Information transfer for learning in non-stationary environments

Pierre-Alexandre Murena
LTCI - Télécom ParisTech, Université Paris-Saclay, Paris, France
murena@telecom-paristech.fr

Abstract

Traditional machine learning setting consists in learning a concept from a learning data set and applying the learned concept on a test data set supposed to be independent from the learning data set but equally distributed.
In practice, this hypothesis does not always hold and some non-stationary environments introduce **changes in the distributions** (concept drift). Two classes of problems belong to this non-stationary category: **transfer learning** and **incremental learning**. In both of them, the acquired knowledge has to be *transferred* and slightly modified to fit new environments.
We present a framework for learning in non-stationary environment based on the notion of algorithmic complexity introducing the idea of minimal transfer of information.

Reminder: Supervised Learning

Traditional supervised learning can be split into two phases:

1. **Training step:** Learn a decision function β from a *labeled* data set $\mathcal{D}_{train} = \{(X_i, Y_i)\}_{i=1,\dots,n}$
2. **Test step:** New data are observed and must be classified: $\mathcal{D}_{test} = \{X_i\}_{i=1,\dots,m}$

A fundamental hypothesis assumes that the data $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ are **independent and identically distributed (i.i.d.)**. It gives sense to empirical risk minimization:

$$\beta^* = \arg\min_{\beta} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\beta(X_i) \neq Y_i) \qquad (1)$$

When $\mathcal{D}_{train}$ is large enough, the obtained decision function β^* is guaranteed to be close to the optimal decision function.

If the distribution changes, the classifier β is not adapted anymore!

Transfer learning

In Transfer Learning, the learner has to learn a concept from a labeled source data set $\mathcal{D}_S = \{(X_i^S, Y_i^S)\}$ and to "*transfer*" the concept to a target data set $\mathcal{D}_T = \{X_i^T\}$.

Our framework [1]:
We propose to apply the minimum description length principle and to encapsulate knowledge transfer inside the model transfer:

$$\min_{M_S, M_T} \quad C(M_S)+C(X_S|M_S)+C(Y_S|M_S,X_S)+C(M_T|M_S)+C(X_T|M_T)$$

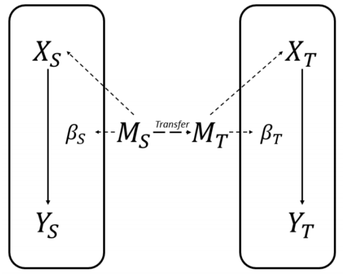

Figure 1: Transfer learning framework

- The information transfer is measured by the quantity $C(M_T|M_S)$.
- The framework remains valid for i.i.d. data (limit case).
- The framework can be used for analogical reasoning ("*ABC is transformed into ABD. What is the result of the same transformation on IJK?*")

Minimum Description Length Principle

The best theory chosen to describe observed data is the theory minimizing the sum of the description length (in bits) of the theory description and of the data encoded using the theory.
The description length of an object X is measured by its **Kolmogorov complexity**. Kolmogorov complexity $C_{\mathcal{M}}(X)$ of X is defined as the length (in bits) of the shortest program generating X on the universal Turing machine $\mathcal{M}$.

Incremental learning

Incremental learning designates online learning of a model from streaming data. In non-stationary environments, the process generating these data may change over time, hence the learned concept becomes invalid. This phenomenon is called "*concept drift*".

Our framework [2]:
We use MDL principle with the same data encoding as in transfer learning. Data are described individually with the help of their corresponding model only. Models are transferred from previously learned models (denoted $M_{\Delta_t^{-1}(\{1\})}$. The final objective is given by:

$$\min_{\mathbf{M}} \quad \sum_t C(M_t|M_{\Delta_t^{-1}(\{1\})})+C(X_t|M_t)+C(\beta_t|M_t,X_t)+C(Y_t|M_t,X_t,\beta_t)$$

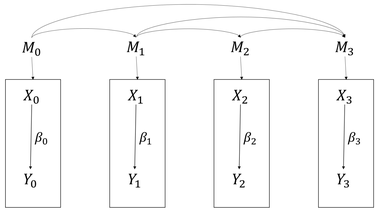

Figure 2: Incremental learning framework

- The chosen framework is agnostic to the nature of data, the learning strategy (active or passive), the modeling strategy and the nature of the drift (abrupt, gradual, recurrent, ...).
- It has been shown to be consistent with state-of-the-art algorithms.
- Implementing a naive algorithm with prototype-based models leads to good experimental results on benchmark data sets.

Conclusion

- We proposed a strong multi-domain and probability-free framework for describing learning problems in non-stationary environments which relies on a minimal transfer of information and on the use of Kolmogorov complexity.
- Several models can be incorporated into our framework, in particular probabilistic models (by the relation $C(x|\mu) = -\log_2 \mu(x)$).
- Competitive performances are obtained with naive models and algorithms, both on real and artificial data.
- The proposed framework can be applied to a wide variety of problems and is not restricted to numerical data. Applications to character strings, ontologies and categorical data follow directly.

Perspectives

- Non-Euclidean nature of learning: Which link with information geometry?
- Learning theory: Toward a generalization of PAC learning?
- Applications: Analogical reasoning, Neural models, application to large data sets.

References

[1] Pierre-Alexandre Murena and Antoine Cornuéjols. Minimum description length principle applied to structure adaptation for classification under concept drift. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2842–2849. IEEE, 2016.

[2] Pierre-Alexandre Murena, Antoine Cornuéjols, and Jean-Louis Dessalles. Incremental learning with the minimum description length principle. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 1908–1915. IEEE, 2017.

Acknowledgements

This work has been supported by the program Futur & Ruptures (Institut Mines-Télécom).