Swarm and Evolutionary Computation 26 (2016) 1–7

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

journal homepage: www.elsevier.com/locate/swevo

Semantic genetic programming for fast and accurate data knowledge discovery

Mauro Castelli [a,*], Leonardo Vanneschi [a], Luca Manzoni [b], Aleš Popovič [c,a]

[a] NOVA IMS, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal
[b] DISCo, Universitá degli Studi di Milano Bicocca, 20126 Milano, Italy
[c] Faculty of Economics, University of Ljubljana, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 8 January 2015
Received in revised form
23 May 2015
Accepted 15 July 2015
Available online 26 July 2015

Keywords:
Genetic programming
Semantics
Knowledge discovery

ABSTRACT

Big data knowledge discovery emerged as an important factor contributing to advancements in society at large. Still, researchers continuously seek to advance existing methods and provide novel ones for analysing vast data sets to make sense of the data, extract useful information, and build knowledge to inform decision making. In the last few years, a very promising variant of genetic programming was proposed: geometric semantic genetic programming. Its difference with the standard version of genetic programming consists in the fact that it uses new genetic operators, called geometric semantic operators, that, acting directly on the semantics of the candidate solutions, induce by definition a unimodal error surface on any supervised learning problem, independently from the complexity and size of the underlying data set. This property should improve the evolvability of genetic programming in presence of big data and thus makes geometric semantic genetic programming an extremely promising method for mining vast amounts of data. Nevertheless, to the best of our knowledge, no contribution has appeared so far to employ this new technology to big data problems. This paper intends to fill this gap. For the first time, in fact, we show the effectiveness of geometric semantic genetic programming on several complex real-life problems, characterized by vast amounts of data, coming from several different application domains.

1. Introduction

Big data is an emerging research field that has gained considerable attention from both academia and practitioners. Recently, the quest for discovering knowledge from big data has proliferated to various corners of our society and attracted researchers from different disciplines, such as biology [1,2], health [3,4] and business [5], just to mention a few. Still, researchers continuously seek to advance existing methods and provide novel ones for analyzing rich data sets to make sense of the data, extract useful information, and build knowledge to inform decision making. We regularly witness the emergence of new forms of computation, combining statistical analysis, optimization, and artificial intelligence for constructing statistical models from large collections of data in an efficient way. Extant research from various fields has considered machine learning (ML) as a key technology in the analysis of big data [6]. One of the youngest paradigms of ML is Genetic Programming (GP) [7]. Despite its infancy, GP has already shown promising results and characteristics as the method of choice for analyzing and solving complex problems, including industrial ones [8,9]. Yet, with big data utilization opportunities constantly arising, the challenge for decision makers of choosing suitable methods that provide timely knowledge discovery and above average levels of understandability persists. Moreover, many algorithms do not scale beyond data sets of a certain size, whereas real-world situations often call for capabilities commonly handling much larger data sets. In a recent interview for the MIT press [10], Una-May O'Reilly, leader of the AnyScale Learning For All (ALFA) group of the MIT, presented GP as one of the most promising emerging technologies for addressing some of the most pressing open issues of big data, an idea that had already been fostered two years earlier by O'Reilly herself and her co-authors in [11]. In her interview, O'Reilly talks about the great versatility and ability to adapt to changes, that should create a competitive advantage for all the methods based on artificial evolution in managing vast amounts of data. O'Reilly's work encourages and motivates us to pursue the investigation of GP as a method to manage big data. As an integration of O'Reilly's considerations, we also assert that GP has several advantages and drawbacks compared to other machine learning (ML) methods.

The first important advantage of GP consists, in our opinion, in the fact that, contrary to several other approaches (among which

[*] Corresponding author. Tel.: +351 213828628; fax: +351 213872140.
E-mail addresses: mcastelli@novaims.unl.pt (M. Castelli), lvanneschi@novaims.unl.pt (L. Vanneschi), luca.manzoni@disco.unimib.it (L. Manzoni), apopovic@novaims.unl.pt (A. Popovič).

http://dx.doi.org/10.1016/j.swevo.2015.07.001
2210-6502/© 2015 Elsevier B.V. All rights reserved.

Neural Networks and Support Vector Machines, just to name two of the most popular) GP is generally able to produce models of data that are directly readable by humans. In this way, experts in the applicative domain should easily interpret and validate these models, eventually having the possibility of manually amending them in some parts, if needed. Secondly, GP is able to perform an implicit feature selection that happens at the same time as the process of learning of the data model. Thus, when using GP it is usually not necessary to preprocess data with any explicit feature selection algorithm. Once again, this is not the case for other ML methods (like for instance, again, Neural Networks and Support Vector Machines), where all the features entered in input to the system will be used by the produced model. This ability of GP of performing an automatic feature selection allows the final user to reduce the overhead of time of explicit feature selection algorithms, also alleviating from the choice of the most appropriate one, a choice that is well-known to be highly depending on the particular application we are trying to solve. Thirdly, as opposed to other ML methods, GP has a much bigger versatility, given by the fact that the learning process is guided by the fitness function, which is hand-tailored by the designer. Lastly, GP is widely recognized to outperform many of the existing ML methods, in particular for complex problems, where little or nothing is known about the model underlying data. This makes GP one of the most promising methods for facing the big data challenge.

Yet, we have to caution decision makers against the recognized drawbacks of GP compared to other ML methods. To begin with, GP is generally a slow and time and computational resource consuming process, and the fact that GP allows us to save the computational overhead given by explicit feature selection only partially solves this problem. The gap in the training time is particularly visible when GP is compared to methods like linear or least square regression, aimed at solving systems of equations, even though GP is often able to produce solutions of better quality than them on complex problems. Next, GP is non-deterministic; at each different execution, GP produces a different model. This fact differentiates GP from other methods, like linear or least square regression and Support Vector Machines, and historically did not help the diffusion of GP in many applicative domains, where the experts for a long time interpreted the non-determinism as a lack of reliability. Only recently, GP has been coupled with rigorous statistic methods, like the ones used in this work, which make them a strongly reliable computational method. Lastly, even though GP has strong theoretical foundations, being at least as solid as the ones from other ML methods, compared to several other ML methods these theoretical bases are generally more complex, involving concepts like Markov chains or other sophisticated statistical studies. As a consequence, it is also more difficult to convince a beginner about the reliability and conceptual rigor of GP.

Considering these drawbacks, we believe that the standard version of GP (STGP), originally defined by John Koza in [7] should be improved to become a widely recognized standard method for big data. Thus, in this paper we present a variant of GP, namely Geometric Semantic GP (GSGP), which has been recently introduced [12] and has the very interesting property of inducing a unimodal error surface for any supervised learning problem. In other words, it is possible to prove that, when using GSGP, the error function does not have any local minimum. This fact should help GSGP to speed up the trial-and-error process of search for good quality solutions, that is typical of many Computational Intelligence methods, including standard GP, and that very often stagnates in local optima. Previous work (see for instance [13,14]) has already highlighted the promising characteristics of GSGP on an interesting set of real-life applications, encouraging us to pursue the work by applying GSGP to problems characterized by larger sets of data. The existence of an efficient implementation of GSGP, like the one discussed in [15], is a further support that pushes us to test the power of GSGP on big data. GSGP is able to produce models that are very precise, often outperforming standard GP (STGP), but also rather “big” in terms of code size. For this reason, even though the implementation proposed in [15] produces a compact representation and allows us to understand some characteristics of the final model (like for instance the features it uses), the firstly emphasized advantage of GP is strongly limited by GSGP. On the other hand GSGP maintains the remaining advantages of GP, and even incrementing the last one, as demonstrated also by the experimental results presented later in this paper. Concerning the drawbacks of GP, the first drawback is strongly reduced by GSGP. In fact, GSGP is generally much faster than STGP. At the same time, while the second limitation still exists (even though it is possible to show that GSGP generally presents a smaller variance of the generated results compared to STGP), the lastly highlighted disadvantage is mitigated, given that the theoretical foundations of GSGP are possibly more intuitive and thus simple to understand than the ones of STGP.

To validate the proposed system, we use a set of complex problems, many of them characterized by vast amount of data. One of the problems we used as test cases come from the field of Biology. Specifically, high-throughput microarrays have become one of the most important tools in functional genomics studies and they are commonly used to address various biological questions like disease classification and treatment prognosis. One of the most important consequences of the massive use of microarray technology has been the huge proliferation of data, which makes the problems and analysis of big data strategies of paramount importance for Biology [16]. The literature is rich of studies involving GP for addressing complex problems, both from the Biology field and others. For instance, in [17] a new classifier that uses GP, explicitly designed to improve generalization ability in presence of vast amounts of data was proposed. Its performance was tested on a rich set of test cases, several of which real-life applications from Biology. In [18] classifiers systems based on GP were instead applied to the field of financial forecasting, another domain that is usually characterized by vast amounts of data, often disordered and unstructured. The other two problems are taken from the e-commerce industry. In particular, we focus on predicting the average reviews score of products for two categories (kindle store and Amazon music) of the http://amazon.com webstore, considering all the reviews up to 2013. Since that webstore is one of the largest in its sector, the chosen problems are particularly suitable to assess the performance of GSGP over vast datasets.

The remainder of this paper is organized as follows. We firstly present the geometric semantic operators used by GSGP. More specifically, we highlight the benefits of the operators on the search process and their suitability for handling big data (Section 2). We then outline the experimental setting of the case study problems under analysis. This is followed by our findings on how semantic genetic programming aids big data knowledge discovery (Section 3). Finally, we conclude the paper by discussing the main contributions and implications of our work and by exploring avenues for future research (Section 4).

2. Geometric semantic genetic programming

Despite numerous human-competitive results achieved by GP [19], researchers still continue to investigate new methods to improve the power of GP as a problem solving techniques [20]. In recent years, one of the emerging ideas is to integrate semantic awareness in the evolutionary process (a survey of methods to integrate semantic awareness in GP can be found in [13]). Even