ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Human Algorithmic Stability and Human Rademacher Complexity

Mehrnoosh Vahdat[1,2,*], Luca Oneto[1], Alessandro Ghio[3], Davide Anguita[3],
Mathias Funk[2] and Matthias Rauterberg[2]

1 - DITEN - University of Genova
Via Opera Pia 11A, I-16145 Genova - Italy

2 - Department of Industrial Design - Technical University Eindhoven
P.O. Box 513, 5600 MB Eindhoven - The Netherlands

3 - DIBRIS - University of Genova
Via Opera Pia 13, I-16145 Genova - Italy

Abstract. In Machine Learning (ML), the learning process of an algorithm given a set of evidences is studied via complexity measures. The way towards using ML complexity measures in the Human Learning (HL) domain has been paved by a previous study, which introduced Human Rademacher Complexity (HRC): in this work, we introduce Human Algorithmic Stability (HAS). Exploratory experiments, performed on a group of students, show the superiority of HAS against HRC, since HAS allows grasping the nature and complexity of the task to learn.

1 Introduction

Exploring the way humans learn is a major interest in Learning Analytics (LA) and Educational Data Mining (EDM) [1]. New advances in LA enable measuring, collecting and analyzing data about learners and their contexts: this allows exploring people's learning behavior, for example through state-of-the-art Machine Learning (ML) approaches, opening the door towards optimized and personalized education [2, 3].

While Analytics and Data Mining can effectively support learning through data analysis tools, recent works in cognitive psychology [4, 5, 6, 7] highlight how cross-fertilization between Machine Learning (ML) and Human Learning (HL) can also be widened to the extents of how people tackle new problems and extract knowledge from observations. For example, inquiry-guided HL focuses on contexts where learners are meant to discover knowledge rather than passively memorizing [8, 9]: the instruction begins with a set of observations to interpret, and the learner tries to analyze the data or solve the problem by the help of the guiding principles [10], resulting in a more effective educational approach [11]. Measuring the ability of a human to capture information rather than simply memorizing is thus key to optimize HL. In this sense, the parallelism with ML is straightforward: in this framework, several approaches in the last decades dealt

[*] This work was supported in part by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA Agency of the European Commission under EMJD ICE FPA n^o 2010-2012.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

with the development of measures to assess the generalization ability of learning algorithms, in order to minimize risks of overfitting.

As a consequence, the mash-up of ML studies on generalization ability estimation and HL has been proposed: for example, Zhu et al. [7] proposed the application of Rademacher Complexity (RC) approaches [12] to estimate human capability of extracting knowledge (Human Rademacher Complexity – HRC). Unfortunately, (H)RC requires that a set of models is aprioristically defined, which includes the models to be explored by the learner (being either an algorithm or a human) [13]. While this hypothesis is not always satisfied by ML methods (e.g. by k-Nearest Neighbors [14]), aprioristically defining a list of alternative models for humans is an even tougher task [15]. This leads to formulating further assumptions [7], which do not often hold in practice.

As an alternative, we propose to exploit Algorithmic Stability (AS) [16, 13] in the HL framework to compute the Human Algorithmic Stability (HAS), which does not rely on the definition of a set of models and does not require any additional assumptions. By comparatively benchmarking HRC and HAS, experiments performed by analyzing the way a group of students learns tasks of different difficulties show that using HAS leads to beneficial outcomes in terms of value of the performed analysis. In particular, HAS is influenced by the nature and the complexity of the task to learn. Moreover, contrarily to HRC, HAS is also able to capture the fast-learning ability of a human when dealing with simple tasks: this allows providing new perspectives with reference to human tendency to overfit training data depending on the nature of the problem faced. These results can thus play, in a virtuous loop cycle between ML and HL, as a measure of the propensity of the learner towards inquiry-based learning versus simple memorization.

2 Rademacher Complexity (RC) and Algorithmic Stability (AS) in Machine Learning (ML)

Let $\mathcal{X}$ and $\mathcal{Y} \in \{\pm 1\}$ be, respectively, an input and an output space. We consider m sets $\mathcal{D}^m = \{\mathcal{S}_n^1, \dots, \mathcal{S}_n^m\}$ of n labeled i.i.d. data $\mathcal{S}_n^j : \{Z_1^j, \dots, Z_n^j\}$ $(j = 1, ..., m)$, where $Z^j_{i \in \{1,\dots,n\}} = (X_i^j, Y_i^j)$, where $X_i^j \in \mathcal{X}$ and $Y_i^j \in \mathcal{Y}$, sampled from an unknown distribution μ. A learning algorithm $\mathcal{A}$ is used to train a model $f : \mathcal{X} \to \mathcal{Y} \in \mathcal{F}$ based on $\mathcal{S}_n^j$.

In this framework, the ability of $\mathcal{A}$ to identify an effective f is assessed through complexity measures, which indicate the tendency of the models to overfit the training data. Rademacher Complexity (RC) is a now-classic measure, used for this purpose:

$$\hat{R}_n(\mathcal{F}) = \frac{1}{m} \sum_{j=1}^{m} \left[1 - \inf_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \ell(f, (X_i^j, \sigma_i^j)) \right] \tag{1}$$

where $\ell(\cdot, \cdot)$ is the hard loss function which counts the number of misclassified examples [13], while σ_i^j (with $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$) are ± 1 valued random variable for which $\mathbb{P}[\sigma_i^j = +1] = \mathbb{P}[\sigma_i^j = -1] = 1/2$ [12]. In other words, the complexity of the model is assessed through its capacity in learning noise (i.e. random labels); however, the set of functions $\mathcal{F}$ must be aprioristically defined, and this is not always possible in practice.

As an alternative, Algorithmic Stability (AS) has been recently proposed: it measures the ability of $\mathcal{A}$ to select similar models, even though training data are (slightly) perturbed. This guarantees that the algorithm is actually learning from data, and it is not simply memorizing them. In order to measure AS, we define the sets $\mathcal{S}_n^{j\setminus i} = \mathcal{S}_n^j \setminus \{Z_i^j\}$, where the i-th sample is removed. AS is computed as [16, 13]:

$$\hat{H}_n\left(\mathcal{A}\right) = \frac{1}{m}\sum_{j=1}^{m} |\ell(\mathcal{A}_{\mathcal{S}_n^j}, Z_{n+1}^j) - \ell(\mathcal{A}_{\mathcal{S}_n^{j\setminus i}}, Z_{n+1}^j)|. \qquad (2)$$

It is worth underlining that AS does not require $\mathcal{F}$ to be known, while only the algorithm $\mathcal{A}$ must be defined to compute the measure.

3 From Machine Learning (ML) to Human Learning (HL): Experimental Design & Results

A previous work [7] depicts an experiment targeted towards identifying the Human Rademacher Complexity (HRC), which is defined as a measure of the capability of a human to learn noise, thus avoiding to overfit data. Given the drawbacks of RC with respect to AS, highlighted in the previous section, we built on the experiments and experience of these activities to design and carry out a new experiment, which aims at estimating the average HRC and Human Algorithmic Stability (HAS) for a class of students: the latter is defined as a measure of the capability of a learner to understand a phenomenon, even when he is given slightly perturbed observations. We target comparing the two quantities and verifying which one is the most informative for getting more insights on HL: 307 undergraduate students were involved through questionnaires, designed as described in the following. Filled questionnaires were collected and analyzed: (anonymized) examples can be found at `http://smartlab.ws/files/questionnaires.zip`.

In particular, we followed the approach designed in [7] and we modified it where appropriate to estimate HAS. The first step consists in defining the phenomena and the rules, that must be grasped by students. Two domains are defined: Shape and Word. The former domain consists of 321 computer-generated 3D shapes, parametrized by $\alpha \in [-8, +8]$, such that a small value of α leads to spiky shapes, while a large α allows to obtain smooth ones. A label is assigned to each shape, and two problems are defined in accordance with ad hoc rules to depict tasks of increasing complexity: Shape Simple (SS), where $Y = +1$ if $\alpha \leq 0$ and $Y = -1$ otherwise; Shape Difficult (SD), where $Y = +1$ if $-4 \leq \alpha \leq 4$ and $Y = -1$ otherwise. The Word domain, instead, consists of 321 words[1], sampled from the Wisconsin Perceptual Attribute Ratings Database [17], which includes words rated by 350 undergraduates based on their emotional valence. Two rules

[1] Having to deal with Italian students only, words have been translated into Italian.

are defined for labelling data, analogously to [7] and to what done above: in Word Simple (WS), words are sorted by their length and the 161 longest ones are assigned $Y = +1$; in Word Difficult (WD), words are sorted by their emotion valence and the 161 most positive ones are assigned $Y = +1$. The probability distribution throughout both domains is uniform.

HRC requires two assumptions to be made: every individual picks up the model from the same set of alternatives (i.e., in ML terms, from the same $\mathcal{F}$); every individual always performs at his best (i.e. in ML terms, the error is minimized). In order to compute HRC, the same procedure of [7] has been adopted. In particular, two domains are identified – Rademacher Shape (RS) and Rademacher Word (RW), while labels are not contemplated when deriving HRC.

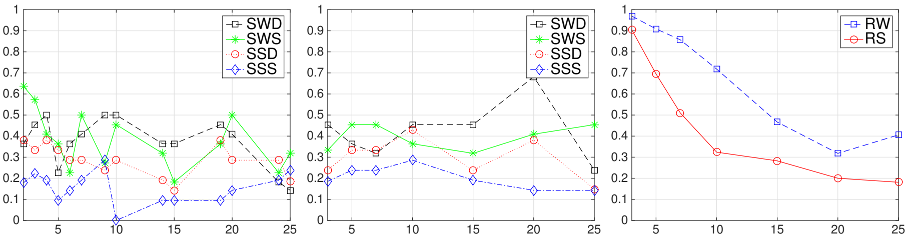
A new experimental protocol has been designed, instead, for HAS. With reference to the defined domains and since labels influence the estimation of HAS, four different tasks are identified: Word Simple (SWS), Word Difficult (SWD), Shape Simple (SSS) and Shape Difficult (SSD). Given a domain and a rule, a dataset $\{Z_1^j, \dots, Z_n^j\}$ with $j = \{1, \dots, 307\}$, is sampled for each participant. The size of the sets are varied for different individuals, and randomly chosen in $n \in \{3, 5, 7, 10, 15, 20, 25\}$. Moreover, for computing HAS, a further pattern $Z_{n+1}^j = \{X_{n+1}^j, Y_{n+1}^j\}$ is also selected (as indicated by Eq. (2)).

Students have been initially asked to complete the questionnaires according to the following protocol: (i) 2 minutes are given to students in order to capture the underlying rule from $n-1$ labeled samples $\{Z_1^j, \dots, Z_{n-1}^j\}$; (ii) participants perform a filler task consisting of some two-digit addition / subtraction questions, to reduce risks of memorization; (iii) student must classify X_{n+1}^j; (iv) students are asked to describe the rule they identified, and to estimate the confidence of their decision; (v) the complete set $\{Z_1^j, \dots, Z_n^j\}$ is given to students (in shuffled order), but participants are not aware that $n-1$ samples are the same as step (i); (vi) another filler task is provided; (vii) sample X_{n+1}^j is given to the individuals for labeling; (viii) participants are asked again for describing the rule they inferred. Unfortunately, a first trial, conducted on 70 volunteers, showed that students were able to mentally link steps (iii) and (vii). We thus modified the protocol, so that the eight phases are equally split (i÷iv and v÷viii) between two students. Thanks to this procedure, all quantities, necessary to compute HAS, can be derived.

In the end, each experiment was consisted of two subjects, who worked on a unique combination of $\{\text{SSS}_{n-1}, \text{SWD}_{n-1}, \text{SSD}_{n-1}, \text{SWS}_{n-1}, \text{RW}_n\}$ and $\{\text{SSS}_n, \text{SWD}_n, \text{SSD}_n, \text{SWS}_n, \text{RS}_n\}$, for the different set sizes. For each student, it is thus possible to compute $\hat{R}_n(\mathcal{F})$ and $\hat{H}_n(\mathcal{A})$. In order to obtain further insights, it is also possible to compute the average empirical error, performed by all students when classifying samples at steps (iii) and (vii):

$$\hat{L}_n(\mathcal{A}) = \tfrac{1}{m} \textstyle\sum_{j=1}^m \ell(\mathcal{A}_{\mathcal{S}_n^j}, Z_{n+1}^j). \tag{3}$$

Figure 1(a) shows the trend of $\hat{L}_n(\mathcal{A})$, as n is varied: as expected from ML



(a) $\hat{L}_n(\mathcal{A})$ (b) $\hat{H}_n(\mathcal{A})$ (c) $\hat{R}_n(\mathcal{F})$

Fig. 1: Experimental results, as n is varied.

theory, $\hat{L}_n(\mathcal{A})$ is smaller for simple tasks than for difficult ones in HL as well. However, analogies end here. While the error of ML models usually decreases with n, results on HL are characterized by oscillations, even for small variations of n. This can be due to the small sample considered, and especially to the fact that only a subset of the students are willing to perform at their best when completing the questionnaire: these phenomena could be explored, in the future, by analyzing the filler tasks, in order to verify the students' level of attention. Another result is in contrast with what expected from the ML point of view: oscillations in terms of $\hat{L}_n(\mathcal{A})$ mostly (and surprisingly) affect simple tasks (SSS, SWS). Moreover, errors performed with reference to the Shape domain are generally smaller than those recorded for the Word domain. Broadly speaking, humans are not learning algorithms, thus more vertical HL interpretations of these effects by experts will be necessary.

Figure 1(b) presents the results obtained when computing $\hat{H}_n(\mathcal{A})$ as n is varied. Despite having being designed in the ML framework, it is worth highlighting how HAS is able to grasp the nature and peculiarities of HL. As a matter of fact, we note that: simple tasks are characterized by smaller values of $\hat{H}_n(\mathcal{A})$; HAS for the Shape domain is generally smaller than for the Word domain. Both results are in accordance with the trend of the error, registered in HL, and the nature of the analyzed phenomenon: in this sense, HAS offers interesting insights on HL, because it raises questions about the ability of humans to learn in different domains.

Finally, Figure 1(c) shows the trend for $\hat{R}_n(\mathcal{F})$. Contrarily to HAS, HRC is not able to grasp the complexity of the task, since labels are neglected when computing $\hat{R}_n(\mathcal{F})$. Moreover, the two assumptions, underlying the computation of HRC, do not hold in practice: in fact, the learning process of an individual should be seen as a multifaceted problem, rather than a collection of factual and procedural knowledge, targeted towards minimizing a "cost" [15]. This leads to less significative results with respect to HL: HRC decreases with n (as in ML), and this trend is substantially uncorrelated with the errors for the considered domains.

Future evolutions will allow to analyze students' level of attention through the results of the filler task, and to include other heterogeneous domains (e.g.

mathematics) and additional rules, of increasing complexity: this will enable to more carefully explore how the domain and the task complexity influence Human Algorithmic Stability, and how this is related to the error performed on classifying new samples. Furthermore, we will make the dataset of anonymized questionnaires publicly available.

References

[1] Z. Papamitsiou and A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4):49–64, 2014.

[2] G. Siemens and P. Long. Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5):30–32, 2011.

[3] M. Bienkowski, M. Feng, and B. Means. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *US Department of Education, Office of Educational Technology*, pages 1–57, 2012.

[4] J. Feldman. Minimization of boolean complexity in human concept learning. *Nature*, 407(6804):630–633, 2000.

[5] N. Chater and P. Vitányi. Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, 7(1):19–22, 2003.

[6] T. L. Griffiths, B. R. Christian, and M. L. Kalish. Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, 32(1):68–107, 2008.

[7] X. Zhu, B. R. Gibson, and T. T. Rogers. Human rademacher complexity. In *Neural Information Processing Systems*, pages 2322–2330, 2009.

[8] A. Kruse and R. Pongsajapan. Student-centered learning analytics. In *CNDLS Thought Papers*, 2012.

[9] V. S. Lee. What is inquiry-guided learning? *New directions for teaching and learning*, 2012(129):5–14, 2012.

[10] V. S. Lee. The power of inquiry as a way of learning. *Innovative Higher Education*, 36(3):149–160, 2011.

[11] T. DeJong, S. Sotiriou, and D. Gillet. Innovations in stem education: the go-lab federation of online labs. *Smart Learning Environments*, 1(1):1–16, 2014.

[12] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[13] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Fully empirical and data-dependent stability-based bounds. *IEEE Transactions on Cybernetics*, 10.1109/TCYB.2014.2361857:in–press, 2014.

[14] P. Klesk and M. Korzen. Sets of approximating functions with finite vapnik–chervonenkis dimension for nearest-neighbors algorithms. *Pattern Recognition Letters*, 32(14):1882–1893, 2011.

[15] D. L. Schacter, D. T. Gilbert, and D. M. Wegner. *Psychology, Second Edition*. Worth Publishers, 2010.

[16] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[17] D. A. Medler, A. Arnoldussen, J. R. Binder, and M. S. Seidenberg. *The Wisconsin Perceptual Attribute Ratings Database*. http://www.neuro.mcw.edu/ratings/, 2005.