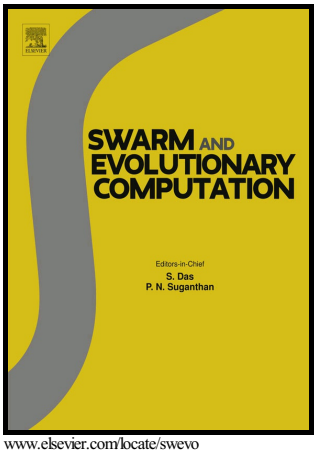

www.elsevier.com/locate/swevo

Maintaining regularity and Generalization in data using the minimum description length principle and genetic algorithm: Case of grammatical inference

Hari Mohan Pandey, Ankit Chaudhary, Deepti Mehrotra, Graham Kendall

PII: S2210-6502(16)30024-4
DOI: http://dx.doi.org/10.1016/j.swevo.2016.05.002
Reference: SWEVO215

To appear in: *Swarm and Evolutionary Computation*

Received date: 23 April 2015
Revised date: 4 May 2016
Accepted date: 16 May 2016

Cite this article as: Hari Mohan Pandey, Ankit Chaudhary, Deepti Mehrotra and Graham Kendall, Maintaining regularity and Generalization in data using the minimum description length principle and genetic algorithm: Case of grammatical inference, *Swarm and Evolutionary Computation*, http://dx.doi.org/10.1016/j.swevo.2016.05.002

Maintaining Regularity and Generalization in Data using the Minimum Description Length Principle and Genetic Algorithm: Case of Grammatical Inference

Hari Mohan Pandey[a*], Ankit Chaudhary[b], Deepti Mehrotra[c], Graham Kendall[d],

[a]Department of Computer Science & Engineering, Amity University Uttar Pradesh,Sector 125, Noida, India
[b]Department of Computer Science, Truman State University, USA
[c]Amity School of Engineering & Technology, Amity University, Sector 125, Noida, India
[d]The University of Nottigham Malaysia CampusJalan Borga, 43500 Semenyih, Selangor Darul Ehsan, Malaysia

[*]Corresponding author.

ABSTRACT

In this paper, a genetic algorithm with minimum description length (GAWMDL) is proposed for grammatical inference. The primary challenge of identifying a language of infinite cardinality from a finite set of examples should know when to generalize and specialize the training data. The minimum description length principle that has been incorporated addresses this issue is discussed in this paper. Previously, the e-GRIDS learning model was proposed, which enjoyed the merits of the minimum description length principle, but it is limited to positive examples only. The proposed GAWMDL, which incorporates a traditional genetic algorithm and has a powerful global exploration capability that can exploit an optimum offspring. This is an effective approach to handle a problem which has a large search space such the grammatical inference problem. The computational capability, the genetic algorithm poses is not questionable, but it still suffers from premature convergence mainly arising due to lack of population diversity. The proposed GAWMDL incorporates a bit mask oriented data structure that performs the reproduction operations, creating the mask, then Boolean based procedure is applied to create an offspring in a generative manner. The Boolean based procedure is capable of introducing diversity into the population, hence alleviating premature convergence. The proposed GAWMDL is applied in the context free as well as regular languages of varying complexities. The computational experiments show that the GAWMDL finds an optimal or close-to-optimal grammar. Two fold performance analysis have been performed. First, the GAWMDL has been evaluated against the elite mating pool genetic algorithm which was proposed to introduce diversity and to address premature convergence. GAWMDL is also tested against the improved tabular representation algorithm. In addition, the authors evaluate the performance of the GAWMDL against a genetic algorithm not using the minimum description length principle. Statistical tests demonstrate the superiority of the proposed algorithm. Overall, the proposed GAWMDL algorithm greatly improves the performance in three main aspects: maintains regularity of the data, alleviates premature convergence and is capable in grammatical inference from both positive and negative corpora.

Abbreviations

ANS, Accepting negative sample; ; APS. Accepting positive sample; BMODA. Bit masking oriented data structure; BNF. Backus Naur Form; BBP. Boolean based procedure; CFL. Context free language; CFG. Context free grammar; CS. Chromosome size; CM. Crossmask/crossover mask; CR. Crossover rate; DFA. Deterministic finite automata; DL. Description Length; DSL. Domain-Specific Language; EA. Evolutionary algorithm; EMP. Elite Mating Pool Genetic Algorithm; GI. Grammatical inference; GA. Genetic algorithm with minimum description length; GAW. Genetic Algorithm without Minimum Description Length; GP. Genetic Programming; GA. Genetic algorithm; ITBL. Improved Tabular Representation Algorithm; M. Model; MM. Mutmask/mutation mask; MDL. Minimum description length; NN. Neural Network; MA. Memetic Algorithm; MR. Mutation rate; NPR. Maximum number of allowable grammar rules; PAC. Probably Approximately Correct; PRL. Production rule length; PDA. Pushdown automata; PS. Population size; RNN. Recurrent Neural Network; RNS. Rejecting negative sample; RPS. Rejecting positive sample; RL. Regular language; SOM. Self-organizing Map; SNR. Signal to noise ratio; TBLA. Tabular Representation Algorithm

Keywords: Bit-masking oriented data structure, Context free grammar,Genetic Algorithm,Grammar induction,Learning algorithm, Minimum description length principle

1. Introduction

The problem with inductive and statistical inference systems is to maintain regularity in the data. In other words "How to take decisions for selecting an appropriate model that should present the competing explanation of the data using limited observations?" Figure 1 shows a scenario where a sender who want to transmit some data to the receiver and, is interested in selecting the best model which can maximally compress the observed data and deliver it to the receiver using as few bits as possible.

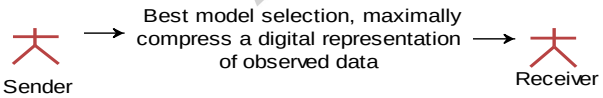

Fig. 1. A scenario showing the rationale of using the MDL principle. The sender wants to transmit some data to the receiver.

Formally, the selection of the best model is the process of deciding among the model classes based on the data. The Principle of Parsimony (Occam's razor) is the soul of the model selection, states that "given a choice of theories, the simplest is preferable" [4] [5]. The purpose of implementing the Parsimony Principle is to find a model, which can best fit the data. Rissanen extracted the essence of the Occam's theory and presented the Principle of Minimum Description Length states that "choose the model that gives the shortest description of data" [4] [12].

The domain of inquiry in this paper is the GI problem. A grammar can be constructed without using the MDL principle, but does not reflect any regularity in the data (Figure 2 (a)). In addition, it is difficult to know when to generalize and specialize the training data. In such situations, the constructed grammar is considered as a very simple grammar, because it simply provides the validity of any