Research Summary: Knowledge Transfer in Artificial Learning

Pierre-Alexandre Murena

Télécom ParisTech - Université Paris Saclay,
46 rue Barrault, 75013 Paris, France,
murena@telecom-paristech.fr

1 Introduction

This document presents the main research problems addressed during my PhD studies. All these researches are led inside the two teams DBWeb in Télécom ParisTech and LInK (Learning and Integration of Knowledge) in AgroParisTech, both located in Paris, and supervised by Pr. Jean-Louis Dessalles and Pr. Antoine Cornuéjols.

My researches focus on learning theory both in the perspective of symbolic machine learning and of learning in continuous domains. I aim at finding an information-theoretic principle guiding information transfer in learning.

The start point of these researches is the idea that most machine learning takes a strong stationary hypothesis for granted. The general framework of statistical learning (mainly supervised and semi-supervised learning) considers two data sets: a *learning data set* (from which the concepts have to be learned) and a *test data set* (on which quality of the learned concepts is evaluated). The key idea of current learning methods and theories is to assume that *training data and test data are independent and identically distributed (i.i.d.)*. However this strong hypothesis does not hold in many cases: either the data generation process evolves over time (aging effect, trending effect...) or the data belong to a different domain. Because similar questions of transfer and domain adaptation had already been addressed in analogical reasoning, we proposed to use an approach based on Kolmogorov complexity instead of probabilities. Kolmogorov complexity is a measure of the information contained inside an object. The use of Kolmogorov complexity in machine learning is accepted by the community, but mainly in a stationary point of view (when the key concept does not vary); we proposed to extend its use to non-stationary environments, in the same way as done in analogical reasoning. A presentation of these issues is given in section 2.

The strong similarity between transfer learning and analogical reasoning led me to consider this issue in my researches. Analogical reasoning consists in situations of the form "*'b' is to 'a' as 'd' is to 'c'*". Because its value has already been demonstrated, I focus on Hofstadter's micro-world, made up of strings of alphabetical characters that can be described with simple concepts like 'predecessor', 'successor' or 'repetition'. The use of Kolmogorov complexity for analogical reasoning had already been considered, but our approach is slightly different. We

developed a small descriptive language for Hofstadter's problems and convert it into a binary code, the length of which corresponds to Kolmogorov complexity. Our work on analogy is presented in section 3. Finally, because of its very global perspective, our research topic leads naturally to collaborations on various topics related to learning. This side aspect of my research is presented in section 4.

2 A global approach of learning

Statistical machine learning in its current form is often considered to be based mainly on three inductive principles: Empirical Risk Minimization (ERM), Bayesianism and Minimum Description Length (MDL). The validity of ERM has been demonstrated under the strong i.i.d. assumption using several frameworks, all inspired by the Probably Approximately Correct learning framework. Besides strong links have been stated between Bayesianism and MDL.

When the i.i.d. hypothesis is not verified, these three principles are not valid anymore and have to be replaced by new principles. Exploring this direction, we considered the most straightforward transfer learning problem and the classification problem, the purpose of which is to associate data to labels. Given source data X_S associated to their classes Y_S and a target data X_T, we aim at finding the corresponding classes Y_T. The idea is to find a classification function β_S such that $\beta_S(X_S) = Y_S$ and to transfer this function into a classification function β_T available on the target data X_T. Because this problem is the same as analogical reasoning, we used a simplification of the general MDL principle in the context of analogy [2].

The mathematical tool used to measure the *description length* in MDL is Kolmogorov complexity [4]. Kolmogorov complexity (also called algorithmic complexity) of an object x is an information theoretic measure defined as the minimal length of a program defined on a Universal Turing Machine (UTM) and the output of which is object x. This quantity can be shown to be machine independent, but non calculable.

The idea we developed is to consider an upper-bound of this complexity based on a restricted Turing machine. The choice of a restricted Turing machine corresponds to an inductive bias, inevitable in any inductive reasoning (see for example the *no-free-lunch theorem*). This choice also raises the problem of machine dependency which seems crucial in human learning.

Our first contribution is a direct use of analogical MDL in the context of transfer learning [5]: I presented a two-part MDL equation based on a data representation called *model*. A model is any object which may be used to compress data. In transfer learning, our purpose is to infer a source model M_S and a target model M_T:

$$\min_{M_S, M_T} \quad C(M_S) + C(X_S|M_S) + C(Y_S|M_S, X_S) + C(M_T|M_S) + C(X_T|M_T) \quad (1)$$

where $C(.)$ designates Kolmogorov complexity, X the data, Y_S the source labels and M the model. This equation applies both for continuous data (X is a matrix,

the rows of which correspond to a vector data point) and for symbolic data (X is a sequence of symbols, for example a sequence of letters). The different terms in the equation present a strong similarity with usual machine learning terms: $C(M)$ corresponds to a model penalization based on its complexity; $C(X|M)$ corresponds to a likelihood term, ie. a fitness of the model toward data; and the term $C(Y|M,X)$ corresponds to an empirical risk. In the paper, we also proposed experimental validations on two toy data sets with a simple prototype-based model. They present good results and high performance on these data sets.

A direct variant of this formula has been proposed for incremental learning [6]. In incremental learning, the system faces a sequence of questions $X_1, X_2, \ldots$ and has to find the solution to each problem one by one. The model used to describe data can be updated if it is outdated and does not correspond to the current data anymore. We propose the following simplified MDL objective:

$$\min_{M} \quad \sum_{t} C(M_t|M_{\Delta_t^{-1}(\{1\})}) + C(X_t|M_t) + C(\beta_t|M_t, X_t) + C(Y_t|M_t, X_t, \beta_t) \quad (2)$$

where Δ_t is a model association function such that $\Delta_t(u) = 1$ if model M_t can be described with model M_u, and $\Delta_t(u) = 0$ otherwise. The consistency of our framework with existing heuristics state-of-the-art methods has been established, as well as the validity of a naive algorithm based on the same neural model as presented for transfer learning.

The successful results obtained with MDL so far encourage future research tracks. Several problems emerge from the developed framework. First, the transfer objective 1 is valid for one target only. In practice, several target problems may occur, hence a multi-target variant of transfer has to be given. In particular, i.i.d. hypothesis consists in assuming infinitely-many targets with specific distributions. I am currently exploring an approach based on Pareto-optimality, implying a prior over the future and thus a new learning concept: *concern for future question*.

Another question of interest is the theoretical validity of such approaches. Unlike statistical learning which has a clear measure of quality (given by the risk), an approach based on MDL does not present any natural quality measure. Such a function has to be found before an equivalent of PAC learning can be developed. Additionally, incremental learning methods do not have access to the whole objective function 2 at all steps: only local optimizations are possible. A measure of the impact of this algorithmic simplification appears as a direct consequence. Finally, we aim at finding a geometric interpretation of these equations. An interesting track is offered by the domain of information geometry and probability distribution manifolds. Under some specific conditions, Kolmogorov complexity may be associated to a probability distributed, hence considered in the perspective of information geometry.

3 Approaches of analogy

Because of its crucial role in my researches on learning, I attach great importance to studying analogical reasoning. For now on, my researches focused on Hofstadter's micro-world [3] which presents highly general characteristics of analogical thought. I will work on this research track with Dr. Laurent Orseau.

Preliminary works have already given insightful results and promising perspectives. I chose to work on the development of a small prototype language generating Hofstadter's analogies (of the form **ABC : ABD :: IJK : IJL**, which has the be read "**ABD** is to **ABC** what **IJL** is to **IJK**"). Among other specifications, the proposed language had to be generative, ie. describe a dynamic generation rather than a static description (as opposed to the description in [2] for instance). For example, the string **ABC** will be generated by the program `alphabet,sequence,3`, which can be read as "*Consider the alphabet and take the sequence of first three letters*". Once such a language is defined, it is turned into a prefix-delimited binary code, the length of which measures an upper-bound of Kolmogorov complexity.

A more elaborated and general version of the language has been recently proposed. This memory-based language offers a flexibility in the management of prior knowledge of the user and offers a simple way to set priority to operations: its grammar enables any possible operator, as long as the operator can be put in long-term memory. The complexity of an element in memory is defined as the complexity of its depth in memory. A more rigorous presentation of this new framework including the considerations on memory will be presented at ICCBR 2017 Computational Analogy Workshop [7].

We propose a validation of our approach with a human experiment. In an online survey, we submitted a few Hofstadter's problems and asked participants for their most intuitive solution. We shew that the majority answers correspond to local minima of Kolmogorov complexity. These results are not yet published.

The proposed approach offers a tool to compare two results of an analogical problem when the generative instructions are given to the system. A first logical direction is to provide automatic instruction generation, hence software able to produce an optimal generative instruction for any complete analogy. Once this will done, I have to find a solution to an analogical problem (eg. find the best solution to **ABC:ABD::IJK:?**): because the space of solutions is infinite, it cannot be explored naively and research biases have to be found. In order to address these issues, I am currently working on a Python interpreter for the developed language.

4 Collaborations and side problems

In the context of my research, I have the opportunity to collaborate on several projects related to non-stationarity and transfer.

A first project is led jointly with Dr. Jérémie Sublime (ISEP) and Dr. Basarab Matei (Université Paris 13) and concerns collaborative clustering. Classic clustering consists in associating similar data together in *clusters*. Collaborative and

multi-view clustering is a framework in which several clustering algorithms are involved and try to influence each other. The algorithms do not produce the same number of clusters, the same underlying model nor the same final solution. This problem is closely related to transfer because it involves sharing information between several different domains. A first contribution has been proposed using operational research tools to select relevant collaborators in an existing probabilistic framework [9]. A brand new approach, based on complexity, is being developed: we expressed the problem of collaborative clustering in terms of data compression and worked on minimal assumptions to obtain a tractable model. This new perspective offers a theoretical background for a wide range of heuristic state-of-the-art approaches and inspires a new algorithm [8].

A new related collaboration has been engaged recently with Dr. Cristina Manfredotti, Pr. Juliette Dibie (AgroParisTech) and Dr. Fatiha Sais (LRI), exploring common approaches in transfer learning and structural mappings in knowledge bases (ontology alignment).

Finally, I am exploring the problem of Transfer Learning using boosting algorithms with Pr. Antoine Cornuéjols and Sema Akkoyunlu. The use of boosting may offer new perspectives on transfer in general and be beneficial to my understanding of transfer mechanisms. A first contribution has been submitted [1].

References

1. Cornuéjols, A., Akkoyunlu, S., Murena, P.A., Olivier, R.: Transfer learning by boosting projections from the target domain to the source domain, submitted to NIPS 2017
2. Cornuéjols, A., Ales-Bianchetti, J.: Analogy and induction : which (missing) link? In: Workshop "Advances in Analogy Research : Integration of Theory and Data from Cognitive, Computational and Neural Sciences". Sofia, Bulgaria (1998)
3. Hofstadter, D.: The copycat project: An experiment in nondeterminism and creative analogies. AI Memo 755, Artificial Intelligence Laboratory, Massachusetts Institute of Technology (1984)
4. Li, M., Vitanyi, P.M.: An Introduction to Kolmogorov Complexity and Its Applications. Springer Publishing Company, Incorporated, 3 edn. (2008)
5. Murena, P., Cornuéjols, A.: Minimum description length principle applied to structure adaptation for classification under concept drift. In: 2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016. pp. 2842–2849 (2016)
6. Murena, P.A., Cornuéjols, A., Dessalles, J.L.: Incremental learning with the minimum description length principle, accepted in International Joint Conference on Neural Networks 2017
7. Murena, P.A., Dessalles, J.L., Cornuéjols, A.: A complexity based approach for solving hofstadter's analogies, accepted at Computational Analogy Workshop, ICCBR 2017
8. Murena, P.A., Sublime, J., Matei, B., Cornuéjols, A.: Multi-view clustering with the principle of minimum description length, submitted to ECML 2017
9. Sublime, J., Matei, B., Murena, P.A.: Analysis of the influence of diversity in collaborative and multi-view clustering, accepted in International Joint Conference on Neural Networks 2017