On the Use of Context Sensitive Grammars in Grammatical Evolution For Legal Non-Compliance Detection

Carl Im
iolex
Hong Kong, China
carl.im@iolex.com

Erik Hemberg
MIT, CSAIL
Cambridge, USA
hembergerik@csail.mit.edu

ABSTRACT

We extend the context-free grammar mapping method in the Grammatical Evolution search heuristic. Grammatical Evolution guarantees the generation of transparent and syntactically correct sentences(phenotypes), but not necessarily semantically correct or feasible ones. Generating syntactically valid phenotypes with post-processing to filter out semantically invalid ones suffers from some issues, e.g. introduction of bias toward short phenotypes and loss in search efficiency. These issues become significant in legal application domains. We first demonstrate that applying Grammatical Evolution with a context free grammar to legal non-compliance detection problems is not a tenable solution. Then we demonstrate how the addition of context sensitivity improves both the search efficiency and achieves a greater diversity in the case of the `iBoB` problem regarding legal non-compliance.

ACM Reference Format:
Carl Im and Erik Hemberg. 2019. On the Use of Context Sensitive Grammars in Grammatical Evolution For Legal Non-Compliance Detection. In *Proceedings of the Genetic and Evolutionary Computation Conference 2019 (GECCO '19).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In legal compliance context and details matter. In this paper we propose an extension to the genotype-to-phenotype mapping method in Grammatical Evolution (GE) [1] that is well-suited for imposing additional contextual information. In GE, one is guaranteed to generate syntactically correct sentences(phenotypes), but not necessarily semantically correct ones.

One class of problems for which the traditional GE approach is not a tenable solution can be found in the Governance, Compliance, and Risk Management industry. For example, in the case of payment-transparency regulation for Korean pharmaceutical companies, there are over 150 regulatory limits. Most scenarios involve only a part of these requirements as in the case of a pharmaceutical employee giving a medical sample to a doctor. But there are also limits and the need to check for violation of these limits require additional information. Thus, applying a context free approach of creating random syntactically correct phenotypes and checking afterwards for their contextual validity becomes intractable:

1) the sheer number of terminals for the 1,000 non-terminals will demand extremely long genomic sequence.

2) because the conditional probability that a randomly chosen syntactically correct phenotype is also semantically correct tends to be exceedingly small in this case, the chance of a longer phenotype being semantically correct exponentially diminishes as a function of the length of the phenotype, thus introducing bias toward the shortest phenotypes(i.e., of transaction length close to 1).

3) compensating for this bias simply by means of hardware tends to be difficult and expensive computationally, because, the probability of generating phenotypes of a given transaction length decreases exponentially with the transaction length.

A large and complex grammar, expressed in Backus-Naur Form(BNF), increases the complexity of the code for semantic validation of sentences. Again the traditional approach has often been to allocate this part to the programmers as well. But this part should really be owned by lawyers(subject matter experts). Which means that the genotype-to-phenotype mapping should be refactored into two parts, *1)* the first part being a context independent genotype-to-abstract phenotype transcription, owned by programmers. *2)* the second part is owned by the lawyers that instantiates the abstract phenotypes to phenotypes that involve concrete terminals only.

The main difference between an application in the legal domain that is an academic exercise and an application that has real world relevance is the associated legal liability. It is in this context that the common phrase of "legal / non-legal collaboration" gains a technical meaning. The current GE architecture where the semantic validation or validation based on legal context is just another static class of validation methods does not help the lawyers, the rightful owners of this issue, to attest to an accurate and transparent phenotype validation. What this consideration suggests is that the genotype-to-phenotype mapping should be divided, with the second part being maintained and managed by the lawyers.

In previous work regarding legal applications the approach generating grammatically correct phenotypes with post-processing to filter out semantically invalid ones suffers from two issues [1], namely introduction of bias toward short phenotypes and significant loss in search efficiency. In this paper, we attempt to address these challenges and propose a potential solution. The research question we formulate from these issues is: *How can context sensitivity be efficiently introduced into Grammatical Evolution for legal non-compliance problems?*

In the `STEALTH` framework [1] each transaction needs to be analyzed for legality/feasibility before it can be executed. A simple check is to validate whether or not an entity has an asset before it

GECCO '19, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

can be transferred. Feasibility checks are divided into two broad categories namely *impossible transactions* and *economically unviable transactions*. The STEALTH framework simply ignored infeasible and invalid transactions during the evaluation, leading to a potential loss of search efficiency.

To alleviate this we introduce a context sensitive phenotype(*CSP*), that has two mapping steps for the genotype-to-phenotype. First, a standard GE mapping is performed. However, some predetermined terminal values are unbound until evaluation within the context(state). Once the terminal value needs to be evaluated the genome might be needed to determine a choice in the context sensitive transactions. This additional mapping of the genome is required since there can be non-empty sets of e.g. buyers and sellers for a transaction.

2 EXPERIMENTS

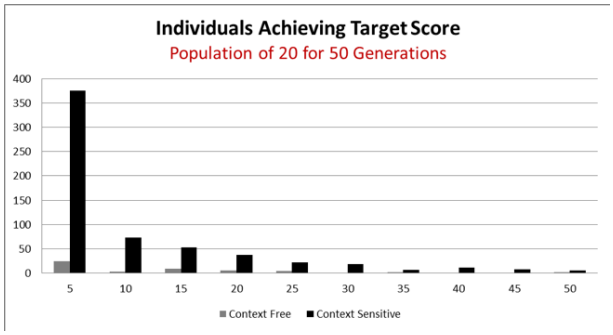
In these experiments we are concerned with the efficiency of a context sensitive representation for Grammatical Evolution in the legal non-compliance detection. We perform experiments with *CSP* and standard GE with a context free phenotype, which we call *CFP*, on a legal non-compliance problem called `iBoB` [1].

Figure 1 show the number of correct solutions found over the generations. The general trend shows that *CSP* finds better solutions earlier than *CFP* for all the variations of population size and generations. In addition, *CSP* finds more correct solutions than *CFP*, see Table 1. For the different settings of population size and generations we see that as the population size increases more individuals achieve the target score in the early generations. Moreover, when the population size is smaller *CSP* is capable of achieving the target score as the generations progress at a higher rate than *CFP*. This indicates the utility of *CSP* for finding high performing solutions. Note that with a large population size, *CSP* finds the target score in the first generation.

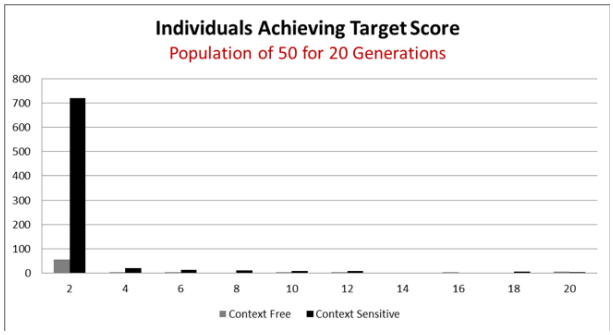
Table 1 summarizes some other statistics from the experiments and *CSP* is outperforming *CFP* on these measures as well. From Table 1 we can postulate that the high rate of success is partly attributable to the high ratio of semantically valid phenotypes, i.e. *CSP* manages to efficiently use the provided fitness evaluations to construct solutions. In addition, there is a slight difference in average length and the most number of transaction in the phenotypes are found by *CSP*. Finally, the number of unique solutions is greater for *CSP* due to its ability to find more valid and longer transactions.

An interpretation is that it is simply a method to reduce the size of the search space. Within the *CFP* BNF `Transaction`[E,E,A,A], there are $|E|^2\ |A|^2$ possible instances. However, once we impose legal constraints, it is reduced to `Transaction`[E,e(E,g),a(E),a(e(E,g))], where e(E,g) = `ListOfNonAffiliates`[g mod Number of Nonaffiliates], which is a deterministic function requiring just one element from the genome, and a(*) is the asset owned by the entity * at the time of transaction.

In the future we will try more complex problems and settings of the *CSP* method. A key component to investigate further is how to introduce degrees of context sensitivity. For example a parameter to control the level of context sensitivity.



(a) CSP vs CFP populations size 20, generations 50



(b) CSP vs CFP population size 50, generations 20

Figure 1: Number of `iBoB` solutions with the target score found (Y-axis) at a generation (X-axis) for the Context Free Phenotype(*CFP*) and Context Sensitive Phenotype(*CSP*) for varying population size and generation with total number of fitness evaluations fixed to 1,000.

Table 1: Results on `iBoB`. 1,000 simulations for *CFP* as well as for *CSP*, the population size and number of generations are indicated by (population size, generations). Ratio Valid is the average number of valid phenotypes as a percentage of total number of individuals generated.

Method	Ratio Valid	Average Transactions	Max Transactions
CFP (100, 10)	0.2	6.08	8
CFP (50, 20)	0.381	6.10	8
CFP (20, 50)	0.53	6.16	7
CFP (10, 100)	0.475	6.18	7
CSP (100, 10)	0.905	6.91	22
CSP (50, 20)	0.912	6.92	18
CSP (20, 50)	0.932	6.91	18
CSP (10, 100)	0.956	6.86	16

REFERENCES

[1] Erik Hemberg, Jacob Rosen, Geoff Warner, Sanith Wijesinghe, and Una-May O'Reilly. 2016. Detecting tax evasion: a co-evolutionary approach. *Artificial Intelligence and Law* 24, 2 (2016), 149–182.