Domain Adaptation from a Pre-trained Source Model

Application on fraud detection tasks

Presenter: Luxin Zhang (Worldline & Inria)

Supervisors: Christophe Biernacki (Inria), Pascal Germain (Inria), Yacine Kessaci (Worldline)

CMStatistics 2019

Nov 15, 2019

Fraud Detection in Transactions

Fraud Detection Problem:
Detect if a transaction is issued by the customer or not.

Fraud Detection Model:
A binary classification model based on the historical transactions of a customer.

Characteristic of Fraud Detection Dataset:

- Huge number of examples (600 thousand per day).
- Extremely imbalanced class (0.2% of fraud).
- Categorical and numerical attributes.
- Highly dependent manually generated attributes.
- Numerical attributes are very skew.

Why to Transfer

Existing Market (Country)

- Well trained classification model.
- The pattern of fraudster evolves.

Expanding Market (Country)

- Consumer behaviors are different from country to country.
- Not enough label information in a new country.
- The pattern of fraudster evolves.

Technology used to face the challenge: **Domain Adaptation**

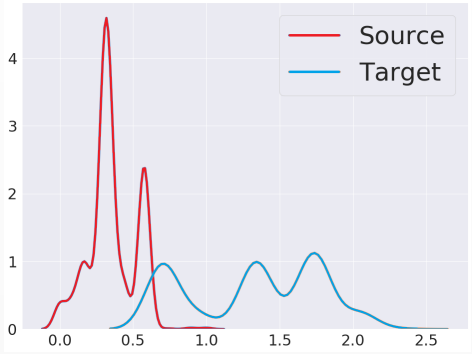
Plan

Introduction of Domain Adaptation

What is Domain Adaptation?

Domain adaptation is a technique of transfer learning to reduce the drift between distributions of data from different domains (Pan and Yang [3])

- Why to transfer? (Just Answered)
- What to transfer?
- How to transfer?



Context

Simplified Dataset:

- Encode categorical attributes by historical risk score.
- Use log-transformation to fix the skew numerical attributes.

Notations:

- $\mathcal{X} = \mathbb{R}^d$: input space. $\mathcal{Y} = \{0, 1\}$: output space.
- $X_s, X_t \in \mathcal{X}$: input data of two domains. $Y_s, Y_t \in \mathcal{Y}$: output data of two domains.
- $h : \mathcal{X} \to [0, 1]$: classifier that returns the probability of being fraud.
- $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$: loss function.
- $R_s^l(h)$, $R_t^l(h)$: True risk of classifier h.

What to Transfer

Our Proposition:
Target to Source Domain Adaptation.

Target to Source Domain Adaptation

Assumption:

$$\text{No label shift} \implies P(Y_s) = P(Y_t)$$

Proposition:

$$P(X_s|Y_s) = P(\mathcal{G}(X_t)|Y_t) \implies R_t^l(h_s^* \circ \mathcal{G}) = R_t^l(h_t^*)$$

$\mathcal{G}$ is the transformation that we are looking for and h_s^* and h_t^* are respectively the true risk minimizers of two domains.

Characteristic of Fraud Detection:

- Proportion of fraud is nearly the same.
- No (not enough) Y_t.

Justification of Assumptions:

- No label shift.
- $\mathcal{G}$ does not depend on Y.

What to Transfer

Related Works:

- Source to target adaptation.
- Common space adaptation.

Advantages of Target to Source Transformation:

- Leverage the improvement of source model.
- No more retraining for every new country.
- A robust model needs investment and expertise.

How to Transfer

Difficulties:
Y_t is not enough to directly estimate $\mathcal{G}$.

Industrial Requirements:

- Better understand consumer behaviors in new country.
- Transactions dataset is large.

Transformation $\mathcal{G}$:

- Interpretability.
- Modularity.
- Scalability.

How to Transfer

Intuition

$$P(X_s|Y_s) = P(\mathcal{G}(X_t)|Y_t) \iff P(X_s) = P(\mathcal{G}(X_t))$$

The function $\mathcal{G}$ who minimizes the "marginal transformation efforts" aligns also the conditional distribution.

$$\mathcal{G} = \underset{\mathcal{G}}{\operatorname{argmin}}\, \mathcal{W}_p\Big(P\big(X_s\big), P\big(\mathcal{G}(X_t)\big)\Big)$$

$\mathcal{W}_p$ is the l_p wasserstein distance. The domain adaptation is formulated to be an optimal transport problem (Courty et al. [1]).

Details of the Transformation

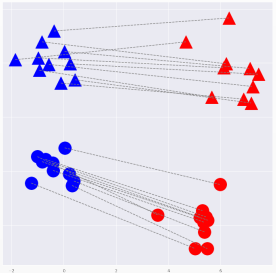
Wasserstein Distance on Empirical Dataset:

$$\mathcal{W}_p(P_s, P_t) = min_{\gamma \in \Gamma(P_s, P_t)} < C_p, \gamma >$$

$< C_p, \gamma >$: the sum of element wise product of matrix C_p and γ.
C_p: a l_p norm matrix between all pairs of examples.
$\Gamma(P_s, P_t)$: a set of joint probability matrix of $P(X_s)$ and $P(X_t)$.

Optimal Transport:

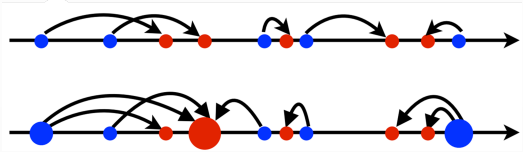
- Aligns the distributions.
- Easy to interpret.
- **Not scalable on big dataset.**(even with entropy regularization [2])



Details of the Transformation

1D Optimal Transport:

It is well known that 1D optimal transport has a closed-form solution where $\mathcal{G}_{1D}(x) = (F_{P_s}^{-1} \circ F_{P_t})(x)$, F is a cumulative distribution function. This solution is also known as the increasing arrangement. (Peyré et al. [4])



Compositions of $\mathcal{G}$:

Assumption: All attributes are independent (or move towards the same direction).

$$\mathcal{G} = \Big[\mathcal{G}_1 \Big| \mathcal{G}_2 \Big| ... \Big| \mathcal{G}_i \Big| ... \Big| \mathcal{G}_{k-1} \Big| \mathcal{G}_k \Big] \text{ where } \mathcal{G}_i = \underset{\mathcal{G}}{\operatorname{argmin}}\, \mathcal{W}_p\Big(P\big(X_{s,i}\big), P\big(\mathcal{G}(X_{t,i})\big)\Big)$$

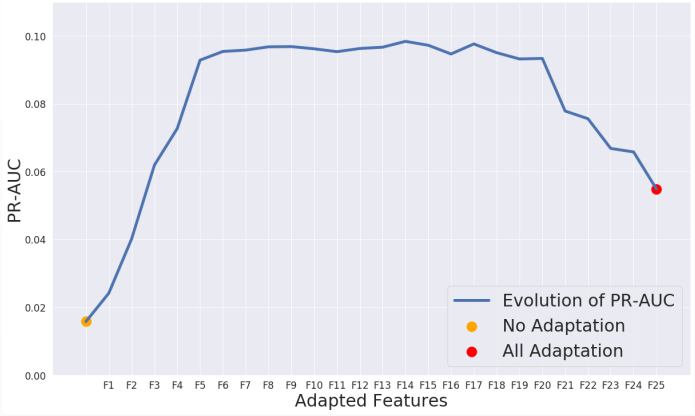
$X_{s,i}$ and $X_{t,i}$ are the i-th attributes of input data X.

Target to Source Domain Adaptation

Which attribute to transfer?

Feature selection using accessible labeled target data.

- Separate attributes into different groups.
- A greedy search based on classifier's performance.
- Keep the attributes the most significant for adaptation.



Experimental Results

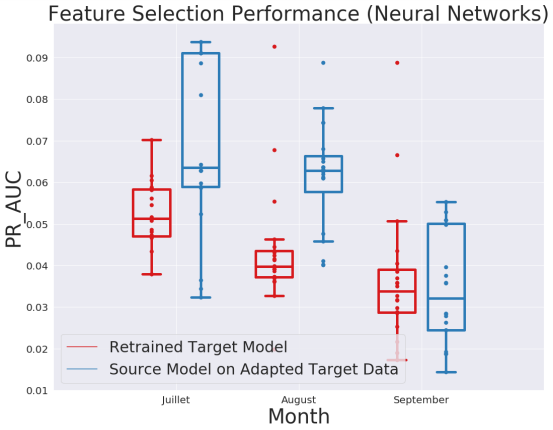
	No Adaptation	**All Adaptation**	**Selected Adaptation**
Juillet	0.016	0.055	$\mathbf{0.070 \pm 0.009}$
August	0.061	**0.077**	0.061 ± 0.006
September	0.013	**0.052**	0.034 ± 0.006

Table: Performance of adaptation based on Neural Networks

	No Adaptation	**All Adaptation**	**Selected Adaptation**
Juillet	0.038	0.045	$\mathbf{0.054 \pm 0.002}$
August	0.063	**0.072**	0.062 ± 0.003
September	0.019	0.038	$\mathbf{0.048 \pm 0.002}$

Table: Performance of adaptation based on Xgboost.

Experimental Results



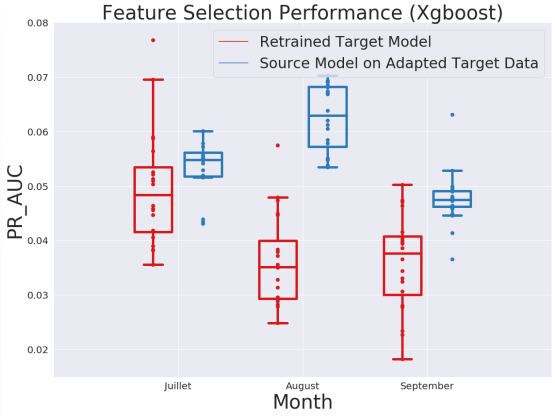


Figure: Comparison of feature selection performance to retrained target model.

Prospects

- Transfer directly the categorical attributes.
- Take into account the imbalance of class.
- Take into account the dependence of attributes.
- Take into account the characteristic of the source classifier.

References

[1] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39 (9):1853–1865, 2016.

[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[3] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[4] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.