Twenty Years of Random Forest: preliminary results of a systematic literature review

Venti anni di Random Forest: una review sistematica preliminare

Massimo Aria, Agostino Gnasso and Luca D'Aniello

Abstract The Random Forest (RF) model consists of an ensemble classifier that produces many decision trees through the use of a randomly selected subset of samples and training variables. The RF model has assumed importance within the scientific community thanks to its performance. The accuracy of its classifications and prediction has allowed the use of RF in several research domains, which have benefited from it. The present study aims to provide a preliminary review of the whole scientific production characterized by all the publications citing the article "Random Forest" by Breiman, 2001, in the last 20 years (2001-2021).

Abstract *L'approccio Random Forest (RF) consiste in un classificatore di ensemble che produce un grande numero di alberi decisionali, attraverso l'uso di un sottoinsieme di variabili, casualmente selezionato. Il modello RF ha assunto importanza all'interno della comunità scientifica grazie alle sue prestazioni, nonchè granzie all'accuratezza delle sue classificazioni e previsioni. Il presente studio mira a fornire un'antemprima preliminare dell'intera produzione scientifica caratterizzata da tutti i lavori accademici che hanno citato l'articolo "Random Forest" di Breiman, 2001, negli ultimi 20 anni (2001-2021).*

Key words: Random Forest, Bibliometrics, Systematic literature review, Science mapping

Massimo Aria
Department of Economics and Statistics, University of Naples Federico II, Italy
email: massimo.aria@unina.it

Agostino Gnasso
Department of Economics and Statistics, University of Naples Federico II, Italy
email: agostino.gnasso@unina.it

Luca D'Aniello
Department of Social Sciences, University of Naples Federico II, Italy
email: luca.daniello@unina.it

1 Introduction

Machine learning is a data analysis approach that automates the construction of analytical models. It is a branch of Artificial Intelligence based on the idea that systems can learn from data, identify patterns on their own and make decisions with minimal human intervention [10]. The use of Ensemble methods in Machine Learning provides different predictive models with different results for the same inputs. Ensemble Learning approaches increase predictive performance models by combining the outputs of a set of induced hypotheses, also called base learners, into a single predictive model. It has the purpose of decreasing variance, altering bias, and improving predictions. An ensemble learner can match any machine learning algorithm such as the decision tree, neuronal network, or a linear regression model. Classification and Regression Trees (CART) are supervised learning techniques that use a nonparametric approach [5]. The process of building trees is intuitive and simple for the human mind, which implies a simple and useful interpretation, but it is not competitive in terms of accuracy concerning other regression and classification approaches. However, predictive performance can be substantially improved by aggregating many decision trees.

In 2001 Leo Breiman proposed Random Forest (RF) method [4], a non-linear approach aims to achieve greater accuracy by averaging multiple decision trees, each of which is grown according to two random steps: the first step consist of the use of a bootstrap sample to train each tree, while the second is the use, at each internal node, of a random subset of variables to generate splits. RF is an evolution of Bagging which aims to reduce the variance of a statistical model, simulates the variability of data through the random extraction of bootstrap samples from a single training set and aggregates predictions on a new record [3].

The purpose of this work is to present a systematic literature review of the last twenty years – from the publishing of RF papers to date - and identify the main research domain that use RF method through quantitative and longitudinal analysis.

2 Materials and methods

Bibliometrics has the potential to introduce a systematic, transparent, and reproducible review process based on the statistical measurement of science, scientists, or scientific activity [6] [12]. Bibliometric analysis involves quantitative methods for exploring, monitoring, and measuring of published research with a set of tools within one or more specified fields over a given period of time [15].

This work aims to perform a bibliometric analysis to investigate thought the knwoledge of scientific literature of all publications citing the RF article with the science mapping approach. As defined by Tijssen and van Raan [14], science mapping plays a crucial role in the study of knowledge structures underlying research and development (R&D) developments. It depicts the structural and dynamic aspects of a scientific research domain [8] from a quantitative and qualitative viewpoint.

As described by Chen [7], the unit of analysis in science mapping is a domain of scientific knowledge that is reflected through an aggregated collection of intellectual contributions from members of a scientific community or more precisely defined specialties.

The review process was performed using bibliometrix, an R-package (`http://www.bibliometrix.org`) that provides a set of tools for quantitative research in bibliometrics and scientometrics [1]. Bibliometrix is a unique tool according to a logical bibliometric workflow and incorporates a wide variety of different analyses.

3 Data collection and findings

To retrieval the bibliometric data, we queried the Web of Science (WoS) indexing database on October 2021. WoS was launched by the Institute for Scientific Information (ISI) and now it is maintained by Clarivate Analytics. It represent of the main databases allowing to explore the literature of several scientific domains. It includes several citation databases specialised on specific fields covering more than 20000 journals, conference proceedings and books [2].

This systematic literature review and meta-analysis is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines, illustrating the outcomes of the literature searches and article selection process [11].

We downloaded 43.887 publications citing the RF article from WoS. We filtered our collection by selecting only the publications classified as articles and reviews from January 2001 to October 2021. Moreover, we retrieved only the publications written in English. The final collection included 34.713 publications.

In Table 1 there are some descriptive information about the whole collection.The publications has been published on 5.491 sources and collected over one million of references. The collection is charaterized by only the 3.25 % of reviews. The number of single-authored documents is 911. This means that most of publications were written by at least two authors. Indeed the average number of authors per document is 3.

The Figure 2 shows the annual scientific production which provides an overview of the number of papers that have cited RF in the last 20 years. The annual growth rate is on average 49.6%. To date, the number of publications is almost equal to the number of publications published in the 2020. This highlights the constant need for researchers to use this methodology in their works.

In Figure 3 a word cloud of the most frequent is reported. To perform this analysis, we considered the KeyWords Plus (ID) used in the different documents. The ID are words or phrases that frequently appear in the titles of an article's references but do not appear in the title of the publication itself. Their generation is based upon a special algorithm [9] that is unique to WoS databases. The most frequent ID is "classification". This means that the RF is used in many classifications works such as text classification and image classification. [13].

MAIN INFORMATION ABOUT DATA	
Timespan	2001:2021
Sources (Journals, Books, etc)	5491
Documents	34713
Average years from publication	3.38
Average citations per documents	21.88
Average citations per year per doc	4.082
References	1066621
DOCUMENT TYPES	
Article	33586
Review	1127
DOCUMENT CONTENTS	
Keywords Plus (ID)	39276
Author's Keywords (DE)	63024
AUTHORS	
Authors	109632
Author Appearances	200304
Authors of single-authored documents	911
Authors of multi-authored documents	108721
AUTHORS COLLABORATION	
Single-authored documents	1092
Documents per Author	0.317
Authors per Document	3.16
Co-Authors per Documents	5.77
Collaboration Index	3.23

Fig. 1 Main information about data

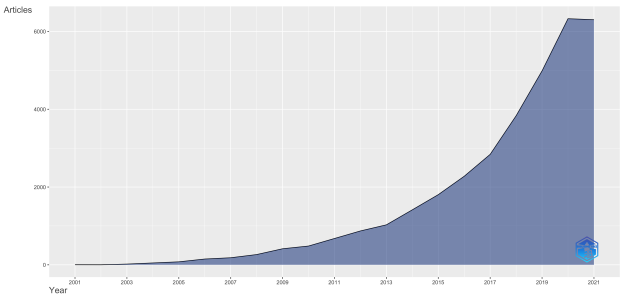

Fig. 2 Annual Scientific Production - Timespan 2001:2021

4 Conclusion

The RF is a user-friendly, intuitive, and very fast approach. It less training time than Decision Tree and Support Vector Machine. RF becomes functional in cases of large datasets because it helps avoid the problem of overfitting by managing the noises present in datasets. As seen in our work, thanks to these capabilities, RF, proposed by Breiman [4], is gaining more and more popularity in the research community for classification and prediction tasks, even twenty years after its publication.

Future developments will be devoted on a detailed analysis of the massive collection used in this work through quantitative and statistical advanced methods. In

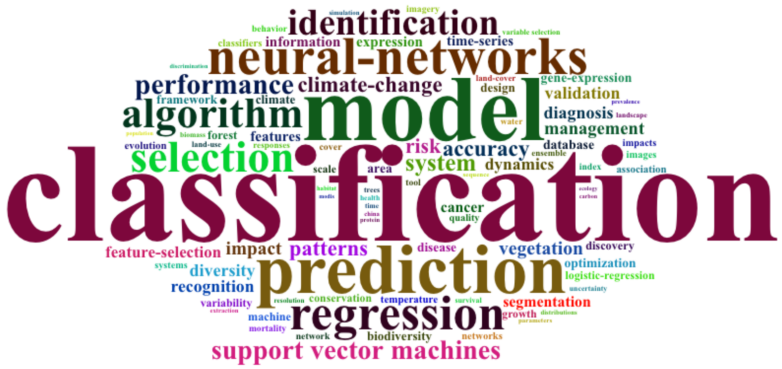

Fig. 3 Word cloud of the most frequent ID

particular, the research domains that use the RF method will be identified to understand what is its main use in different research areas.

References

1. Aria, M., Cuccurullo, C.: bibliometrix: An r-tool for comprehensive science mapping analysis. Journal of informetrics, **11**,(4), pp. 959–975 (2017)
2. Aria, M., Misuraca, M., Spano, M.: Mapping the evolution of social research and data science on 30 years of social indicators research. Social Indicators Research, **149**,(3), pp. 803–831 (2020)
3. Breiman, L.: Bagging predictors. Machine learning, **24**, (2),pp. 123–140 (1996)
4. Breiman, L.: Random forests. Machine learning, **45**, (1), pp. 5–32 (2001)
5. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.: Classification and regression trees. Routledge (2017)
6. Broadus, R. N.: Toward a definition of "bibliometrics". Scientometrics, **12**, (5-6), pp. 373–379 (1987)
7. Chen, C.: Science mapping: a systematic review of the literature. Journal of data and information science, **2**, (2) (2017)
8. Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., Herrera, F.: Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for information Science and Technology, **62**, (7), pp. 1382–1402 (2011)
9. Garfield, E.: Current eamments. Curr. Contents, **32**, pp. 3–7 (1990)
10. Mitchell, T.: Machine Learning. McGraw Hill (1997)
11. Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J. Clin. Epidemiol. **62** (10), pp. e1-e34, (2009)
12. Pritchard, A.: Statistical bibliography or bibliometrics. Journal of documentation, **25**(4), pp. 348–349 (1969)

13. Song, Q., Liu, X., Yang, L.: The random forest classifier applied in droplet fingerprint recognition. In 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) pp. 722-726, IEEE, (2015)
14. Tijssen, R. J., Van Raan, A. F.: Mapping changes in science and technology: Bibliometric co-occurrence analysis of the R&D literature. Evaluation Review, **18**,(1), pp. 98–115 (1994)
15. Van Raan, A. F.: Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics, **62**,(1), pp. 133–143 (2005)