

Understanding the Difference between Decision Tree and Random Forest Algorithms

Random forests have become an important tool for improving accuracy in regression and classification problems since their inception by Leo Breiman in 2001. In this paper, we revisit a historically important random forest model, called centered random forests, originally proposed by Breiman in 2004 and later studied by Gérard Biau in 2012, where a feature is selected at random and the splits occurs at the midpoint of the node along the chosen feature. If the regression function is d -dimensional and Lipschitz, we show that, given access to n observations, the mean-squared prediction error is $O((n(\log n)(d-1)/2) - 1d\log 2 + 1)$. This positively answers an outstanding question of Biau about whether the rate of convergence for this random forest model could be improved beyond $O(n - 1d(4/3)\log 2 + 1)$. Furthermore, by a refined analysis of the approximation and estimation errors for linear models, we show that our new rate cannot be improved in general. Finally, we generalize our analysis and improve current prediction error bounds for another random forest model, called median random forests, in which each tree is constructed from subsampled data and the splits are performed at the empirical median along a chosen feature.

Machine Learning is a significant technique to realize Artificial Intelligence. The Random Forest Algorithm can be considered as one of the Machine Learning's representative algorithm, which is known for its simplicity and effectiveness. It is also can be defined as a Decision Tree-Based Classifier that chooses the best classification tree as the final classifier's classification of the algorithm via voting. Random Forest is the most accepted group classification technique because of having excellent features such as Variable Importance Measure, Out-of-bag error, Proximities, etc. Currently, it is in the new classification, intrusion detection, content information filtering, and sentiment analysis that is why there is an extensive range of applications in image processing. In this paper, the construction process of Random Forests and the study status of Random Forests would primarily be introduced in terms of capacity enhancement and performance indicators. The use of Random Forest in different fields such as Medicine, Agriculture, Astronomy, etc. is often mentioned.

In this paper, we propose energy-based least squares projection twin support vector machines (ELSPTSVM) for classification problems. The proposed ELSPTSVM models seek projection directions for each class in such a manner that the data of each class are well separated from the data of other class in its respective subspace. To further enhance the performance, the proposed ELSPTSVM model uses a recursive procedure to generate multiple projection axis for each class. The proposed ELSPTSVM model uses energy parameters and solves a system of linear equations to boost the classification performance. The experimental results show the efficiency of the proposed ELSPTSVM model on

benchmark UCI datasets with binary class and some non-UCI datasets of fecundity estimation for fisheries data of binary class.

Machine learning predictive models are nowadays widely applied in various systems in both commercial and public areas. The need to understand and comprehend their behavior has arisen and was, through an extensive increase in research in Explainable AI (XAI), noticed in the last years. One of the problems in the aforementioned XAI spectrum is understanding the large rule sets, such as those mined from big datasets or induced through Random Forest algorithms. This work explores the possibility of tackling such an issue using multiple nonoverlapping decision trees created with incorporated knowledge provided by the set of rules. Random Forest is being used as a source of such rules. Performance evaluation conducted on commonly used datasets shows that such a model could be competitive with an equally deep Random Forest while providing explanations that can be adjusted and interpreted as a rule list.

In the current age of the Fourth Industrial Revolution (4IR or Industry 4.0), the digital world has a wealth of data, such as Internet of Things (IoT) data, cybersecurity data, mobile data, business data, social media data, health data, etc. To intelligently analyze these data and develop the corresponding *smart and automated* applications, the knowledge of artificial intelligence (AI), particularly, *machine learning (ML)* is the key. Various types of machine learning algorithms such as supervised, unsupervised, semi-supervised, and reinforcement learning exist in the area. Besides, the *deep learning*, which is part of a broader family of machine learning methods, can intelligently analyze the data on a large scale. In this paper, we present a comprehensive view on these *machine learning algorithms* that can be applied to enhance the intelligence and the capabilities of an application. Thus, this study's key contribution is explaining the principles of different machine learning techniques and their applicability in various real-world *application* domains, such as cybersecurity systems, smart cities, healthcare, e-commerce, agriculture, and many more. We also highlight the challenges and potential *research directions* based on our study. Overall, this paper aims to serve as a reference point for both academia and industry professionals as well as for decision-makers in various real-world situations and application areas, particularly from the technical point of view.

References

1. Klusowski, J. (2021). Sharp Analysis of a Simple Model for Random Forests. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (PMLR 130), 757–765.
2. Abdulkareem, N. M., & Abdulazeez, A. M. (2021). Machine Learning Classification Based on Random Forest Algorithm: A Review. International Journal of Science and Business, 5(2), 128–142.

3. Ganaie, M. A., Tanveer, M., Suganthan, P. N., & Snasel, V. (2021). Oblique and rotation double random forest. arXiv preprint arXiv:2111.02010.
4. Cic-ddos2019 [online]. available: <https://www.unb.ca/cic/datasets/ddos-2019.html/> (Accessed on 28 March 2020).