Journal of Computer and System Sciences 79 (2013) 622–639

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Inferring descriptive generalisations of formal languages ☆

Dominik D. Freydenberger [a,*], Daniel Reidenbach [b]

[a] Institut für Informatik, Johann Wolfgang Goethe-Universität, Postfach 111932, 60054 Frankfurt am Main, Germany
[b] Department of Computer Science, Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom

ARTICLE INFO

Article history:
Received 17 January 2011
Received in revised form 18 October 2012
Accepted 30 October 2012
Available online 11 December 2012

Keywords:
Inductive inference
Descriptive generalisation
Pattern languages
Descriptive patterns
Upper approximate identification from positive data

ABSTRACT

In the present paper, we introduce a variant of Gold-style learners that is not required to infer precise descriptions of the languages in a class, but that must find descriptive patterns, i.e., optimal generalisations within a class of pattern languages. Our first main result characterises those indexed families of recursive languages that can be inferred by such learners, and we demonstrate that this characterisation shows enlightening connections to Angluin's corresponding result for exact inference. Furthermore, this result reveals that our model can be interpreted as an instance of a natural extension of Gold's model of language identification in the limit. Using a notion of descriptiveness that is restricted to the natural subclass of terminal-free E-pattern languages, we introduce a generic inference strategy, and our second main result characterises those classes of languages that can be generalised by this strategy. This characterisation demonstrates that there are major classes of languages that can be generalised in our model, but not be inferred by a normal Gold-style learner. Our corresponding technical considerations lead to insights of intrinsic interest into combinatorial and algorithmic properties of pattern languages.

1. Introduction

In Gold's intensively studied learning paradigm of language identification in the limit from positive data (cf. Gold [12]), it is a requirement for the computational learner to infer, for any positive presentation of any language in some class, an *exact* description of that language. While this maximum accuracy of the output of the inference procedure is clearly a natural goal, it has a number of downsides, the most obvious one being the fact that it can lead to significant limitations to the learning power of the model. From a more applied point of view, there is another important reason why one might wish to relax it and settle for receiving an approximation of the language from the learner: depending on the class of languages to be inferred, the corresponding grammars or acceptors might have undesirable properties, i.e., they might have computationally hard decision problems or be incomprehensible to a (human) user. Thus, in various settings it might be perfectly acceptable for an inference procedure to output a compact and reasonably precise approximation of the language instead of producing a precise yet arbitrarily complex grammar.

In the present paper, we introduce and study such a variant of Gold's model, where the requirement of exact language identification is dropped and replaced with that of inference of easily interpretable approximations. More precisely, we consider a learner that, for any language it reads, must converge to a *consistent pattern*, i.e., a finite string that consists of variables and of terminal symbols and that can be turned into any word of the language by substituting arbitrary strings

☆ A preliminary version [10] of this work was presented at COLT 2010.

* Corresponding author.
E-mail addresses: freydenberger@em.uni-frankfurt.de (D.D. Freydenberger), D.Reidenbach@lboro.ac.uk (D. Reidenbach).

0022-0000/$ – see front matter
http://dx.doi.org/10.1016/j.jcss.2012.10.001

of terminal symbols for the variables. In addition to being seen as mere descriptions of common features of words in a given language, such a pattern α can also be interpreted as a generator of a formal language $L(\alpha)$, the so-called *pattern language* (cf. Angluin [1]), which is simply the maximum set of words the pattern is consistent with. Hence, referring to this terminology, we can state that our learner has to output a pattern generating a language that is a superset of the input language, which means that our approach does not yield an arbitrary approximation of a language, but rather a *generalisation*. Even though many classes of pattern languages have a number of NP-complete or undecidable basic decision problems (see, e.g., Angluin [1], Jiang et al. [16], and Freydenberger and Reidenbach [8]), patterns (or related concepts, such as regular expressions and their various extensions implemented in today's programming languages and text editors, see Câmpeanu et al. [4]) are widely used when commonalities of words are to be specified or interpreted by a human user, which demonstrates that they are a worthwhile concept in the context of our paper.

When inferring consistent patterns instead of precise descriptions, it is of course vital to develop and employ a notion of high-quality patterns, so that the inference procedure does not lead to an overly imprecise result. Otherwise, the learner could always output the pattern $\alpha := x_1$ (where x_1 is a variable), which is consistent with every language, and this approach would obviously neither lead to a rich theory nor to practically relevant results. In our model, the inference procedure shall therefore be required to converge to a pattern δ that is *descriptive* of the language L (with respect to a class $\mathrm{PAT}_\star$ of pattern languages). This means that δ must be consistent with L, $L(\delta)$ must be included in $\mathrm{PAT}_\star$, and there is no pattern δ' satisfying $L(\delta') \in \mathrm{PAT}_\star$ and $L \subseteq L(\delta') \subset L(\delta)$; in other words, a pattern is descriptive of a language if there is no other pattern providing a closer match for the language. Since descriptiveness captures a natural understanding of patterns providing a desirable generalisation of languages and, furthermore, descriptive patterns can be used to devise Gold-style learners precisely identifying classes of pattern languages from positive data, this concept has been thoroughly investigated (see, e.g., Angluin [1], Jiang et al. [16], and Freydenberger and Reidenbach [9]), and the same holds for optimal approximations of other types of languages (see, e.g., Arimura et al. [3]). While established definitions of descriptiveness often restrict their view to patterns covering finite languages and normally use the full class of E- or NE-pattern languages (to be formally introduced in Section 2) as the class $\mathrm{PAT}_\star$ of admissible pattern languages, we allow a descriptive pattern to cover a finite or an infinite language, and we have a class $\mathrm{PAT}_\star$ that can be arbitrarily chosen. Both of these extensions of the original definition are absolutely straightforward.

To summarise our model of inference, we consider a learner that reads a positive presentation of a language and, after having seen a new input word, outputs a pattern, the so-called *hypothesis*. We then say that, for a class $\mathcal{L}$ of languages and a class $\mathrm{PAT}_\star$ of pattern languages, the learner $\mathrm{PAT}_\star$-*descriptively generalises* $\mathcal{L}$ if and only if, for every positive presentation of every language $L \in \mathcal{L}$, the sequence of hypotheses produced by the learner converges to a pattern δ that is descriptive of L with respect to the class $\mathrm{PAT}_\star$. A more formal definition of our model is given in Section 4.1.

While the focus of research in inductive inference from positive data has been on exact identification, there are quite a few of studies of paradigms where – either directly or indirectly – approximations of languages are inferred; however, the motivation for this research often differs substantially from ours as described at the beginning of the present section. We now shall briefly summarise these approaches in order to highlight the differences to our model. Mukouchi [23] introduces the concept of *strong-minimal inference* and *minimal inference*, where the learner needs to converge to a minimum generalisation for *any* language (or, respectively, for any language where such a minimum generalisation exists among the admissible hypotheses). Hence, the model does not support an explicit restriction to a specific class $\mathcal{L}$ of languages that need to be generalised by the learner. The notion of *upper approximate identification* by Kobayashi and Yokomori [19] considers the inferrability of minimum generalisations of languages and features an explicit split between a class $\mathcal{L}$ of languages to be generalised and a class of admissible hypotheses. Hence – apart from the fact that our model, unlike upper approximate identification, is not restricted to indexed families $\mathcal{L}$, but rather restricts the nature of the hypotheses – this model is virtually identical to our approach. However, the focus of the paper [19] (and of subsequent studies; see, e.g., Kobayashi and Yokomori [20] and Fernau [6]) is on a different inference paradigm, namely *upper-best approximate identification*, where the topology of the class of hypotheses is restricted, as it needs to contain, for every language to be generalised, a semantically unique minimal generalisations. Such a property does not hold for the classes of pattern languages we shall use as a hypothesis space, and many of our technical results deal with exactly this aspect of *competing* descriptive generalisations, i.e., the existence of several descriptive patterns with incomparable languages for the language to be inferred. Finally, Jain and Kinber [14] introduce the model of *ResAllMWSubEx*, which does not directly draw its motivation from the wish of investigating the inference of approximations of languages, but is nevertheless similar to upper approximate identification studied by Kobayashi and Yokomori [19]. More precisely, this model again considers convergence to a minimal generalisation of the language that is presented to the learner, but it allows hardly any control over these languages, since they can be any sublanguage of any language in the hypothesis space.

In summary, the main difference between our notion of descriptive generalisation and all related approaches described above is that we have a distinct split between a class $\mathcal{L}$ of languages to be inferred and an arbitrary class $\mathrm{PAT}_\star$ of pattern languages determining the set of admissible hypotheses. This leads to a compact and powerful model that yields interesting insights into the question of to which extent the generalisability of $\mathcal{L}$ depends on properties of $\mathcal{L}$ or of $\mathrm{PAT}_\star$. We discuss this topic in Section 4.2, and we demonstrate in Section 4.3 that descriptive generalisation can be interpreted as a natural instance of a very general and simple inference model which, to the best of our knowledge, has not been considered so far.

In Section 5, we investigate our model for a fixed and rich class $\mathrm{PAT}_\star$, namely the class of *terminal-free E-pattern languages*, i.e., the class of all pattern languages generated by patterns not containing any terminal symbols, where the empty