Target-Free Domain Adaptation through Cross-Adaptation (Student Abstract)

Aleksander Obuchowski [*1], **Barbara Klaudel**[*2], **Piotr Frackowski** [2], **Sebastian Krajna** [2], **Wasyl Badyra** [2], **Michał Czubenko** [2], **Zdzisław Kowalczuk** [2]

[1]Polish-Japanese Academy of Information Technology
[2]Gdańsk University of Technology
obuchowskialeksander@gmail.com, klaudel.b@gmail.com, s175486@student.pg.edu.pl,
s174136@student.pg.edu.pl,s175565@student.pg.edu.pl,michal.czubenko@pg.edu.pl, kova@pg.edu.pl

Abstract

The population characteristics of the datasets related to the same task may vary significantly and merging them may harm performance. In this paper, we propose a novel method of domain adaptation called "cross-adaptation". It allows for implicit adaptation to the target domain without the need for any labeled examples across this domain. We test our approach on 9 datasets for SARS-CoV-2 detection from complete blood count from different hospitals around the world. Results show that our solution is universal with respect to various classification algorithms and allows for up to a 10pp increase in F1 score on average.

Introduction

A fundamental problem of machine learning in healthcare is the transition from research to a real clinical setting. The number of medical devices and algorithms approved by the Food and Drug Administration still constitutes only a small number compared with the number of research projects developed yearly(Benjamens, Dhunnoo, and Meskó 2020). One of the reasons behind the difficulties of transferring from research to clinical practice is dataset bias. It occurs when the training data differs from the target population (Varoquaux and Cheplygina 2022). It can be revealed after training and testing the model on the dataset from different sources. This problem is especially relevant in the development of open-source solutions as open-source datasets usually come from a single centre and may fail to reflect the properties of populations with different demographics. The problem of dataset bias was particularly evident in the models tackling the problems of COVID-19. During the pandemic, a remarkable number of machine-aided diagnosis of COVID-19 papers was released in a very short time. However, the majority of produced models are of limited clinical use due to underlying biases (Roberts et al. 2021). The diversity of the available datasets for SARS-CoV-2 detection from complete blood count (CBC) makes them particularly suitable for testing domain generalization methods. The feasibility of routine blood tests for complete blood count has been tested before (Kistenev et al. 2022). However, the majority of the studies relied on the data from a single centre and only on internal validation. We propose a novel method of domain adaptation called "cross-adaptation". It allows for implicit adaption to the target domain without the need for any labelled examples across this domain. We test our method on 9 CBC detests, 5 algorithms and 3 adaptation methods. Our code is publicly available at: https://github.com/TheLion-ai/cross-adaptation

[*]Authors contributed equally

Background

Domain adaptation belongs to a group of methods known as transfer learning. Transfer learning is the process of adaptation to new situations, new tasks, and new environments (Yang et al. 2020). It helps to improve the learning of a target task for the target domain using the knowledge from the source domain. Domain adaptation is a type of transductive transfer learning. It tackles a problem where the source and target tasks are the same but the domains differ. It attempts to align the distributions of the data to minimize the discrepancies between the domains. It assumes that feature and label spaces are the same while their probability distributions differ. It has been applied to medical scenarios in several works (Guan and Liu 2021).

Our Solution

Algorithm 1: Target-free domain adaptation through cross-adaptation.

Initialize an empty set of transformed datasets $D_x = \{\}$
For each domain D_i in the set of domains $D = \{D_1, D_2, ..., D_n\}$:

1. Treat the dataset D_i as a source dataset $D_s := D_i$.
2. Treat datasets from other domains as the target dataset $D_t := D - D_i$.
3. Transform source dataset based on the target with the selected domain adaptation method – d.

$$D_{x_i} = g(D_s, D_t)$$

4. Add D_{x_i} to the transformed datasets D_x.

Use the transformed dataset to train the general model.

Cross-adaptation performs domain adaptation across a set of domains of all the available training datasets $\{D_1, D_2, ..., D_n\}$ (where n is the number of training datasets) by iteratively changing which dataset is the source and which datasets are the target. It follows the case of multi-target domain adaptation.

The algorithm produces transformed datasets $D_x = \{D_{x_1}, D_{x_2}, ..., D_{x_n}\}$. The procedure for transforming the datasets is described in Algorithm 1.

The resulting datasets D_x that is a concatenation of all transformed datasets $\{D_{x_1}, D_{x_2}, ..., D_{x_n}\}$ can then be paired with its respective labels to train the model f. This model can then be successfully used on a new domain not present in the training domains D. As shown in the Algorithm 1, this method works irrespective of the domain adaptation method g. We can denote the source and target domains as D_s and D_t.

Experiments and Results

We tested our method on the task of SARS-CoV-2 detection based on Complete Blood Count (CBC) tests. Additionally, we used 5 different machine learning algorithms and 3 domain adaptation algorithms (Kernel Mean Matching, Kullback-Leibler Importance Estimation Procedure, Transfer AdaBoost for Classification) to test the robustness of our method. Our dataset consists of 4870 tests from 8 publicly available datasets and 1 closed-source dataset (Cabitza et al. 2021b) (Cabitza et al. 2021a) (Klaudel et al. 2023). We selected 10 input features: white blood cells, hemoglobin, mean corpuscular volume, mean corpuscular hemoglobin concentration, platelets, monocytes count %, basophils count %, lymphocytes count %, eosinophils count %, and sex. The features are the most common CBC parameters often measured as a part of a routine examination. For each dataset, the target variable (infection with the SARS-CoV-2 virus) was determined based on the RT-PCR result. Only the data from training sets were transformed. The test and validation data required no transformations. The transformed datasets were not used as targets, but their untransformed counterparts. The training dataset consisted of data from 2 hospitals in Brazil, 3 in Italy, and 1 in Poland. The tests were performed on the datasets from 1 hospital in Ethiopia and 1 in Spain. Table 1 presents the results of the experiment comparing 2 scenarios (1) training on untransformed source datasets, (2) training on source datasets transformed with cross-adaptation. The results for cross-adaptation present average results for all 3 domain adaptation algorithms. We used hyperparameter tuning for both scenarios.

Model	Untransformed source	CA
KNN	55%	73%
Decision Tree	65%	73%
Random Forest	62%	70%
XGBoost	62%	73%
MLP	65%	72%

Table 1: The results (F1 score) for training on untransformed source datasets compared with training on the source data transformed with cross-adaptation (CA).

Conclusions and Future Work

In the paper, we present a novel method for target-free domain adaptation. It is most suitable for training the models for the healthcare domain, where the training dataset consists of the records from different hospitals around the world and the final solution is meant to be deployed globally. Preliminary results from CBC datasets show is able to improve F1 score of the model by 10pp on average. In the future, we are planning to test the method across different modalities and tasks.

References

Benjamens, S.; Dhunnoo, P.; and Meskó, B. 2020. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1): 1–8.

Cabitza, F.; Campagner, A.; Ferrari, D.; Di Resta, C.; Ceriotti, D.; Sabetta, E.; Colombini, A.; De Vecchi, E.; Banfi, G.; Locatelli, M.; et al. 2021a. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 59(2): 421–431.

Cabitza, F.; Campagner, A.; Soares, F.; de Guadiana-Romualdo, L. G.; Challa, F.; Sulejmani, A.; Seghezzi, M.; and Carobene, A. 2021b. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, 208: 106288.

Guan, H.; and Liu, M. 2021. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3): 1173–1185.

Kistenev, Y. V.; Vrazhnov, D. A.; Shnaider, E. E.; and Zuhayri, H. 2022. Predictive models for COVID-19 detection using routine blood tests and machine learning. *Heliyon*, e11185.

Klaudel, B.; Obuchowski, A.; Dabrowska, M.; Sałaga-Zaleska, K.; and Kowalczuk, Z. 2023. Machine-aided detection of SARS-CoV-2 from complete blood count. In *International Conference on Diagnostics of Processes and Systems*, 17–28. Springer.

Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A. I.; Etmann, C.; McCague, C.; Beer, L.; et al. 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3): 199–217.

Varoquaux, G.; and Cheplygina, V. 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1): 48.

Yang, Q.; Zhang, Y.; Dai, W.; and Pan, S. J. 2020. *Introduction*, 3–22. Cambridge University Press.