BioF4C-Frame: A Framework for Feature Preparation and Selection applied to Ecological Data Classification

Luma Rios Delponte[1], Carina F. Dorneles[1] (Advisor), Simone S. Werner[1] (Co-advisor)

[1]Department of Informatics and Statistics – INE – Federal University of Santa Catarina – Brazil

rios.luma@posgrad.ufsc.br, carina.dorneles@ufsc.br, simone.werner@ufsc.br

Abstract. Feature selection is key to improving classification models for biological databases, especially those with high dimensionality and inconsistencies, such as species incidence data for plants, algae, and fungi. These datasets often have redundant variables, class imbalance, and taxonomic inconsistencies, hindering performance and interpretability. Despite its potential, FS is underexplored in ecological and botanical contexts, particularly in biodiversity hotspots like Brazil's Atlantic Forest. This study presents BioF4C-Frame, a modular framework combining preprocessing, similarity functions, FS strategies, and classification models. Empirical analysis assesses how methods like LASSO interact with classifiers such as Random Forest, Decision Trees, Naïve Bayes, ANN, and Logistic Regression. Random Forest showed the best performance, handling complex, imbalanced data effectively. LASSO and ANN offered modest but relevant gains, especially after deduplicating taxonomic authorship fields using similarity functions. The study proposes FS practices for biodiversity data, improves classification reliability, and highlights future research gaps.

Resumo. A seleção de atributos é essencial para melhorar modelos de classificação em bases biológicas, especialmente aquelas com alta dimensionalidade e inconsistências, como dados de incidência de espécies de plantas, algas e fungos. Esses dados apresentam variáveis redundantes, desequilíbrio entre classes e inconsistências taxonômicas, afetando desempenho e interpretabilidade. Apesar do potencial, a seleção de atributos é pouco explorada em contextos ecológicos e botânicos, sobretudo em hotspots como a Mata Atlântica. Este estudo apresenta o BioF4C-Frame, um framework modular que integra pré-processamento, funções de similaridade, estratégias de seleção e modelos de classificação. A análise empírica avalia como métodos como o LASSO interagem com classificadores como Random Forest, Árvores de Decisão, Naïve Bayes, Redes Neurais e Regressão Logística. O Random Forest obteve o melhor desempenho, lidando bem com dados ecológicos complexos e desbalanceados. LASSO e Redes Neurais apresentaram ganhos modestos, especialmente após deduplicação de campos de autoria taxonômica com funções de similaridade. O estudo propõe práticas de seleção para dados de biodiversidade, aprimora a confiabilidade da classificação e indica lacunas para pesquisas futuras.

1) Thesis Defense Data and Highlights:

Category: Master's Degree Defense Date: February 26, 2025
Advisor: Carina F. Dorneles (UFSC)
Co-advisor: Simone S. Werner (UFSC)
Board Members: Ronaldo Mello (UFSC), Renato Fileto (UFSC), Sérgio Lifschitz (PUC-Rio)

Highlights:

- Modular pipeline for FS and classification in ecological data
- Taxonomic deduplication using ICN-aligned similarity metrics
- Empirical benchmark of FS strategies and ML classifiers applied to over 16,000 records of Atlantic Forest *Begoniaceae*
- Hybrid feature selection excels in imbalanced, high-dimensional ecological data
- Random Forest proved effective in biodiversity modeling

Dissertation: https://repositorio.ufsc.br/handle/123456789/265521
Publications: https://sol.sbc.org.br/index.php/sbbd/article/view/30709

2 Context and Problem

Ecological datasets, especially those from biodiversity hotspots like Brazil's Atlantic Forest, are marked by taxonomic inconsistencies, high dimensionality, and significant class imbalance [15, 11]. In the context of *Begoniaceae* species, such issues hinder automated analysis and conservation planning [1, 2]. The shortage of taxonomists and the variety of sources further complicate modeling [16, 8, 9]. Feature Selection (FS) can reduce dimensionality and improve model interpretability [10, 3], but comparative studies in biodiversity contexts remain rare [13]. This work addresses this gap with BioF4C-Frame—a framework integrating taxonomic deduplication, FS strategies, and classification models tailored to ecological datasets.

3 Objective

To design, implement, and evaluate BioF4C-Frame, a reproducible and modular framework for feature preparation and selection in biological datasets. It addresses noise, redundancy, and class imbalance in ecological records to improve species identification and distribution. Goals include taxonomic data cleaning, FS method benchmarking, and classifier comparison on high-dimensional, imbalanced data.

4 Contribution

BioF4C-Frame is a modular and replicable framework that combines preprocessing, taxonomic deduplication, feature selection (FS), and classification strategies tailored for ecological data analysis. It uniquely integrates authorship normalization based on the International Code of Nomenclature (ICN), using seven domain-adapted text similarity functions (e.g., Smith-Waterman, Levenshtein, Metaphone), and evaluates their effectiveness for author field deduplication. The FS module supports four strategies (filter, wrapper, embedded, and hybrid), including empirical benchmarking of LASSO, RFE, and mutual information methods. Five classification models—Random Forest, ANN, Naïve Bayes, Logistic Regression, and Decision Tree—are tested on 16,608 records of *Begoniaceae* from SPLINK and Reflora. Two target variables, `stateprovince` and `scientificname`, were prioritized for evaluation. The solution achieved robust predictive performance (F1-scores up to 0.77), and demonstrated substantial improvements after deduplication. The entire codebase is publicly available, supporting reproducibility and transparency. This framework contributes to biodiversity conservation by improving classification reliability and highlighting the role of semantic data curation in ecological databases.

5 Advancement over the State of the Art

Most existing studies treat feature selection (FS) and classification as separate tasks, without addressing the semantic inconsistencies common in biodiversity data [7, 13]. Although FS methods such as Mutual Information, RFE, and LASSO are widely used in medical and environmental contexts [17, 5, 12], their application in ecological scenarios—especially with taxonomic inconsistencies and class imbalance—remains limited. This work advances the state of the art by proposing **BioF4C-Frame**, a modular pipeline that integrates data harmonization, ICN-guided taxonomic deduplication [14], FS benchmarking, and classification tailored to ecological datasets. The framework uniquely evaluates seven string similarity functions under real biodiversity data, demonstrating their direct impact on predictive performance.

Unlike prior works such as Bourel and Segura [4] or Cutler et al. [6], which focus on classifier performance without semantic preprocessing, BioF4C-Frame incorporates a preprocessing layer that improves model input quality and interpretability. It also systematically compares filter, wrapper, embedded, and hybrid FS strategies across five classifiers, with empirical validation using species incidence data from the Atlantic Forest. To the best of our knowledge, this is the first framework to jointly address semantic deduplication, FS strategy benchmarking, and ecological classification in a reproducible manner, offering practical contributions to biodiversity modeling and conservation.

6 Solution Overview

The BioF4C-Frame pipeline consists of five major stages: (1) **Data Harmonization**, which standardizes fields and addresses inconsistencies in botanical records; (2) **Authorship Deduplication**, employing seven string similarity functions (e.g., Smith-Waterman, Jaro-Winkler, Fingerprinting) to reduce redundancy in the `scientificnameAuthorship` field, following ICN guidelines; (3) **Feature Selection**, implemented through filter (e.g., mutual information), wrapper (e.g., RFE), embedded (e.g., LASSO), and hybrid strategies, with emphasis on feature stability and dimensionality reduction; (4) **Classification**, using five models with varying degrees of complexity and interpretability (RF, ANN, NB, LR, DT); and (5) **Validation**, performed via stratified 10-fold cross-validation, ensuring balanced representation of classes. The framework evaluates models using precision, recall, F1-score, and confusion matrices. Performance comparisons before and after deduplication revealed that semantic preprocessing directly enhances classification outcomes—especially when using Smith-Waterman on complex author strings. Feature importance analysis further highlighted that the `firstAuthor` variable, once normalized, became a key predictor for taxonomic classification.

7 Evaluation

Stratified cross-validation showed that Random Forest achieved the highest F1-scores: 0.77 for `stateprovince` and 0.74 for `scientificname`. LASSO proved effective for dimensionality reduction, preserving performance while simplifying models. Deduplication significantly improved results, with Smith-Waterman yielding the best author name disambiguation. Redundant authorship data impaired accuracy, reinforcing the need for domain-specific preprocessing. The normalized first author emerged as the most predictive feature. Hybrid FS methods combining wrapper and embedded strategies performed best in imbalanced scenarios. These results highlight the impact of integrating semantic preprocessing, FS, and interpretability in ecological modeling.

8 Scientific and Technical Production, Awards

Scientific Production

- Luma G. R. Cerqueira, Carina F. Dorneles, Simone S. Werner *Optimizing Botanical Data Integrity: A Comparative Study of Text Similarity Methods*, SBBD 2024.
- Delponte, L. R., Dorneles, C. F., Werner, S. S. *Beyond Species: Enhancing Botanical Data Integrity Using Similarity Metrics in Authorship Attibuition* Submitted and Accepted to Journal of Information and Data Management, 2025.

9 Acknowledgements

A special thanks to the brazilian National Council for Scientific and Technological Development (CNPq) and the University of Santa Catarina (UFSC) for the financial support that made this work possible. Also, the Department of Informatics and Statistics (INE) at UFSC and the Systems Integration and Application Laboratory (LISA) for the shared knowledge.

References

[1] Reflora - herbário virtual, 2020. URL `http://floradobrasil.jbrj.gov.br/reflora/herbarioVirtual/`. Accessed: 2020-12-14.

[2] Alexandre Antonelli, Chris Fry, Richard J Smith, James Eden, Rafaël H A Govaerts, Paul Kersey, Eimear Nic Lughadha, and Andrea R Zuntini. State of the world's plants and fungi 2023. *Royal Botanic Gardens, Kew*, 2023.

[3] V. Bolón-Canedo, N. Sánchez-Maroño, et al. Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, 8(2):93–110, 2019. doi: 10.1007/s13748-015-0080-y.

[4] M. Bourel and A.M. Segura. Multiclass classification methods in ecology. *Ecological Indicators*, 85:1012–1021, 2018. doi: 10.1016/j.ecolind.2017.11.031.

[5] Rung-Ching Chen, Christine Dewi, Su-Wen Huang, and Rezzy Eko Caraka. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1):52, 2020. doi: 10.1186/s40537-020-00327-4. URL `https://doi.org/10.1186/s40537-020-00327-4`.

[6] D. Richard Cutler, Thomas C. Edwards Jr., Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, and Joshua J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[7] Dimitrios Effrosynidis and Avi Arampatzis. An evaluation of feature selection methods for environmental data. *Ecological Informatics*, 61:101224, 2021. doi: 10.1016/j.ecoinf.2021.101224.

[8] Zoë A Goodwin, David J Harris, Denis Filer, John RI Wood, and Robert W Scotland. Widespread mistaken identity in tropical plant collections. *Current biology*, 25(22):R1066–R1067, 2015.

[9] J. Hortal, F. Bello, J.A.F. Diniz-Filho, T.M. Lewinsohn, J.M. Lobo, and R.J. Ladle. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46:523–549, 2015.

[10] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):923–936, 2017. doi: 10.1109/TPAMI.2018.2858826.

[11] N. Myers, R.A. Mittermeier, C.G. Mittermeier, G.A.B. Fonseca, and J. Kent. Biodiversity hotspots for conservation priorities. *Nature*, 403:853–858, 2000.

[12] Elnaz Pashaei and Nizamettin Aydin. Binary black hole algorithm for feature selection and classification on biological data. *Applied Soft Computing*, 56, 03 2017. doi: 10.1016/j.asoc.2017.03.002.

[13] P. Schratz, J. Muenchow, E. Iturritxa, et al. Monitoring forest health using hyperspectral imagery: Does feature selection improve the performance of machine-learning techniques? *Remote Sensing*, 13(23):4832, 2021. doi: 10.3390/rs13234832.

[14] Nick J Turland, John Harry Wiersema, Fred R Barrie, Werner Greuter, David L Hawksworth, Patrick Stephen Herendeen, Sandra Knapp, Wolf-Henning Kusber, De-Zhu Li, Karol Marhold, et al. *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017.* Koeltz botanical books, 2018.

[15] R.J. Whittaker, M.B. Araújo, P. Jepson, R.J. Ladle, J.E.M. Watson, and K.J. Willis. Conservation biogeography: Assessment and prospect. *Diversity and Distributions*, 11:3–23, 2005.

[16] E.O. Wilson. Biodiversity research requires more boots on the ground. *Nature Ecology & Evolution*, 1:1590–1591, 2017. doi: 10.1038/s41559-017-0346-0.

[17] Yongbo Zheng, Yueqiang Peng, Yingying Gao, Guo Yang, Yu Jiang, Gaojie Zhang, Linfeng Wang, Jiang Yu, Yong Huang, Ziling Wei, and Jiayu Liu. Identification and dissection of prostate cancer grounded on fatty acid metabolism-correlative features for predicting prognosis and assisting immunotherapy. *Computational Biology and Chemistry*, 115:108323, 2025. ISSN 1476-9271. doi: https://doi.org/10.1016/j.compbiolchem.2024.108323. URL `https://www.sciencedirect.com/science/article/pii/S1476927124003116`.