POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Machine Learning for Species Diversification Dynamics

Laurea Magistrale in High Performance Computing Engineering - Ingegneria del Calcolo ad Alte Prestazioni

Author: Mattia Colbertaldo

Advisor: Prof. Cristina Silvano

Co-advisor: Prof. Ernst Wit, Dr. Francisco Richter Mendoza (Università della Svizzera Italiana)

Academic year: 2023-2024

---

1. Abstract

Abstract

This study explores the application of deep learning techniques for inferring parameters of diversification models from extant phylogenies, providing an alternative to the slow and computationally expensive Maximum Likelihood Estimation (MLE) methods, which often lack analytical solutions. The research evaluates Convolutional Neural Networks (CNNs), Linear models, Feed Forward Neural Networks (FFNNs), Gated Recurrent Units (GRUs), and Long Short-Term Memory (LSTM) Networks. Despite computational constraints, promising results were achieved. The analysis focused on two types of encoding: CDV and Summary Statistics. A significant contribution of this work is the introduction of a Random Forest Classifier for parameter inference, which exhibited superior predictive accuracy and faster training times compared to other deep learning models. The study also highlighted the limitations of previous research due to overly stringent constraints. A novel tool capable of inferring the mathematical model of a given phylogeny was developed, dramatically reducing computational time from minutes to milliseconds and achieving an impressive test accuracy of 88%. This advancement enables efficient and accurate inference of diversification models, even in scenarios where MLE was infeasible.

The research leverages the diversitree package for simulating phylogenetic trees under various models, selecting it for its comprehensive and customizable functions. This study represents a significant step towards developing sophisticated and scalable analytical tools for evolutionary biology.

The necessary files to reproduce all the following steps can be found at `https://github.com/Mattia-Colbertaldo/PhyloML`.

This thesis has been developed at the end of the Double Master Degree in High Performance Computing Engineering at Politecnico di Milano and Università della Svizzera Italiana as part of the EUMaster4HPC European Program.

2. Models

Diversification models are analytical frameworks used in evolutionary biology to study the

processes that drive the emergence and extinction of species over time. These models help scientists understand how different factors, such as physical traits, behaviors, and environmental changes, influence the rates of speciation (the formation of new species) and extinction (the loss of species). By analyzing phylogenetic trees—diagrams that show the evolutionary relationships among species—diversification models can reveal patterns and trends in biodiversity. Examples of diversification models include the CRBD (Cladogenetic Birth-Death), BiSSE (Binary State Speciation and Extinction), and MuSSE (Multiple State Speciation and Extinction) models. These models are essential for reconstructing evolutionary histories, predicting future biodiversity changes, and informing conservation strategies by identifying key traits and environmental factors that impact species survival and diversification.

For instance, woodpeckers with diverse drumming behaviors (Fig. 1) exemplify a Multiple State Speciation and Extinction model (MuSSE). We can split the phylogeny in three states: regular drummers, occasional drummers and non-drummers. Their drumming intricacy and frequency correlate with speciation and extinction probabilities. Complex drumming aids territory defense and mate attraction, reducing gene flow and fostering speciation. Conversely, species with simpler drumming might face higher extinction risks, particularly in densely populated areas. This interplay showcases how macro behavioral or physical traits, called states, impact evolutionary dynamics, underscoring the significance of diversification models in understanding biodiversity trends.

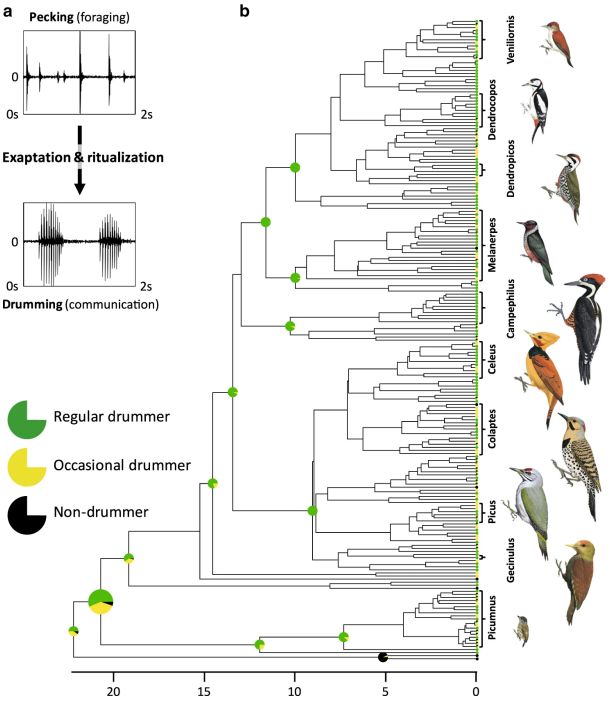

Figure 1: Drumming, among woodpeckers species, can be seen as a state differentiation, as it can lead to different speciation and extinction rates. From the figure, it can be seen that non-drummers have a lower speciation and an higer extinction rates, compared to drummers.

3. Simulations

Gillespie's Stochastic Simulation Algorithm (SSA) is utilized across all models for simulating birth-death processes, with or without character state transitions, as well as incorporating geographical considerations and quantitative traits. Gillespie's algorithm enables the sampling of waiting times until the next event (speciation, extinction, or transition) from exponential distributions. It calculates the total rate of events and selects the actual event

Model	Parameters	Description
const BD	λ, μ	Models pure birth-death processes without character state transitions.
BiSSE	$\lambda_1, \lambda_2, \mu_1, \mu_2, q_{12}, q_{21}$	Constant rates of speciation, extinction, and transition between two character states.
MuSSE	$\lambda_1, \ldots, \lambda_K, \mu_1, \ldots, \mu_K, q_{ij}$	Generalizing BiSSE to K character states with constant rates.
QuaSSE	Arbitrary functions of trait values	Allows for variation of speciation and extinction rates as functions of trait values.
GeoSSE	$s_A, s_B, s_{AB}, x_A, x_B, d_A, d_B$	Accounts for geographical considerations with constant rates of speciation, extinction, and dispersal.
ClaSSE	λ_{klm} (where $1 \leq k, l, m \leq n$), μ_k (where $1 \leq k \leq n$), q_{lm} (where $1 \leq l \neq m \leq n$)	Distinguishes between anagenetic and cladogenetic speciation with variable rates. There will be n^3 cladogenesis rates and n extinction rates plus $n(n-1)$ transition rates, making a total of $n^3 + n + n(n-1)$ parameters.
BiSSEness	$\lambda_0, \lambda_1, \mu_0, \mu_1, q_{12}, q_{21}, p_{0c}, p_{0a}, p_{1c}, p_{1a}$	Estimates rates of change occurring along lineages and probabilities of change during speciation.

Table 1: Overview of the main phylogenetic models.

based on its relative proportion to the total rate. Additionally, each model outlines specific adaptations and extensions of Gillespie's algorithm to handle various aspects such as character state transitions, geographical factors, and arbitrary functions of quantitative traits. An all-in-one tool was created to simulate phylogentic trees for any model and parameter range. The ranges used in this work were obtained by analyzing real-world phylogenetic trees and cleaning the data, in order to create a neural network suited to work in a realistic scenario.

4. Extending Deep Learning Techniques to Complex Phylogenetic Models

This paper[2] introduces a novel approach to diversification inference, based on deep learning. The method is applied to the time constant homogeneous birth-death model and the Binary-State Speciation and Extinction model. The authors show the results obtained with the deep learning models trained on the CDV encoding and the SS encoding, and with the MLE method: an error boxplot is shown in Fig. **??** for BD model.

The goal was to train various deep learning models on the more complex previously discussed phylogenetic models to accurately infer the parameters of the trees.

Additionally, I introduced a Random Forest Classifier to the models, that led to the best results, as shown in the following sections.

4.1. Addressing Ill-Posedness of the Problem

In this section, we address the challenge of ill-posedness inherent in the parameter estimation problem. Specifically, we encounter a scenario where multiple combinations of the speciation rates, denoted as lambda (λ), extinction rates, denoted as mu (μ), and transition rates, denoted as (q), can produce the same tree topology. This ambiguity impedes the model's ability to effectively learn the relationship between the trees and the parameters.

To mitigate this issue, we have implemented a reordering strategy for the parameters associated with each tree. The reordering is based on the lambda values, ensuring that the parameters

are consistently arranged in increasing order of lambda. By imposing this ordering constraint, we provide the model with a more structured input, facilitating its learning process.
Through this approach, we aim to enhance the model's capacity to discern meaningful patterns and relationships between the tree topologies and the corresponding parameters. By organizing the parameters in a systematic manner, we strive to improve the robustness and accuracy of the parameter estimation process in the face of the inherent ill-posedness of the problem.

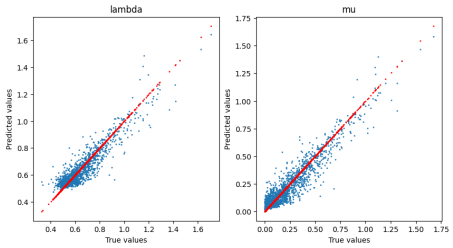

Figure 2: Results obtained with the Random Forest Classifier for the BD model. The red scatter points represent the true values, while the blue points represent the predicted values. The points are really close to the line $y = x$, which means that the predictions are accurate.

4.2. Results

The results obtained with the deep learning models (FFNN, CNN, and Linear) were promising, demonstrating these models' ability to accurately predict the parameters of the phylogenetic models. The Random Forest Classifier, in particular, exhibited exceptionally accurate results and proved to be the most effective approach for parameter estimation, with the added advantage of faster training times.
Predicting the individual states of parameters for the MuSSE model yielded results that were similar or worse compared to predicting all parameters simultaneously. This suggests that the parameters are not independent, and failing to consider their interrelationships can lead to suboptimal predictions. Recurrent Neural Networks models (GRU and LSTM), on the other hand, were unable to learn any meaningful relationships between the trees and the parameters.
A single Dense Layer model consistently showed slightly better performance in every scenario, closely rivaling the Random Forest Classifier. Given its faster training time compared to other models, including the Random Forest Classifier, the Dense Layer model emerges as a viable choice for this type of problem. The encoding method (CDV vs. SS) did not significantly impact the results, with both encodings yielding comparable performance across the models.
The distribution of parameters generated by the simulations was not uniform, leading to biased predictions towards the mean of the parameters. Addressing this issue could be an area for further work, especially if more data become available. The models were trained on a relatively small dataset due to limited computational resources, and the results could likely be improved with training on a larger dataset.
A notable limitation is that tree structures do not contain information about the character states of the species, so many more trees are needed to infer state-related parameters. Consequently, the models cannot accurately predict transition rates or the probabilities of character state shifts during speciation or lineage development. This issue is particularly relevant for models dependent on character states, such as BiSSEness and ClaSSE. The results obtained with these models, and for transition rates in general, reflect this limitation. Future work could explore alternative encoding methods that incorporate character states, enhancing the accuracy of predictions for transition rates and probabilities of character state shifts during speciation or lineage development.
In conclusion, the Random Forest Classifier was the best-performing model, demonstrating good accuracy in predicting parameters compared to MLE results provided by the authors, with remarkably short training times.

5. Parameter constraints and data manipulation

The paper "A Comparison of Deep Learning Architectures for Inferring Parameters of Diversification Models from Extant Phylogenies" [1] explores the use of deep learning for phylogenetic inference, with a focus on understanding the ideal network architecture and data representation for using deep learning in phylogenetic inference.

5.1. Data manipulation

Upon further examination of the code provided by the authors, It has come to our attention that the true values were incorporated into the training set, thereby inflating the results artificially. This practice is unequivocally prohibited as it fundamentally undermines the integrity of the evaluation process. Consequently, the results obtained under these conditions are rendered entirely invalid and without merit.

5.2. Strict Constraints

The constraints imposed by the authors on the BD and BiSSE model make the model non-fittable to real-world applications: the speciation rate of state 1 is twice the speciation rate of state 0, the extinction rates are fixed to 0 (pure birth model), and the transition rates are symmetric. This configuration obviously led to incredibly good results, obtained with the MuSSE model, with a similar architecture to the ones used in the previous experiment, and especially with only 1000 trees. Since the assumptions were not realistic, I decided to remove the constraints and train the models without them. More specifically, I uniformly sampled every parameter in the given range, without deriving any relation between them, and the same was done for the transition rates, making them unsymmetric. The predictions were, at this point, completely off the true values, consistently with the previous experiments. This confirms that the constraints imposed by the authors are necessary to obtain good results with the deep learning models, and that the models are not able to learn the relation between the parameters without them, even considering the manipulation of the training set that the authors did.

6. Model Predictor

The model predictor is a new introduction in this field: it is a tool that allows to predict the model that generated a given tree. Without this new approach, the only way to guess it would be to run Maximum Likelihood Estimation for every possible model, with an high probability of not having an analytical solution for some of them. Various classification models were trained on the trees generated with the different models, and the best model was chosen based on the accuracy of the predictions. I tried both the CDV and SS encoding to compare the results obtained with the different encodings.

The trees were then concatenated, and the targets were stored in a numpy array. The data was then split into training and testing sets, and various classification models were trained on the training set. The accuracy of the predictions was then evaluated on the testing set, and the best model was saved. The models used were a Random Forest Classifier, a Multi-Layer Perceptron Classifier, a Deep Neural Network, and a Convolutional Neural Network. The accuracy of the predictions was evaluated on 10 different seeds for each model. The number of available trees, in both cases, was around 30'000, equally distributed among the different models.

The performance comparison on the test dataset of the models trained on CDV encoding is shown in Fig. 3. The one with the SS encoding is shown in Fig. 4. All the models were trained on 10 different seeds, to get the mean accuracy and standard deviation. In both cases the best model is the **Random Forest Classifier**, with a mean accuracy of **62% for the CDV encoding and 68% for the SS encoding**, and an incredibly low standard deviation of 0.0014, that makes the model really reliable: the 95% confidence interval is (0.6787, 0.6844)%, which means that, on average, 95% of the times the accuracy of this model is inside that range. The baseline accuracy is 20%, since there are 5 models to choose from, and the accuracy is $1/5 = 20\%$ for a random guess. This new method not only drastically reduces the computational time and resources previously needed to accomplish this task (from minutes to milliseconds) but also makes the inference possible even in cases where it was not, achieving all this with incredibly high accuracy.

7. Conclusions

This work aimed to investigate the use of deep learning techniques for inferring parameters of diversification models from extant phylogenies to substitute Maximum Likelihood Estimation methods, which are slow, computationally expensive, and often lack an analytical solution. The analysis included Convolutional Neural Networks, Linear models, Feed Forward Neural Networks, Gated Recurrent Units (GRU), and

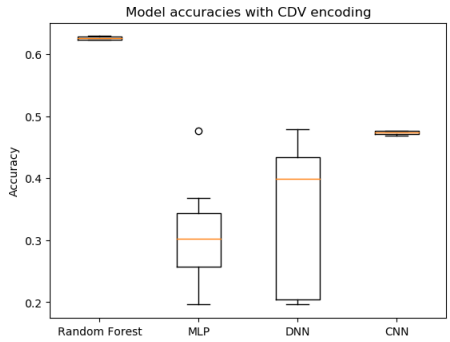

Figure 3: Test accuracy of model predictors trained on CDV-encoded trees, on 10 different seeds each. As you can see from the standard deviation of the random forest classifier, its high accuracy is also really reliable.

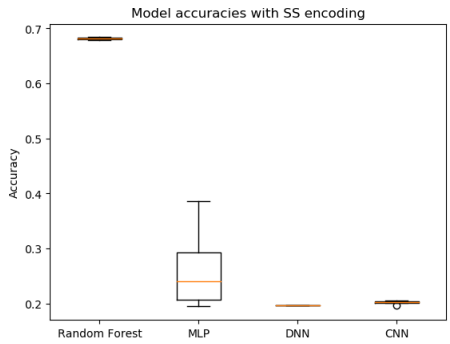

Figure 4: Test accuracy of model predictors trained on SS-encoded trees, on 10 different seeds each. As you can see from the standard deviation of the random forest classifier, its high accuracy is also really reliable.

Long Short-Term Memory (LSTM) Networks. It led to positive results, only limited by the computational constraints this work was subjected to.

The study mainly involved two types of encoding (CDV and Summary Statistics). Since these encodings do not follow a sequential type of representation, they were not suited for Recurrent Neural Networks (GRU and LSTM), which performed the worst. All the other networks, however, showed good accuracy, even considering the limit on computational resources.

A novel approach was introduced in the field: using a Random Forest Classifier to infer parameters from the trees. Its performance in prediction accuracy was consistently the best compared to other deep learning models, and its training time was a fraction of that of other models.

Additionally, while aiming to extend the work of another research on the matter, it emerged that the declared results were caused by the imposition of strong constraints, which limit the applicability of these models to real-world scenarios.

Furthermore, this work presents a completely new approach in the field: a tool capable of inferring the mathematical model of a given phylogeny. Without it, the only way to guess the diversification model of a phylogeny is to run Maximum Likelihood Estimation for every possible model, with a high probability of not having an analytical solution for some of them. The study involved the analysis of a Multi-Layer Perceptron, a Linear model, a Convolutional Neural Network, and a Random Forest Classifier. The latter outperformed the others, achieving a really high mean test accuracy of 68%, with a 95% confidence interval of (0.6787, 0.6844)%. This new method not only drastically reduces the computational time and resources previously needed to accomplish this task (from minutes to milliseconds) but also makes the inference possible even in cases where it was not, achieving all this with incredibly high accuracy.

References

[1] Ismaël Lajaaiti, Sophia Lambert, Jakub Voznica, Hélène Morlon, and Florian Hartig. A comparison of deep learning architectures for inferring parameters of diversification models from extant phylogenies, 2023. bioRxiv.

[2] S. Lambert, J. Voznica, and H. Morlon. Deep learning from phylogenies for diversification analyses. 2023.