Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Generalization bounds for learning weighted automata

Borja Balle [a,*], Mehryar Mohri [b,c]

[a] Department of Mathematics and Statistics, Lancaster University, Lancaster, UK
[b] Courant Institute of Mathematical Sciences, New York University, NY, USA
[c] Google Research, New York, NY, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Learning theory
Generalization bounds
Weighted automata
Rademacher complexity

ABSTRACT

This paper studies the problem of learning weighted automata from a finite sample of strings with real-valued labels. We consider several hypothesis classes of weighted automata defined in terms of three different measures: the norm of an automaton's weights, the norm of the function computed by an automaton, and the norm of the corresponding Hankel matrix. We present new data-dependent generalization guarantees for learning weighted automata expressed in terms of the Rademacher complexity of these classes. We further present upper bounds on these Rademacher complexities, which reveal key new data-dependent terms related to the complexity of learning weighted automata.

© 2017 Published by Elsevier B.V.

1. Introduction

Weighted finite automata (WFAs) provide a general and highly expressive framework for representing functions mapping strings to real numbers. The mathematical theory behind WFAs, that of rational power series, has been extensively studied in the past [31,54,39,17] and has been more recently the topic of a dedicated handbook [28]. WFAs are widely used in modern applications, perhaps most prominently in image processing and speech recognition where the terminology of *weighted automata* seems to have been first introduced and made popular [24,46,52,44,48], in several other speech processing applications such as speech synthesis [56,3], in phonological and morphological rule compilation [36,37,50], in parsing [53,47], machine translation [25,2], bioinformatics [30,4], sequence modeling and prediction [22], formal verification and model checking [6,5], in optical character recognition [20], and in many other areas.

The recent developments in spectral learning [34,8] have triggered a renewed interest in the use of WFAs in machine learning, with several recent successes in natural language processing [10,11] and reinforcement learning [19,33]. The interest in spectral learning algorithms for WFAs is driven by the many appealing theoretical properties of such algorithms, which include their polynomial-time complexity, the absence of local minima, statistical consistency, and finite sample bounds *à la* PAC [34]. A refined analysis of such algorithms provides sample bounds which are independent of the size of the Hankel matrices used by the learning algorithm [26]. However, the typical statistical guarantees given for the hypotheses used in spectral learning only hold in the realizable case. That is, these analyses assume that the labeled data received by the algorithm is sampled from some unknown WFA. While this assumption is a reasonable starting point for theoretical analyses, the results obtained in this setting fail to explain the good performance of spectral algorithms in many practical applications where the data is typically not generated by a WFA. See [13] for a recent survey of algorithms for learning WFAs with a discussion of the different assumptions and learning models.

* Corresponding author.
E-mail addresses: b.deballepigem@lancaster.ac.uk (B. Balle), mohri@cs.nyu.edu (M. Mohri).

https://doi.org/10.1016/j.tcs.2017.11.023
0304-3975/© 2017 Published by Elsevier B.V.

Please cite this article in press as: B. Balle, M. Mohri, Generalization bounds for learning weighted automata, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2017.11.023

There exists of course a vast literature in statistical learning theory providing tools to analyze generalization guarantees for different hypothesis classes in classification, regression, and other learning tasks. These guarantees typically hold in an agnostic setting where the data is drawn i.i.d. from an arbitrary distribution. For spectral learning of WFAs, an algorithm-dependent agnostic generalization bound was proven in [12] using a stability argument. This seems to have been the first analysis to provide statistical guarantees for learning WFAs in an agnostic setting. However, while [12] proposed a broad family of algorithms for learning WFAs parametrized by several choices of loss functions and regularizations, their bounds hold only for one particular algorithm within that family.

In this paper, we start the systematic development of algorithm-independent generalization bounds for learning with WFAs, which apply to all the algorithms proposed in [12], as well as to others using WFAs as their hypothesis class. Our approach consists of providing upper bounds on the Rademacher complexity of general classes of WFAs. The use of Rademacher complexity to derive refined generalization bounds is standard [38] (see also [16] and [49]). It has been successfully used to derive learning guarantees for classification, regression, kernel learning, ranking, and many other machine learning tasks (e.g. see [49] and references therein). A key benefit of Rademacher complexity analyses is that the resulting generalization bounds are data-dependent.

Our main results consist of upper bounds on the Rademacher complexity of three broad classes of WFAs. The main difference between these classes is the quantities used for their definition: the norm of the transition weight matrix or initial and final weight vectors of a WFA; the norm of the function computed by a WFA; and, the norm of the Hankel matrix associated to the function computed by a WFA. The formal definitions of these classes is given in Section 3. Let us point out that our analysis of the Rademacher complexity of the class of WFAs described in terms of Hankel matrices directly yields theoretical guarantees for a variety of spectral learning algorithms. We will return to this point when discussing the application of our results. As an application of our Rademacher complexity bounds we provide a variety of generalizations bounds for learning with WFAs using a bounded Lipschitz loss function; our bounds include both data-dependent and data-independent bounds.

Related work To the best of our knowledge, this paper is the first to provide general tools for deriving learning guarantees for broad classes of WFAs. However, there exists some related work providing complexity bounds for some sub-classes of WFAs in agnostic settings. The VC-dimension of deterministic finite automata (DFAs) with n states over an alphabet of size k was shown by [35] to be in $O(kn \log n)$. This can be used to show that the Rademacher complexity of this class of DFA is bounded by $O(\sqrt{nk \log n/m})$. For probabilistic finite automata (PFAs), it was shown by [1] that, in an agnostic setting, a sample of size $\tilde{O}(kT^2n^2/\varepsilon^2)$ is sufficient to learn a PFA with n states and k symbols whose log-loss error is at most ε away from the optimal one in the class when the error is measured on all strings of length T. New learning bounds on the Rademacher complexity of DFAs and PFAs follow as straightforward corollaries of the general results we present in this paper.

Another recent line of work, which aims to provide guarantees for spectral learning of WFAs in the non-realizable setting, is the so-called low-rank spectral learning approach [41]. This has led to interesting upper bounds on the approximation error between minimal WFAs of different sizes [40]. See [15] for a polynomial-time algorithm for computing these approximations. This approach, however, is more limited than ours for two reasons: first, because it is algorithm-dependent; second, because it assumes that the data is actually drawn from some (probabilistic) WFA, albeit one that is larger than any of the WFAs in the hypothesis class considered by the algorithm.

The rest of this paper is organized as follows. Section 2 introduces the notation and technical concepts used throughout. Section 3 describes the three classes of WFAs for which we provide Rademacher complexity bounds. The bounds are formally stated and proven in Sections 4, 5, and 6. In Section 7 we provide additional bounds required for converting some sample-dependent bounds from Sections 5 and 6 into sample-independent bounds. Finally, the generalizations bounds obtained using the machinery developed in previous sections are given in Section 8.

2. Preliminaries

2.1. Weighted automata, rational functions, and Hankel matrices

Let Σ be a finite alphabet of size k. Let ϵ denote the empty string and $\Sigma^\star$ the set of all finite strings over the alphabet Σ. The length of $u \in \Sigma^\star$ is denoted by $|u|$. Given an integer $L \geq 0$, we denote by $\Sigma^{\leq L}$ the set of all strings with length at most L: $\Sigma^{\leq L} = \{x \in \Sigma^\star : |x| \leq L\}$. Given two strings $u, v \in \Sigma^\star$ we write uv for their concatenation.

A WFA over the alphabet Σ with $n \geq 1$ states is a tuple $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\}_{a\in\Sigma}\rangle$ where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ are the initial and final weights, and $\mathbf{A}_a \in \mathbb{R}^{n\times n}$ the transition matrix whose entries give the weights of the transitions labeled with a. Every WFA A defines a function $f_A : \Sigma^\star \to \mathbb{R}$ defined for all $x = a_1 \cdots a_t \in \Sigma^\star$ by

$$f_A(x) = f_A(a_1 \cdots a_t) = \boldsymbol{\alpha}^\top \mathbf{A}_{a_1} \cdots \mathbf{A}_{a_t} \boldsymbol{\beta} = \boldsymbol{\alpha}^\top \mathbf{A}_x \boldsymbol{\beta} \ , \tag{1}$$

where $\mathbf{A}_x = \mathbf{A}_{a_1} \cdots \mathbf{A}_{a_t}$. This algebraic expression in fact corresponds to summing the weights of all possible paths in the automaton indexed by the symbols in x, where the weight of a single path $(q_0, q_1, \ldots, q_t) \in [n]^{t+1}$ is obtained by multiplying the initial weight of q_0, the weights of all transitions from q_{s-1} to q_s labeled by x_s, and the final weight of state q_t, that is

Please cite this article in press as: B. Balle, M. Mohri, Generalization bounds for learning weighted automata, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2017.11.023

Download English Version:

https://daneshyari.com/en/article/6875592

Download Persian Version:

https://daneshyari.com/article/6875592

Daneshyari.com