Evolutionary Machine Learning: An Essay on Experimental Design

Filipe Assunção
fga@dei.uc.pt
Nuno Lourenço
naml@dei.uc.pt
Bernardete Ribeiro
bribeiro@dei.uc.pt
Penousal Machado
machado@dei.uc.pt

CISUC
Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal

Abstract

Evolutionary Machine Learning (EML) combines Evolutionary Computation (EC) with Machine Learning (ML) to automatically search for the best structure and/or parameterisation of ML models for solving specific tasks. However the results reported by the authors in their articles detail their work, replicating the results and comparing them to other approaches are tasks that tend to be difficult. This happens mainly because of the high number of numeric parameters, and specific technical details. Another issue that prevents the approaches from being replicated is the fact that the code developed is rarely made publicly available. In this essay we discuss and provide some guidelines to address these problems. Our goal is not to provide a unique, right answer, for these issues. Rather, we aim to promote a healthy discussion that can lead to new and innovative ideas and practices.

1 Introduction

When developing Machine Learning (ML) models for performing specific tasks (e.g., an Artificial Neural Network (ANN) for classification) the practitioner often undergoes a long and weary process of trial-and-error, where the structure and/or parameters of the model are continuously tuned in the search for the best performance. To avoid this, practitioners can resort to Evolutionary Machine Learning (EML), which uses Evolutionary Computation (EC) to optimise the ML models. Therefore, a population of individuals (each encoding a solution for the ML model) is continuously evolved, guided by a quality function that defines how well each model performs on solving the task.

However, according to Eiben and Jelasity [2] "verifying results found in the literature is in practice almost impossible". This happens for a number of reasons: the proposed approaches have specific and complex implementation details, parameters are not always clearly explained and detailed, and the code developed is not made publicly available. In addition, the evaluation of the methodologies is not performed in the same way, which makes the comparison between approaches even harder.

The critique by Eiben and Jelasity focuses on EC. But, similarly to what happens in EC, in ML models also have lots of parameters that need to be defined. Thus, the combination of both fields results in a bigger problem, since the number of parameters that must be clearly defined increases, making the reproduction of the results more challenging.

The main goal of this essay is not to propose one unique and right way of specifying the experimental setup. Instead, we want to promote discussion so that better practices can emerge and be used by authors in the area. In the upcoming sections we start by tackling reproducibility, and the comparison of approaches, respectively on Sections 2 and 3. Then, in Section 4, we focus on statistical analysis. To end, in Section 5, we discuss open questions and future directions.

2 Reproducibility

Reproducing the results obtained by EML methods is a difficult task. Since we are are combining two fields, namely EC and ML, there will be more parameters and technicalities than if considering only one in isolation. Therefore, if all the parameters that are used in the algorithms are not clearly pointed out, the reproduction of results and comparison of approaches is very hard.

One of the most common problems in EML is trying to understand how the benchmark dataset was partitioned. We advocate the use of three partitions: (i) train; (ii) validation; and (iii) test. The first two are used during the evolutionary process, and the last (test) must be kept out of it, and is used for measuring how well the evolved models perform beyond the data used for generating them, i.e., their generalisation ability. During the evolutionary process, the training set is used if the evolved ML model has a training phase (usual in supervised learning) for optimising numeric values (e.g., weights of ANNs), and the validation set is used for measuring the fitness value of the model, after training. However, if there is no need for a training phase, the train and validation sets can be merged, and consequently we only need two dataset partitions: validation and test.

We understand that different authors may use different names to mention the same sets that we defined above. The important message is that, the explanation of the dataset partitions, and what they are used for needs to be clear. More importantly, a percentage of the data must be kept outside the evolutionary process; otherwise the results are biased. Moreover, the way the partitions were created should be clear and reproducible.

Still focusing on the benchmark issues, whenever data augmentation, dimensionality reduction, or sampling techniques are used, both the method and its parameterisation must be provided. This applies for all other EC and ML parameters. Therefore we recommend that the experimental setup is reported on a table, in a clear and summarised way, avoiding the need for the reader to scan the entire article to search them. We divide the table into three distinct sections:

EC – specifies all the parameters directly concerned with the used evolutionary engine, such as, number of runs, number of generations, population size, crossover and mutation rates and parent selection mechanism;

ML – details the parameters of the evolved models, and the allowed ranges. Other information, such as the metric used for assessing the quality of the models, or the use of cross-validation can be also included here;

Benchmark – contains all the information regarding the benchmark partition, and when applicable, dataset sampling, augmentation and other parameters regarding any form of pre-processing.

More table sections may be required depending on the problem that is being solved. The same applies to the parameters contained in each section.

Despite a clear specification of the used parameters, the replication of results may still be difficult due to implementation and algorithmic details. Eiben and Jelasity [2] suggest that to avoid the previous a framework could be implemented, and used by all researchers, where only new features would need to be implemented. Although this standardisation of code is likely the ideal approach, it may not be feasible. Researchers use different programming languages, and there are many approaches to encode the same solutions. So, unless all variants are implemented and made available, researchers will tend to avoid standardisation. We defend that the easiest and most simple form of reproducibility would be to open up and share the code developed, by uploading it to repositories and include it within the paper.

3 Comparing Approaches

After specifying the experimental setup and conducting experiments there is the need to compare the approach with similar ones to acknowledge how it performs in the broader scope.

The first decision that needs to be made is concerned with the number of evolutionary runs that will be executed. Evolutionary Algorithmss (EAs) are stochastic search heuristics, and thus different runs can lead to very different solutions. Further discussion on the definition of the number of evolutionary runs is carried out in the next section.

ML results are typically presented in the form of tables, which report various performance metrics on the used models. In addition, in specific domains, e.g. ANNs, specific criteria concerning the structure of the models also tend to be presented: number of neurons, layers and connections. Whatever the chosen metrics are, what is important is to clearly define

them, because those are the properties that are going to be used to compare one methodology with others. Further, we defend that a plot depicting the fitness evolution across generations should be presented, because a table does not allow one to check evolution and convergence speed, which may be important in problems where time is crucial.

Some authors just report the results attained by the best model found throughout evolution. However, as above stated, EML approaches have a stochastic behaviour. Therefore, showing just the best result does not capture the overall picture of the tested methodologies, and in some circumstances the best solutions can be deceiving outliers. By presenting the average of the best individuals found in each evolutionary run, along with its standard deviation it becomes possible to verify if the methodology consistently finds suitable solutions or not. The best result can also be presented separately, but never at the cost of discarding average ones.

Even though results can be shown in terms of averages, they provide little information. It is possible to have an approach A that, on average, is slightly superior to an approach B, but nonetheless the difference is insignificant, and thus the approaches can be considered equal in terms of performance. To effectively acknowledge the superiority of one method over another, based on empirical data, we must use statistical tests.

4 Statistical Analysis

When using EAs it is unlikely that two consecutive runs lead to the same results. Randomness is an important part of the evolutionary process, and thus the stochastic essence of the methods requires multiple repetitions of the experiments in order to gather enough experimental data to apply a sound statistical analysis. Next, we discuss how we think that the experimental analysis should be conducted, considering aspects like the number of runs, initialisation of the populations, and the statistical tests that should be used.

Before starting any experimental study one should define the number of runs, N, the hypothesis, H, that is to be tested, and the significance level, α. The number of runs identifies the amount of executions of the algorithm. Different runs can generate distinct results; if for different runs the algorithm consistently gives similar results it is possible to state that it is robust. Thus, the larger the value of N the better we can assess the robustness of the approach. Additionally, it also defines the size of the sample that will be used by the statistical methods. Therefore, we need a large value of N, which as a rule of thumb should not be lower than 30. Next, the hypothesis H, which is a statement that we want to assert as true, must be defined. The last step consists on the definition of the significance level α. The significance level is used in the statistical tests as the cutoff value to reject the null hypothesis. Commonly used values for α are either the 0.05 or the 0.01. The lower the significance level, the more the data must diverge from the null hypothesis to be significant.

Once we have executed the methods, and gathered all the samples, we need to ascertain what distribution our data follows in order to decide which type of statistical test to use. Some are based on the assumption that the data follows a certain distribution. When this happens it is possible to use a set of statistical procedures called parametric tests. The most common assumption is that the distribution of the samples follows a normal distribution. To check if the samples follow a normal distribution it is possible to use two tests: Kolmogorov-Smirnov and Shapiro-Wilk. If the test is non-significant, it tells us that the distribution of the samples is not different from a normal, and thus the we can assume that probably the data is normal. After this, and before selecting any parametric test, we need to check if the variance is homogeneous.

The last part of the statistical analysis is concerned with the hypothesis testing and the reporting of the results. To test the hypothesis we can use two types of statistical procedures: (i) parametric, which assumes that the sample data comes from a population that follows a probability distribution based on a fixed set of parameters, e.g., a normal distribution; (ii) non-parametric, that makes no assumptions about the data distribution. There are several tests available in each category. We need to check which is the one that suits our assumptions the best. For example, if we are to compare two approaches A and B, with different initial conditions and with no assumptions about the distribution of the samples, we have to select a non-parametric test (most common situation in EML). Based on these assumptions the test that is most appropriated is the Mann-Whitney U test. For other scenarios and statistical procedures, please refer to [3].

Care must be taken with the interpretation of the word "significant", because even if the probability of the effect in our results occurs by chance is small (less than α), it does not imply that the effect is of great importance. Insignificant and unimportant effects can be significant due to the large number of experiments conducted. So the question now is knowing how important an effect is. The solution to this problem is to measure and report the size of the effect that we are testing, known as effect size.

Table 1: Graphical overview of the statistical results with effect sizes.

		Dataset-A	Dataset-B	Dataset-C	Dataset-D
RMSE	Test	++	$\sim$	+++	++
	Validation	$\sim$	$\sim$	+++	$\sim$
Accuracy	Test	++	$\sim$	+++	++
	Validation	++	$\sim$	++	$\sim$
AUROC	Test	++	$\sim$	+++	++
	Validation	++	$\sim$	+++	$\sim$
F-measure	Test	+++	$\sim$	+++	$\sim$
	Validation	++	$\sim$	+++	$\sim$

The effect size is a simple and standardised measure of the magnitude of the observed effect. Since it is a standardised measure it means that we can compare effects sizes across different studies that have different metrics. In the literature there are many methods to compute the effect size; the Pearson's r correlation coefficient is one of the most widely used ones. One of the advantages is that it is constrained between 0 (no effect) and 1 (a perfect effect). Cohen [1] has made the following suggestions for a scale of the effects: small ($0.1 \leq r < 0.3$), medium ($0.3 \leq r < 0.5$), and large ($r \geq 0.5$). We recommend reporting the effect size in the form of a table, where the approaches are compared according to the following graphical overview: $\sim$ indicates no statistical difference between the compared approaches, and $+$ signals that approach A is statistically better than approach B. The effect size is denoted by the number of $+$ signals, where $+$, $++$ and $+++$ correspond respectively to small, medium and large effect sizes. A $-$ signals scenarios where approach A is worse than approach B. Table 1 shows an example following these guidelines.

5 Road Ahead

In this paper we have discussed multiple issues in experimental design, and how they affect EML. In particular:

- The reproduction of experimental results is nearly impossible. As EML merges the EC and ML fields, the number of parameters that needs to be set up is extremely high. The same happens to implementation details, which if not clearly specified, and if the code is not shared, make it almost impossible for other authors to replicate the obtained results. In that sense we propose the creation of a platform for sharing the benchmarks and obtained results, where the code may be made available;
- The reporting of results cannot be based only on the presentation of the best models. At least average values should be provided, along with the standard deviation, so that it is possible to analyse the consistency of the evolved models over the different runs;
- The comparison of different approaches using statistical procedures is essential. Most of the current published works do not rely on any statistical inference tools, or, when they do, the report of the analysis is not adequate, and very difficult to follow. In this work, we propose a recipe to fill this void by defining a set of guidelines that aim at improving and easing the comparisons between different works.

The current essay is by no means an extensive review of the literature. There are much more open questions that need to be answered. One of the most prominent ones concerns the increasingly necessity to find objective measures that are able to define what makes a good model. It is true that there are many performance metrics (e.g., accuracy or RMSE); however, when the results of two different methods are very close, should one choose a model that is more complex, despite the small increase in performance?

Acknowledgements

This work is partially funded by: Fundação para a Ciência e Tecnologia (FCT), Portugal, under the grant SFRH/BD/114865/2016.

References

[1] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 1988. ISBN 0805802835.

[2] Agoston E Eiben and Márk Jelasity. A critical note on experimental research methodology in ec. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, volume 1, pages 582–587. IEEE, 2002.

[3] Andy Field. *Discovering statistics using SPSS*. Sage publications, 2009.