## ***Original Research Article***

## **ON ENSEMBLE LEARNING: CONTRIBUTION TO THE DEVELOPMENT OF THE MACHINE LEARNING MODEL BY MODELING AN ESTIMATOR BASED ON AGGREGATION METHODS**

# 1) Abstract

The objective of this article is to contribute to the development of the ensemble model in Machine Learning by creating a method based on homogeneous and heterogeneous families, with the ambition of obtaining better quality predictions.

To illustrate this approach, we can cite already proven techniques such as Random Forest, which uses the Bagging method, as well as Adaboost and Gradient Boosting, which are part of the sequential ensemble method. Other approaches such as SVM, K-means and neural networks also enrich this panorama.

The central idea of our method is to combine multiple ensemble models to optimize errors at each iteration. This process allows for more informed decisions for future predictions. Indeed, by integrating the strengths of various models, we can mitigate their respective weaknesses and thus improve the robustness of predictions.

As the saying goes, “Unity is strength.” This maxim is particularly resonant in the context of building a robust machine learning model, as it highlights the importance of collaboration between different approaches to achieve superior results. By uniting these different techniques, we aspire to create a model that can adapt to the complexities of the data and provide reliable and accurate results.

**Keywords** : Machine Learning, Algorithm, Classifiers, prediction, Variable

**2) Introduction**

Following our extensive research, we found that the Stacking algorithm, which combines multiple estimators from sequential and parallel families into a homogeneous group, has undeniable strengths in the field of Machine Learning. This mechanism aims to create a meta-estimator that predicts the final result based on the predictions of various classifiers. However, we have highlighted a major limitation: this algorithm fails to effectively correct prediction errors at different learning levels. This inability to adjust errors over iterations represents a significant opportunity for improvement.

It is with this in mind that we considered enriching the Stacking ensemble algorithm by integrating reinforcement learning concepts. This innovative approach would allow combining homogeneous and heterogeneous algorithms while optimizing errors at each learning level. Indeed, in our proposal, the errors generated at the first level would be transmitted to the second, and so on, until the formation of a Multi-meta-estimator that we call Lady\_Forest. This method aims to strengthen the learning capacity of models by ensuring that each estimation level learns from the errors of the previous ones, thus improving the accuracy and robustness of predictions.

We know that the fundamental goal of a machine learning algorithm is to minimize learning errors as much as possible. However, when using these algorithms, several challenges arise, such as reducing model complexity, lack of K-fold cross-validation, model regulation, establishing a robust ensemble model, and data augmentation. These issues raise crucial questions for researchers and practitioners in the field:

* How to build a model that minimizes its complexity while maintaining high performance?
* What mechanism should be adopted to prevent overfitting and underfitting during the learning process?
* How to identify and eliminate insignificant variables in model training?
* What method should we apply to ensure accurate and reliable predictions?
* What impact does the number of features have on the effectiveness of model training?

In response to these questions, we propose a significant contribution to the development of machine learning models through a new classifier based on aggregation methods. Our goals are clear: reduce model complexity, introduce K-fold validations during training, regularize the model, establish a robust ensemble model, and augment the data before training. This integrated approach promises not only to improve model performance, but also to address current challenges in machine learning, paving the way for more accurate and reliable predictions in various application domains.

**3) Methodology**

The Lady-Forest method allowed us to combine several meta-estimators through algorithms of homogeneous and heterogeneous classes such as Decision Tree, Random Forest, Adaboost, GradientboostingTree, SVM, Neural Network in order to form a Multi-Hyper-Estimators (MHS).

Figure 1 : The figure below shows how our method works

Hyper-Stimateur

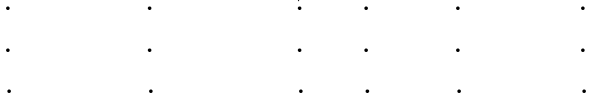
Lady\_Forest

New Data

### **Training Step**

…

*Predictions* *Predictions*



### *Homogeneous Classification Mode Heterogeneous Classification Mode*

Final Predictions

**Interpretation of symbols:**

|  |  |
| --- | --- |
|  | : The different errors of each level of learning. |
|  | : Large set classifiers. |
|  | : The different weights of the large set estimators |
|  | : Mode predictions of homogeneous estimators. |
|  | : Mode predictions of heterogeneous estimators. |
|  | : The different homogeneous and heterogeneous estimators. |
|  | : Meta-Classifier Prediction. |

K-Fold1 and K-Fold2: The different parts of DATASET divided into two to reduce the errors of Overfitting and Underfitting.

**4) Results**

From the methodology described above, integrating several large ensemble estimators into a machine learning framework requires a rigorous and well-defined mathematical approach. We have chosen to focus on the use of Markov chains, in particular non-homogeneous Markov chains. This approach is particularly relevant in the context of machine learning, as it allows to model the temporal evolution of decision processes in a dynamic environment.

Nonhomogeneous Markov chains are distinguished by their ability to account for contextual variations over time. Unlike homogeneous chains, where transition probabilities remain constant, nonhomogeneous chains allow these probabilities to be adapted based on the current state of the system. This is crucial in the field of machine learning, where input conditions can change significantly and unpredictably.

This flexibility brings several benefits. First, it improves the handling of temporal dependency, a key aspect in many learning scenarios, especially those involving sequential or temporal data. By incorporating information about the history of decisions and previous states, we can better anticipate the future behaviors of the estimators, thus increasing their accuracy.

Moreover, this approach ensures optimal convergence of the estimators. By adapting transition probabilities at each step of the training process, non-homogeneous Markov chains facilitate the adjustment of models to the nuances of the data. This not only reduces prediction errors, but also optimizes the model training process by ensuring that they remain responsive to contextual changes.

**4.1. Properties**

Let E be a sequence of random variables with values ​​in the state space .   
Is a Markov chain on *E* when, for all i and j in *E* :

* the event [ =i] depends only on the previous state of , and not on the states that precede it;
* The probability of moving from state i to state j does not depend on n . So to recognize a Markov chain:
* we identify the space of states of the random variable;
* we justify that the state of at step (n+1) depends only on the state of at step n ; and not on the previous states;
* we justify that the probability of moving from one state to another does not depend on the step number.

**4.2. Non-homogeneous Markov process**

Consider ( a continuous-time Markov process, having a finite state space The cumulative intensity measure is another parameter that allows us to define a Markov process. It is a matrix of functions, denoted , such that:

For this purpose, in the case of our method, the cumulative intensity measure could mean the sum of the transition probabilities to a given state, or a group of states during learning in order to adjust the weights of each large estimator according to the rewards obtained for its correct predictions.

1. **Proposed algorithm: “Lady-Forest”**

**Entrance** :

* *Training data:*
* *(m, …, M): The number of iterations.*
* *p: The number of estimators*
* *: The different weights of the large set classifiers*
* *: Great estimator of each level of learning*
* *: Error of each learning level.*

**Output** : A set of estimators

***Step 1: Learning Homogeneous and Heterogeneous Classifiers***

Learn a new estimatorbased on set D

***Step 2: Learning the Large Classifiers***

Learn a new estimator based on its predictions of

1. Initialize the weights= , *i* = 1 *,..., n*
2. Calculate

**For *m* = *1* to *M***:

* 1. Adjusting the weak rule on the sample of estimatorsweighted by the weights *,...,* resulting from this adjustment.
  2. **Calculate the error rate of :**

**(c) Calculation of the weight of the resulting estimators :**

**(d) Readjust the weights of the estimators :**

***Step 3: Construction of a new data set D***

Construction of a new dataset will contain the predictions of the first level

**Step 4: Repeat step (2)**

**Return**

**5) Contribution and Perspective**

We have highlighted the importance of aggregation methods in Machine Learning and proposed an innovative approach based on a set method. This method is based on the following principles:

* Combination of homogeneous and heterogeneous estimators: Integration of models from the same family (decision trees, Random Forest, etc.) and from different families (SVM, neural networks, etc.);
* Large-scale aggregation: Using and combining multiple large estimators to capture complex dependencies in the data;
* Error optimization at each level of interaction: A hierarchical strategy for optimizing intermediate errors, aiming to reduce both variance and bias at each step of the process;
* Building a Hyper-Estimator: An optimized combination of aggregated predictions to form a better performing global model.

Our methodology is inspired by the principles of Q-Learning (due to the sequential optimization of interactions) and Adaboost (for its management of model weights), while distinguishing itself by an improvement of the Stacking algorithm.

This improvement consists of a restructuring and optimized weighting of the basic estimators, which allows to surpass the performances of standard methods. And we have named our estimator approach "Lady-Forest", which is based on a structured organization of the models into interactive subsets, optimized iteratively.

Finally, this contribution opens the way to several perspectives, including the exploration of more advanced optimization strategies for learning model weights, integration of online learning mechanisms for dynamic data and evaluation on specific application domains such as health, finance (cases of fraud or non-fraud), etc.

**Conclusion**

In this paper, we further explore the effectiveness of aggregation methods in Machine Learning by introducing a novel approach, called Lady-Forest, that aims to optimize the performance of ensemble algorithms. Our methodology relies on a carefully optimized combination of homogeneous and heterogeneous estimators, embedded in a hierarchical structure designed to simultaneously reduce the bias and variance of predictions.

By optimizing the interactions between models at each level, we have demonstrated that Lady-Forest outperforms traditional algorithms, such as Stacking or Adaboost, particularly in contexts requiring increased robustness to complex or unbalanced data. This superiority is manifested not only by improved predictive performance, but also by an ability to adapt to diverse and unpredictable environments.

Thus, Lady-Forest represents a significant advance in the field of aggregation methods. By combining the strengths of existing approaches while providing better generalization and increased flexibility, it opens new perspectives for the development of machine learning models. This approach could potentially transform the way we approach the challenges of predictive modeling, laying the foundation for more powerful and adaptive algorithms that can efficiently handle diverse datasets and maximize the accuracy of predictions in real-world applications.

**References**

Bartlett and Traskin M., 2008; Adaboost is consistent, Journal of Machine learning Research, 2347-2368.

Buhlmann P. and Yu B., 2003; Boosting with the L2 Loss: Regression and Classification, Journal of American Statistical Association, 324-339.

Friedman J., Hastie T. and Tibshirani R., 2002; Additive Logistic Regression: A Statical View of Boosting (whith discussion), The Annals of Statistics, 337-407.

Thomas G. Dietterich., 2000; Ensemble Methods in Machine Learning, in Multiple Classifier Systems, pages 1-15 Springer.

ROUVIERE Laurent., 2021; Introduction to aggregation methods: Boosting, Bagging and random forests, Illustrations with R, Lecture notes, University of Rennes 2 Place du Recteur H. Le Modal.

Genuer Robin and PAGGI Jean-Michel., 2017; Cart Trees and Random Forests: Importance and selection of variables, Master's course notes, University of Bordeaux.

MORVANT Emilie., 2013; Majority voting learning for supervised classification and domain adaptation: PAC-BAYESIAN APPROACHES AND COMBINATION OF SIMILARITY2S, Doctoral thesis in mathematics and computer science, Laboratoire d'Informatique Fondamentale UMR 7279-CNRS.

BAHRI Emma., 2010; Improvement of adaptive methods for supervised learning of real data, PhD thesis, University Lyon lumière 2, Faculty of Economics and Management.

DERKAOUI and DERKAOUI Soumia., 2017; Comparative study of set methods for classifying medical data, Master's thesis in Biomedical Engineering, Faculty of Technology, Adou Bakr Belkaid University of Tlemen.

GHODBANE SARA RAYANE., 2021; Law of large numbers: Theories and Application in R, Master's thesis in mathematics, statistics option, faculty of exact sciences and natural and life sciences