Robust Learning from Untrusted Sources

Nikola Konstantinov   Christoph H. Lampert

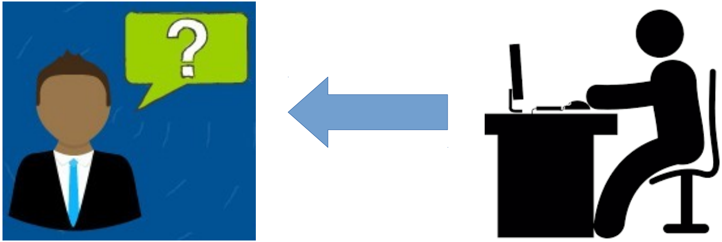
ICML, June 2019

Institute of Science and Technology

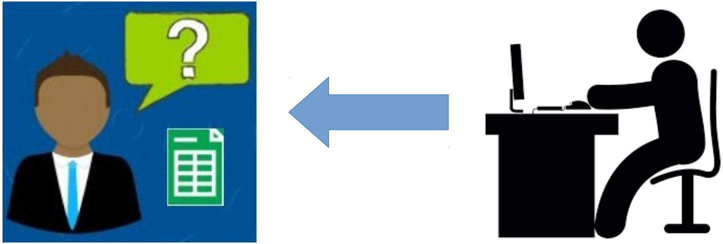
European Research Council

Collecting data for machine learning applications

Collecting data for machine learning applications

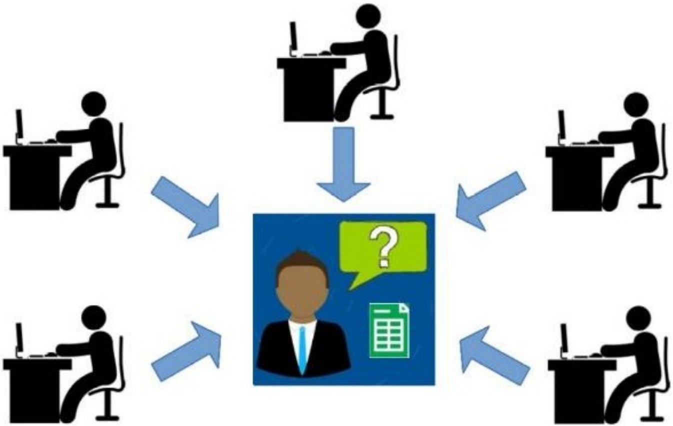


Collecting data for machine learning applications



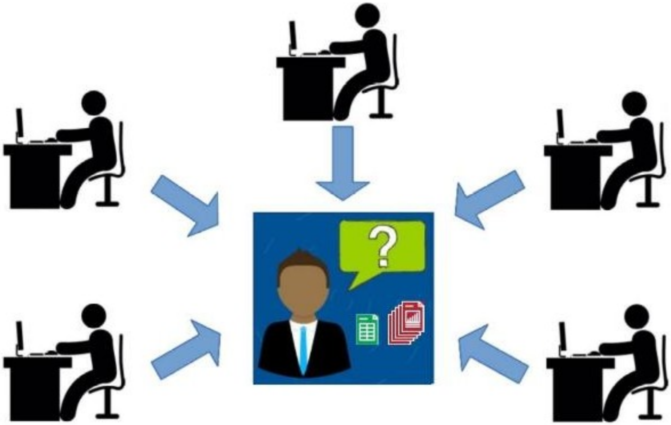
Using multiple data sources

Crowdsourcing



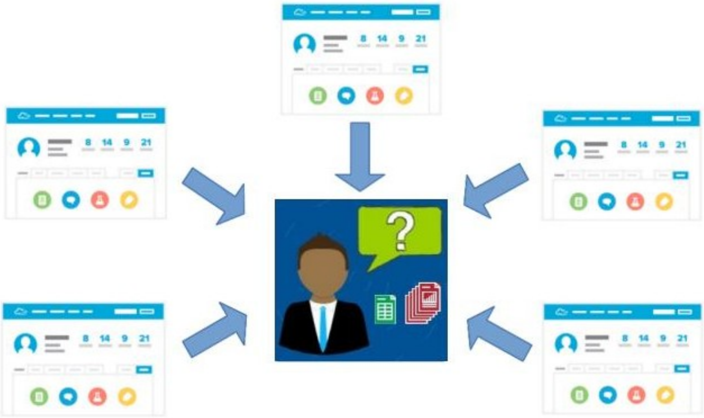
Using multiple data sources

Crowdsourcing



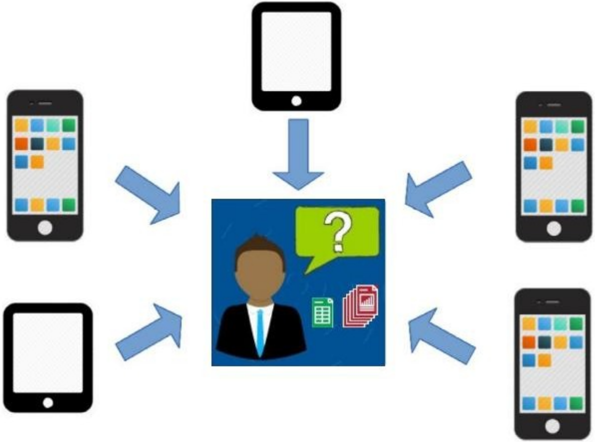
Using multiple data sources

Web crawling



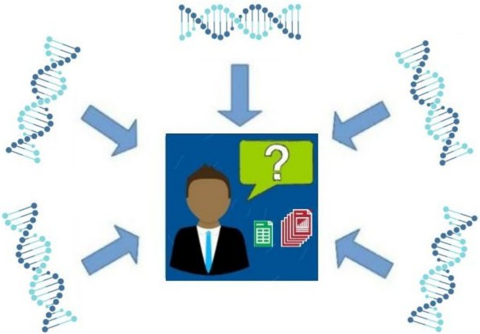
Using multiple data sources

Data from personal devices



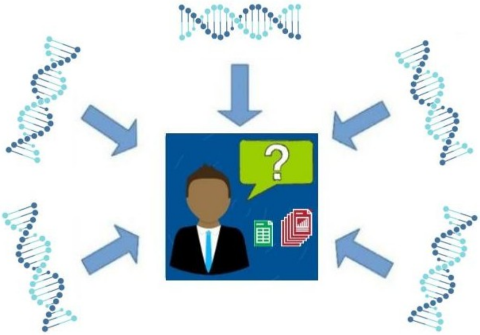
Using multiple data sources

Data from different labs



Using multiple data sources

Data from different labs



How can we learn robustly from such data?

Learning from untrusted sources

Motivation

- Untrusted sources can provide valuable data for training.
- Some of these data batches might be corrupted or irrelevant.

Goal

- Naive approaches are to:
  - Simply train on all data.
  - Train only on the trusted subset.
- Can we do better?

Setup

Learning task

- Unknown target distribution $\mathcal{D}_T$ on $\mathcal{X} \times \mathcal{Y}$.
- Loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.
- Want to learn a predictor $h : \mathcal{X} \to \mathcal{Y}$ from a hypothesis class $\mathcal{H}$.

Given

- Have a small reference dataset:

$$S_T = \left\{ \left(x_1^T, y_1^T\right), \dots, \left(x_{m_T}^T, y_{m_T}^T\right) \right\} \sim \mathcal{D}_T$$

- Also given m_i data points from each source $i = 1, \dots, N$:

$$S_i = \left\{ \left(x_1^i, y_1^i\right), \dots, \left(x_{m_i}^i, y_{m_i}^i\right) \right\} \sim \mathcal{D}_i$$

Approach

- Assign weights $\alpha = (\alpha_1, ..., \alpha_N)$ to the sources, $\sum_{i=1}^{N} \alpha_i = 1$.
- Minimize the α-weighted empirical loss:

$$\hat{h}_\alpha = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\epsilon}_\alpha(h) = \operatorname*{argmin}_{h \in \mathcal{H}} \left(\sum_{i=1}^{N} \alpha_i \frac{1}{m_i} \sum_{j=1}^{m_i} L\left(h\left(x_j^i\right), y_j^i\right) \right)$$

- Want a small expected loss on the target distribution:

$$\epsilon_T\left(\hat{h}_\alpha\right) = \mathbb{E}_{\mathcal{D}_T}\left(L(\hat{h}_\alpha(x), y)\right)$$

- How to decide which sources are *trustworthy*?

Approach

- Discrepancies between the sources (Kifer et al., VLDB 2004; Mohri et al., ALT 2012):

$$\mathrm{disc}_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_T) = \sup_{h \in \mathcal{H}} |\epsilon_i(h) - \epsilon_T(h)|$$

- Small if $\mathcal{H}$ does not distinguish between the two learning tasks.
- Popular in the domain adaptation literature.

Bound on the expected loss

- Given a hypothesis set $\mathcal{H}$, let:
  - $\hat{h}_\alpha = \text{argmin}_{h\in\mathcal{H}}\, \hat{\epsilon}_\alpha(h)$
  - $h_T^* = \text{argmin}_{h\in\mathcal{H}}\, \epsilon_T(h)$
- For any $\delta > 0$, with probability at least $1-\delta$:

$$|\epsilon_T(\hat{h}_\alpha) - \epsilon_T(h_T^*)| \leq$$

$$2\sum_{i=1}^{N} \alpha_i \text{disc}_{\mathcal{H}}\left(\mathcal{D}_i, \mathcal{D}_T\right) + C\left(\delta\right)\sqrt{\sum_{i=1}^{N}\frac{\alpha_i^2}{m_i}} + 4\sum_{i=1}^{N}\alpha_i \mathcal{R}_i\left(\mathcal{H}, L\right)$$

Similar bounds in Ben-David et al., ML 2010; Zhang et al., NIPS 2013.

Bound on the expected loss

- Given a hypothesis set $\mathcal{H}$, let:
  - $\hat{h}_\alpha = \text{argmin}_{h \in \mathcal{H}} \hat{\epsilon}_\alpha(h)$
  - $h^*_T = \text{argmin}_{h \in \mathcal{H}} \epsilon_T(h)$
- For any $\delta > 0$, with probability at least $1 - \delta$:

$$
|\epsilon_T(\hat{h}_\alpha) - \epsilon_T(h^*_T)| \leq
$$

$$
2 \sum_{i=1}^{N} \alpha_i \text{disc}_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_T) + C(\delta) \sqrt{\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}} + 4 \sum_{i=1}^{N} \alpha_i \mathcal{R}_i(\mathcal{H}, L)
$$

Similar bounds in Ben-David et al., ML 2010; Zhang et al., NIPS 2013.

Bound on the expected loss

- Given a hypothesis set $\mathcal{H}$, let:
  - $\hat{h}_\alpha = \text{argmin}_{h \in \mathcal{H}} \hat{\epsilon}_\alpha(h)$
  - $h_T^* = \text{argmin}_{h \in \mathcal{H}} \epsilon_T(h)$
- For any $\delta > 0$, with probability at least $1 - \delta$:

$$|\epsilon_T(\hat{h}_\alpha) - \epsilon_T(h_T^*)| \leq$$

$$2 \sum_{i=1}^{N} \alpha_i \text{disc}_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_T) + C(\delta) \sqrt{\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}} + 4 \sum_{i=1}^{N} \alpha_i \mathcal{R}_i(\mathcal{H}, L)$$

Similar bounds in Ben-David et al., ML 2010; Zhang et al., NIPS 2013.

Bound on the expected loss

- Given a hypothesis set $\mathcal{H}$, let:
  - $\hat{h}_\alpha = \text{argmin}_{h \in \mathcal{H}} \hat{\epsilon}_\alpha(h)$
  - $h^*_T = \text{argmin}_{h \in \mathcal{H}} \epsilon_T(h)$
- For any $\delta > 0$, with probability at least $1 - \delta$:

$$|\epsilon_T(\hat{h}_\alpha) - \epsilon_T(h^*_T)| \leq$$

$$2\sum_{i=1}^{N} \alpha_i \text{disc}_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_T) + C(\delta)\sqrt{\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}} + 4\sum_{i=1}^{N} \alpha_i \mathcal{R}_i(\mathcal{H}, L)$$

Similar bounds in Ben-David et al., ML 2010; Zhang et al., NIPS 2013.

Bound on the expected loss

- Given a hypothesis set $\mathcal{H}$, let:
  - $\hat{h}_\alpha = \text{argmin}_{h\in\mathcal{H}}\, \hat{\epsilon}_\alpha(h)$
  - $h^*_T = \text{argmin}_{h\in\mathcal{H}}\, \epsilon_T(h)$
- For any $\delta > 0$, with probability at least $1-\delta$:

$$|\epsilon_T(\hat{h}_\alpha) - \epsilon_T(h^*_T)| \leq$$

$$2\sum_{i=1}^{N}\alpha_i \text{disc}_{\mathcal{H}}\left(\mathcal{D}_i, \mathcal{D}_T\right) + C\left(\delta\right)\sqrt{\sum_{i=1}^{N}\frac{\alpha_i^2}{m_i}} + 4\sum_{i=1}^{N}\alpha_i \mathcal{R}_i\left(\mathcal{H}, L\right)$$

Similar bounds in Ben-David et al., ML 2010; Zhang et al., NIPS 2013.

Bound on the expected loss

- Given a hypothesis set $\mathcal{H}$, let:
  - $\hat{h}_\alpha = \text{argmin}_{h \in \mathcal{H}} \hat{\epsilon}_\alpha(h)$
  - $h_T^* = \text{argmin}_{h \in \mathcal{H}} \epsilon_T(h)$
- For any $\delta > 0$, with probability at least $1 - \delta$:

$$|\epsilon_T(\hat{h}_\alpha) - \epsilon_T(h_T^*)| \leq$$

$$2\sum_{i=1}^{N} \alpha_i \text{disc}_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_T) + C(\delta)\sqrt{\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}} + 4\sum_{i=1}^{N} \alpha_i \mathcal{R}_i(\mathcal{H}, L)$$

Similar bounds in Ben-David et al., ML 2010; Zhang et al., NIPS 2013.

Algorithm

- Theory suggests:
  - Select α by minimizing:

$$\sum_{i=1}^{N} \alpha_i \text{disc}_{\mathcal{H}}\left(\mathcal{D}_i, \mathcal{D}_T\right) + \lambda \sqrt{\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}}$$

  - Find $\hat{h}_\alpha$ by minimizing the α-weighted empirical risk.
  - Choose λ by cross-validation on the reference dataset.
- Trade-off between exploiting trusted sources and using all data.
- In practice, work with the empirical discrepancies:

$$\text{disc}_{\mathcal{H}}\left(S_i, S_T\right) = \sup_{h \in \mathcal{H}} \left| \frac{1}{m_i} \sum_{j=1}^{m_i} L\left(h\left(x_j^i\right), y_j^i\right) - \frac{1}{m_T} \sum_{j=1}^{m_T} L\left(h\left(x_j^T\right), y_j^T\right) \right|$$

Experiments

- Evaluate empirically on:
  - Multitask Dataset of Product Reviews [1].
  - Animals with Attributes 2 [2].
- Some clean reference data for a target task is available.
- Have other subsets, some of which are corrupted.
- Experimented with various manipulations/problems with the data.

[1] Pentina et al., ICML 2017; McAuley et al., 2015
[2] Xian et al., TPAMI 2018

Results

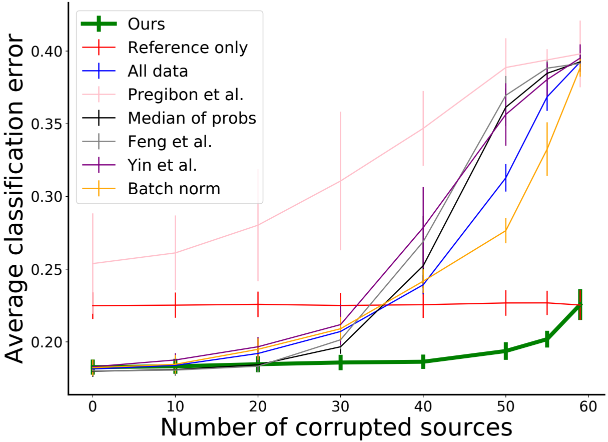


Figure: Animals with Attributes 2: RGB channels swapped

Summary

- Data from different sources is naturally heterogeneous.
- Our method suppresses the effect of corrupted/irrelevant data.
- The approach is theoretically justified and shows good empirical performance.
- The algorithm can be applied even when the data is private and/or distributed.

Summary

- Data from different sources is naturally heterogeneous.
- Our method suppresses the effect of corrupted/irrelevant data.
- The approach is theoretically justified and shows good empirical performance.
- The algorithm can be applied even when the data is private and/or distributed.

Thank you for your attention!

Poster 156

Summary

- Data from different sources is naturally heterogeneous.
- Our method suppresses the effect of corrupted/irrelevant data.
- The approach is theoretically justified and shows good empirical performance.
- The algorithm can be applied even when the data is private and/or distributed.

Thank you for your attention!

Poster 156

Code available at: https://github.com/NikolaKon1994/Robust-Learning-from-Untrusted-Sources

References I

Ben-David, Shai et al. (2010). "A theory of learning from different domains". In: *Machine learning* 79.1-2, pp. 151–175.

Kifer, Daniel et al. (2004). "Detecting change in data streams". In: *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30.*

McAuley, Julian et al. (2015). "Image-based recommendations on styles and substitutes". In: *38th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM.

Mohri, Mehryar et al. (2012). "New analysis and algorithm for learning with drifting distributions". In: *International Conference on Algorithmic Learning Theory.*

Pentina, Anastasia et al. (2017). "Multi-task Learning with Labeled and Unlabeled Tasks". In: *International Conference on Machine Learning (ICML).*

Xian, Yongqin et al. (2018). "Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly". In: *IEEE transactions on pattern analysis and machine intelligence.*

References II

Zhang, Chao et al. (2012). "Generalization bounds for domain adaptation". In: *Advances in neural information processing systems*.