ based on some distance metric ψ in equation 1. We use Gromov-Wasserstein distance here. The Gromov-Wasserstein distance allows us to compute distances with samples that do not belong to the same metric space. As GW is expensive and we require computationally efficient similarity estimates, we adopt the Low-Rank Gromov-Wasserstein distance[22] LR on these transformed distributions, as summarized in equation 2, where r is the selected rank.

$$\mathcal{O} = \psi(\phi(D_a), \phi(D_b)) \tag{1}$$

$$\mathcal{O} = GW - LR^{(r)}(\phi(D_a), \phi(D_b)) \tag{2}$$

The most similar prior dataset $D_s \in \mathcal{D}_{meta}$ is the dataset with the smallest distance to the new dataset D_{new}. LOTUS then assigns the optimal configuration from $\mathcal{A}$: $A_{\lambda_{new}^*}^* = A_{\lambda_s^*}^*$ where $A_{\lambda_s^*}^*$ is predicted as the optimal configuration for D_{new}.

[1] We tried to add AutoCluster in our experiments but encountered errors that we could not resolve, and the authors didn't reply to our questions.

[2] None of the clustering works we studied were reproducible or workable without any errors, therefore we introduce our own baselines for automated clustering

Algorithm 1 Pseudocode for Meta-training

Inputs: $\mathcal{D}_{meta}, L, \boldsymbol{A}, \boldsymbol{\Lambda_A}$

while $D_i \in \mathcal{D}_{meta}$ **do**
  $A^*_{\lambda^* i} \leftarrow \text{argmin}_{\substack{\forall A^j \in \boldsymbol{A} \\ \forall \lambda \in \boldsymbol{\Lambda_A}}} L\left(A^j_\lambda, \{\boldsymbol{X}\}\{\boldsymbol{y}\}\right)$
  $\mathcal{A} \leftarrow A^*_{\lambda^* i}$
end while

Algorithm 2 Pseudocode for LOTUS (meta-testing)

Inputs: $D_{new}, \mathcal{D}_{meta}, \mathcal{A}$

while $D_i \in \mathcal{D}_{meta}$ **do**
  $\mathcal{O}_i \leftarrow GWLR(\phi(D_{new}, D_i))$ {Distance calculation}
end while
$s \leftarrow \text{argmin}\{\mathcal{O}_1, ..., \mathcal{O}_n\}$ {Retrieval of most similar dataset}
$A^*_{\lambda^*_{new}} \leftarrow A^*_{\lambda^*_s}$ {Model Selection}

To find $\mathcal{A}$ we develop two tools for each task (outlier detection and clustering) on top of the GAMA AutoML framework[8]. We construct proxy tasks for outlier detection by retrieving many extremely imbalanced classification tasks from OpenML [27] where examples of the smallest class are considered outliers. For clustering, we follow a similar approach, using many classification tasks and using each class as a cluster. We follow a zero-shot approach, i.e. we recommend the best pipeline based on prior knowledge, without using evaluations on the new dataset to choose the next pipeline to try.

4 Experiments

Outlier detection: To evaluate our method on outlier detection we use the ADBench [10] benchmark, which is a large and comprehensive set of datasets for benchmarking outlier detection algorithms. We compared our method with PyOD baselines[29] and MetaOD[30]. We use the area under the ROC curve (AUC) as the optimization metric L during the search phase.
Clustering: Due to a lack of standard benchmarks for clustering we extracted 57 clustering datasets from OpenML [27]. Though there are a lot of automated clustering methods available, during our experiments we found all of them non-reproducible(AutoCluster crashed, and others did not have open-source implementations and datasets). As baselines, we use all sci-kit learn clustering techniques, with Calinkski-Harbarnz as an internal metric, and optimize their hyperparameters using Adjuster Mutual Information (AMI) as the optimization metric L.

5 Results

We use the Bayesian Wilcoxon signed-rank test (or ROPE test [3, 4]) to analyze the results of our experiments. ROPE defines an interval wherein the differences in model performance are considered equivalent to the null value. Using this test allows us to compare model performances in a more practical sense. We set the ROPE value to 1% for our experiments and use the baycomp library [3] to run the analyses. We first compare the results with the state-of-the-art and then other strong baselines. Table 1 and Table 2 show that our method is significantly better than all existing baselines by a wide margin, which suggests that zero-shot meta-learning outperforms tuning unsupervised techniques with internal metrics. Moreover, since our method also outperforms MetaOD, we can conclude that OT-based distances work better than unsupervised meta-feature-based distances used in earlier work.

6 Conclusion, Discussion, and Limitations

In this work, we show a simple but very effective method to find similar unlabeled tabular datasets and use this to meta-learn pipelines for unsupervised model selection tasks. Our experiments show

Estimator	p(**LOTUS**)	p(**ROPE**)	p(**Estimator**)
MetaOD	**0.740**	0.074	0.186
ABOD	**1.0**	0.0	0.0
OCSVM	**1.0**	0.0	0.0
LODA	**1.0**	0.0	0.0
KNN	**1.0**	0.0	0.0
HBOS	**999.82**$\cdot 10^{-3}$	0.0	$0.18 \cdot 10^{-3}$
IForest	**999.54**$\cdot 10^{-3}$	0.0	$0.46 \cdot 10^{-3}$
COF	**1.0**	0.0	0.0
LOF	**1.0**	0.0	0.0

Table 1: Outlier detection results. Shown are ROPE testing results with LOTUS vs MetaOD and all PyOD baselines with ROPE=1%(Higher is better)

Estimator	p(**LOTUS**)	p(**ROPE**)	p(**Estimator**)
Internal	**1.0**	0.0	0.0
DBSCAN	**1.0**	0.0	0.0
OPTICS	**1.0**	0.0	0.0
Agg.Clustering	**1.0**	0.0	0.0
KMeans	**0.908**	0.003	0.089
MiniBatch Kmeans	**0.959**	0.0	0.040
BIRCH	**0.987**	0.0	0.013

Table 2: Clustering results. Shown are ROPE testing results with LOTUS vs. Sklearn Clustering baselines with ROPE=1%(Higher is better)

that optimal transport-based distances provide a very promising approach for automating tasks such as clustering and outlier detection.

6.1 Using Gromov-Wasserstein Distance as a Similarity Measure

Our experiments show that using optimal transport distances like GW-LR provided a feasible and robust approach for dataset similarity and meta-learning. We would like to emphasize that this similarity measure should only be used as a relative similarity measure. For instance, in our case, we use this similarity measure to find the most similar dataset from a collection of datasets in $\mathcal{D}_{meta}$. To estimate to what degree datasets are similar, Nies et al. [16] propose optimal transport-based correlation measures that could be leveraged as an alternative approach. Our approach assumes that Wasserstein distances can capture the intrinsic properties of datasets and can indicate the similarity between them. OTDD [1] also uses optimal transport distances with feature cost to provide a distance between datasets but this approach requires labels.

6.2 Limitations

Our framework has limitations that should be taken into consideration. First, our method's effectiveness depends on the presence of similar datasets to the meta-test dataset within $\mathcal{D}_{meta}$. In cases where there are no similar datasets, such as for automated clustering with dataset id *42464* in our experiments, our suggested pipeline may not yield favorable results. Second, the time complexity of our system scales linearly with the number of datasets in $\mathcal{D}_{meta}$. This means that as the number of datasets increases, LOTUS may require more time to perform model selection. These limitations are important to note as they may impact the practical application of our approach in certain scenarios.

6.3 Future Work

This work proposed a very effective way to automate clustering with zero-shot recommendation via an Optimal Transport-based distance function. However, this distance is still very expensive to compute. Although the low-rank computation decreases the time complexity, the system still takes around 30 minutes to compute similarities with other datasets. In future work, we aim to explore Wasserstein Embedding Networks [5] or MetaICNN [2] for faster computation, although these networks still do not (yet) support Gromov Wasserstein space.

References

[1] David Alvarez-Melis and Nicoló Fusi. Geometric dataset distances via optimal transport. *ArXiv*, abs/2002.02923, 2020.

[2] Brandon Amos, Samuel Cohen, Giulia Luise, and Ievgen Redko. Meta optimal transport. *ArXiv*, abs/2206.05262, 2022.

[3] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017.

[4] Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1026–1034, Bejing, China, 22–24 Jun 2014. PMLR.

[5] Nicolas Courty, Rémi Flamary, and Mélanie Ducoffe. Learning wasserstein embeddings. In *International Conference on Learning Representations*, 2018.

[6] Marcilio CP De Souto, Ricardo BC Prudencio, Rodrigo GF Soares, Daniel SA De Araujo, Ivan G Costa, Teresa B Ludermir, and Alexander Schliep. Ranking and selecting clustering algorithms using a meta-learning approach. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 3729–3735. IEEE, 2008.

[7] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. Auto-sklearn 2.0: Hands-free automl via meta-learning. *arXiv:2007.04074 [cs.LG]*, 2020.

[8] P. Gijsbers and J. Vanschoren. Gama: A general automated machine learning assistant. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12461 LNAI:560–564, 2021.

[9] Nicolas Goix. How to evaluate the quality of unsupervised anomaly detection algorithms?, 2016.

[10] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench: Anomaly detection benchmark. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[11] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. Automated machine learning: Methods, systems, challenges. *Automated Machine Learning*, 2019.

[12] Yuening Li, Daochen Zha, Na Zou, and Xia Hu. Pyodds: An end-to-end outlier detection system with automated machine learning. *Companion Proceedings of the Web Conference 2020*, 2020.

[13] Yue Liu, Shuang Li, and Wenjie Tian. Autocluster: Meta-learning based ensemble method for automated unsupervised clustering. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11–14, 2021, Proceedings, Part III*, page 246–258, Berlin, Heidelberg, 2021. Springer-Verlag.

[14] Martin Q. Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. A large-scale study on unsupervised outlier model selection: Do internal strategies suffice? *CoRR*, abs/2104.01422, 2021.

[15] Henrique O. Marques, Ricardo J. G. B. Campello, Arthur Zimek, and Jorg Sander. On the internal evaluation of unsupervised outlier detection. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, SSDBM '15, New York, NY, USA, 2015. Association for Computing Machinery.

[16] Thomas Giacomo Nies, Thomas Staudt, and Axel Munk. Transport dependency: Optimal transport based dependency measures. 2021.

[17] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11:355–607, 2019.

[18] Bruno Almeida Pimentel and Andre CPLF De Carvalho. A new data characterization for selecting clustering algorithms using meta-learning. *Information Sciences*, 477:203–219, 2019.

[19] Yannis Poulakis, Christos Doulkeridis, and Dimosthenis Kyriazis. Autoclust: A framework for automated clustering based on cluster validity indices. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1220–1225. IEEE, 2020.

[20] Herilalaina Rakotoarison, Louisot Milijaona, Andry RASOANAIVO, Michele Sebag, and Marc Schoenauer. Learning meta-features for autoML. In *International Conference on Learning Representations*, 2022.

[21] José A Sáez and Emilio Corchado. A meta-learning recommendation system for characterizing unsupervised problems: On using quality indices to describe data conformations. *IEEE Access*, 7:63247–63263, 2019.

[22] Meyer Scetbon and Marco Cuturi. Low-rank optimal transport: Approximation, statistics and debiasing. *NeurIPS 2022*, abs/2205.12365, 2022.

[23] Rodrigo GF Soares, Teresa B Ludermir, and Francisco AT De Carvalho. An analysis of meta-learning techniques for ranking clustering algorithms applied to artificial data. In *International Conference on Artificial Neural Networks*, pages 131–140. Springer, 2009.

[24] David Stern, Ralf Herbrich, Thore Graepel, Horst Samulowitz, Luca Pulina, and Armando Tacchella. Collaborative expert portfolio management. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence AAAI-10 (to appear)*, July 2010.

[25] Dennis Tschechlov, Manuel Fritz, Holger Schwarz, Y Velegrakis, D Zeinalipour-Yazti, PK Chrysanthis, and F Guerra. Automl4clust: Efficient automl for clustering analyses. In *EDBT*, pages 343–348, 2021.

[26] Joaquin Vanschoren. Meta-learning: A survey. *ArXiv*, abs/1810.03548, 2018.

[27] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

[28] Chi Wang, Qingyun Wu, Markus Weimer, and Erkang Zhu. Flaml: A fast and lightweight automl library. In *MLSys*, 2021.

[29] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *J. Mach. Learn. Res.*, 20:96:1–96:7, 2019.

[30] Yue Zhao, Ryan Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4489–4502. Curran Associates, Inc., 2021.