Swarm and Evolutionary Computation 39 (2017) 323–338

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

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

# A semi-supervised Genetic Programming method for dealing with noisy labels and hidden overfitting

Sara Silva[a,b,*], Leonardo Vanneschi[c], Ana I.R. Cabral[d], Maria J. Vasconcelos[e]

[a] *BioISI – BioSystems & Integrative Sciences Institute, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal*
[b] *CISUC, Department of Informatics Engineering, University of Coimbra, Portugal*
[c] *NOVA IMS, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal*
[d] *Department of Natural Resources, Environment and Territory, Instituto Superior de Agronomia, University of Lisbon, Tapada da Ajuda, 1349-017 Lisbon, Portugal*
[e] *Centro de Estudos Florestais, Instituto Superior de Agronomia, Tapada da Ajuda, 1349-017 Lisboa, Portugal*

ARTICLE INFO

*Keywords:*
Data errors
Noisy labels
Classification
Hidden overfitting
Semi-supervised learning
Genetic Programming

ABSTRACT

Data gathered in the real world normally contains noise, either stemming from inaccurate experimental measurements or introduced by human errors. Our work deals with classification data where the attribute values were accurately measured, but the categories may have been mislabeled by the human in several sample points, resulting in unreliable training data. Genetic Programming (GP) compares favorably with the Classification and Regression Trees (CART) method, but it is still highly affected by these errors. Despite consistently achieving high accuracy in both training and test sets, many classification errors are found in a later validation phase, revealing a previously hidden overfitting to the erroneous data. Furthermore, the evolved models frequently output raw values that are far from the expected range. To improve the behavior of the evolved models, we extend the original training set with additional sample points where the class label is unknown, and devise a simple way for GP to use this additional information and learn in a semi-supervised manner. The results are surprisingly good. In the presence of the exact same mislabeling errors, the additional unlabeled data allowed GP to evolve models that achieved high accuracy also in the validation phase. This is a brand new approach to semi-supervised learning that opens an array of possibilities for making the most of the abundance of unlabeled data available today, in a simple and inexpensive way.

## 1. Introduction

This article tells a story. This story takes place in the realm of satellite imagery. It is a story of classification methods yielding unusually bad results, the search for the causes of such odd behavior, the discovery of human errors in the labeling of the data, and finally, the development of a method to overcome them. Why not simply eliminating the errors and redoing the work, in order to achieve the typical good results on this kind of application? Because noisy labels are very common [1–4] and usually go unnoticed, as the results seldom reveal, or even suggest, that something is wrong with the data. And even when they do, it may not be viable to go back and clean the data, and repeat the whole process. So we have to assume the data contains errors, and we have to develop learning methods that can still provide useful and reliable models under these conditions. One can state that Genetic Programming (GP) [5,6] is one of the most resilient learning methods, able to cope with noisy and faulty data, and still provide good results. But as the story will tell, even GP can be highly deceived by a very small percentage of data mislabeling.

The next section is dedicated to reviewing previous and related work on the subjects of data errors and semi-supervised learning. Section 3 describes the problem tackled and the data used in this study, including a description of the errors. Section 4 describes the workings and parameterizations of the two methods used in the beginning of our work, Classification and Regression Trees, and Genetic Programming, while Section 5 specifies the procedures used to assess their performance. Section 6 introduces the new semi-supervised GP method, explaining the differences to standard GP, and Section 7 reports all the results obtained with all the methods. Section 8 discusses these results at length, exploring the reasons for the success of the semi-supervised GP method. Finally, Section 9 summarizes the contributions of this work, and raises many additional related questions.

[*] Corresponding author. BioISI – BioSystems & Integrative Sciences Institute, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal.
*E-mail address:* sara@fc.ul.pt (S. Silva).

https://doi.org/10.1016/j.swevo.2017.11.003
Received 6 April 2017; Received in revised form 29 September 2017; Accepted 12 November 2017
Available online 15 November 2017

## 2. Previous and related work

This section reviews the literature related to both themes addressed by this work: data errors and semi-supervised learning. Inside each theme we begin by addressing work published in the context of the wide machine learning field, followed by work in the context of Evolutionary Algorithms (EAs) and more specifically GP, and finally work related to remote sensing. We do not attempt at performing an exhaustive review of all the work published on such wide research themes, but instead we overview the amount and type of work that has been done in different specific themes, in particular the ones more related to our own work, providing pointers to more thorough surveys whenever possible.

### 2.1. Data errors

The objective of many learning systems is to construct a model of the world which is completely consistent with observations, based on the assumption that the data available is error-free [7]. However, this is seldom the case. According to [7] the many sources of errors may be external or internal. External errors are objective, like random errors (normally called noise) and systematic errors. Random errors are introduced by the inherent unpredictability of the world being observed, or during the transmission of the observations to the learning system. Systematic errors are more predictable, arising for instance from a problem in the device collecting data, like an instrument that is poorly calibrated. Internal errors are subjective and depend mostly on the interpretation of the data. Transversal to this classification is the concept of outlier, *i.e.*, an observation that appears to deviate markedly from other observations in a sample. The importance of outliers in statistical data and machine learning can be inferred by the very large amount of literature dealing with the subject. Interesting surveys can be found in [8] and [9].

Strategies for learning with imperfect data can focus on data cleansing, *i.e.*, identifying and repairing the errors, or on developing and using learning systems that are able to cope with them. Data mining with noisy data is considered in [10], where the authors survey other related works and propose their own error-aware method based on using noise knowledge to rectify the model built from corrupted data. According to this work, data cleansing is a limited procedure that can only be applied to certain error types from certain data sources, may lead to information loss, and constitutes in itself a potential source of additional errors.

Nevertheless, data cleansing has played a critical role in ensuring data quality, particularly with the advent of big data, where errors in data are extremely frequent. Many data cleansing algorithms have been translated into tools to detect and to possibly repair certain classes of errors such as outliers, duplicates, missing values, and violations of integrity constraints [11]. In [12], various views of data cleansing were surveyed and reviewed and a brief overview of existing data cleansing tools was given. A general framework of the data cleansing process was presented, as well as a set of general methods that can be used to address the problem. Other works followed the same path, like [13,14]. Some methods were specifically developed for big data, like [15,16]. Since different types of errors may coexist in the same data set, it is often appropriate to run more than one kind of tool. In [11], a systematic analysis of the existing data cleansing tools was performed, aimed at understanding whether these tools are robust enough to capture most errors in real-world data sets and what is the best strategy to run multiple tools to optimize the error detection effort.

Oblivious to all the efforts in cleaning data, and the problems that erroneous data may cause to learning systems, many machine learning methods are in fact equipped to perform reasonably well in modeling data with inaccuracies, as they rely on soft computing techniques to produce inexact but robust solutions. Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and Genetic Programming (GP) are some of them. In classification problems, these methods can deal not only with errors in the features, but also with errors in the labels, precisely the type addressed in our work.

An excellent review of different types of label noise and their consequences, as well as different algorithms that consider label noise, was published some years ago [17]. Among the large body of work that is reviewed, semi-supervised learning appears as one of the main noise-tolerant approaches, and a number of works on remote sensing are among the target applications. Other work not covered in this review deals with noisy labels in image annotation [18], data factorization [19], labelling pixels in aereal images [20], multiple kernel learning [21] and sentiment detection in Twitter [22].

In [23] a theoretical study on risk minimization bounds is performed on the problem of binary classification in the presence of erroneous labels, and the results are applied in developing noise-tolerant versions of SVM and weighted logistic regression. Other applied theoretical works are presented in [24,25], where the authors develop and analyse an improved logistic regression classifier that is robust to label noise. More recently, [26] studies the conditions in which a consistent classification is possible with label noise, [27] studies the use of importance reweighting to achieve an optimal classifier in the presence of noisy labels, and [28] shows that loss factorization can be directly applied on learning with poorly labeled data.

Among the most recent work, a few studies deal with the identification and correction of noisy labels. In [29] a novel $L_1$-optimization based sparse learning model is used to explicitly detect noisy labels, while [30] does it via a mutual consistency check using a Parzen window classifier. In [31] the unreliable labels are improved using a text label refinement algorithm, while in [32] the noisy labels are recovered as the classifier is built, using a Least-Squares SVM. In [33], the approach of repeated labeling is used in order to improve label quality, including a selective approach based on both labeling and model uncertainty.

A large and diverse body of work has also been published in the past few years focusing on using noisy labels in such varied applications as the detection of malicious network traffic [34], classification of historical notary acts [35], and time-series segmentation [36].

Noisy labels are also tackled with deep learning approaches. The notion of consistency is used in [37] to improve the predictions of a deep ANN when the labeling is missing or is subjective. Deep learning is also used in other works like [38–40]. A number of approaches rely on active learning techniques [41–44].

Compared to the huge effort that was dedicated to the detection and repairing of data errors by the larger machine learning community, the amount of work involving EAs, in particular GP, for these tasks is rather limited. Indeed, to the best of our knowledge, no paper specifically dealing with GP has ever tackled these issues directly. On the other hand, it is quite a common trend to use GP as a feature extraction process and, among the several advantages of this approach, it is typical to show that GP is resistent to data errors, and is often able to generate features that are more robust, more insightful and less prone to errors than the ones contained in the original data. The quality of a set of features can be quantified by using a machine learning method to generate a data model based on those features (and thus the fitness of the evolved features is given by the performance of this method), or by using other criteria that do not depend on any machine learning method. For instance, in [45] a measure based on information gain was employed as fitness function.

Another trend is to incorporate techniques into GP that improve its generalization ability. This was done in [46], where symbolic regression problems were solved by using new measures of fitness based on statistical learning theory, like for instance Akaike Information Criterium, Bayesian Information Criterium and Structural Risk Minimization, based on the Vapnik-Chervonenkis (VC) theory. The authors show the advantages of this type of approach and a better ability of GP to deal with noisy data.