Semi-Supervised Rotation Forest

José Miguel Ramírez-Sanz[1 [0000−0003−0189−8046]], David Martínez-Acha[1 [0009−0005−4214−2845]], Álvar Arnaiz-González[1 [0000−0001−6965−0237]], César García-Osorio[1 [0000−0002−1206−1084]], and Juan J. Rodríguez[1 [0000−0002−3291−2739]]

Department of Computer Engineering, Universidad de Burgos, Spain
dmacha@ubu.es

Abstract. This work is about Semi-Supervised Rotation Forest (SSRotF), a semi-supervised extension of Rotation Forest for classification with scarce labeled data (presented in [1]). By incorporating unlabeled instances into both the rotation and training stages, SSRotF achieves strong performance on 54 UCI datasets, outperforming basic wrapper-based SSL methods. A complementary meta-learning analysis also helps identify when SSL is most beneficial.

Keywords: Machine Learning · Semi-Supervised Learning · Rotation Forest · Meta-Learning.

1 Introduction

Machine learning is traditionally divided into two main branches: supervised learning (SL), which relies entirely on labeled data, and unsupervised learning, which finds patterns without predefined labels. Semi-supervised learning (SSL) places between them, aiming to improve predictive performance by combining a small amount of labeled data (L) with a large pool of unlabeled data (U) [4].

Ensemble methods, such as Random Forest (RF) and Rotation Forest (RotF), are highly popular in SL due to their ability to combine multiple weak learners into a robust model. RotF achieves excellent generalization by applying Principal Component Analysis (PCA) to subsets of features to rotate the original data, introducing high diversity among the base decision trees [2]. Nonetheless, traditional RotF cannot exploit unlabeled data by design. This summary presents Semi-Supervised Rotation Forest (SSRotF), an intrinsically semi-supervised adaptation of RotF that integrates unlabeled instances into both the data rotation phase and the base learner training [4].

2 Semi-Supervised Rotation Forest

Adapting RotF for SSL requires modifying two core components of the original algorithm to benefit from the underlying structure of unlabeled data. Algorithm 1 details the training stage of the proposed method Semi-Supervised Rotation Forest (SSRotF).

Algorithm 1 SSRotF: Training stage of Semi-Supervised Rotation Forest

Require: L (labeled), U (unlabeled), Y (labels), F (features), T (ensemble size), K (subsets), W (target weight)

Calculate centroids for each class in Y from L
for $i = 1$ to T **do**
  Split F into K disjoint subsets $F_{i,j}$
  Pseudo-label U by closeness to the centroids
  for $j = 1$ to K **do**
    Let $L_{i,j}$ and $U_{i,j}$ be the data subsets for features in $F_{i,j}$
    Eliminate a random subset of classes
    Sample 50% of remaining $L_{i,j} \rightarrow L'_{i,j}$
    Sample 10% of remaining $U_{i,j} \rightarrow U'_{i,j}$
    Apply PCA on $L'_{i,j} \cup U'_{i,j} \rightarrow$ coefficients matrix $C_{i,j}$
  end for
  Construct rotation matrix R_i from $C_{i,j}$
  Rearrange columns of R_i to match $F \rightarrow R_i^a$
  Build semi-supervised tree $D_i(W)$ using (LR_i^a, UR_i^a, Y)
end for

2.1 PCA with Unlabeled Data and Semi-Supervised Trees

In SSRotF, we increase diversity by including unlabeled instances in the PCA calculation. To prevent noise, a filtering step is applied: unlabeled instances are pseudo-labeled using the closest class centroids. Only 10% of these filtered instances are subsampled and combined with 50% of the labeled instances to calculate the PCA coefficients.

The base learners are semi-supervised decision trees [3]. These trees evaluate splits using a weighted combination of target space impurity (labeled data) and descriptive space variability (all data):

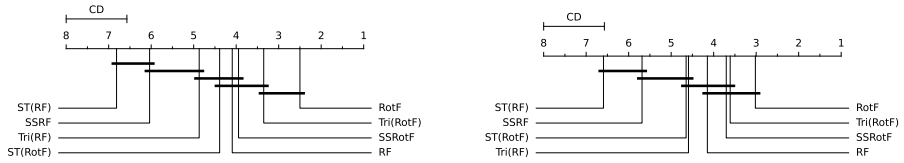
$$Imp(X) = \overbrace{W \cdot ImpTarget(L,Y)}^{Target\ space} + \underbrace{\frac{1-W}{n} \cdot \sum_{i=1}^{n} ImpFeat(X, f_i)}_{Descriptive\ space} \quad (1)$$

In our experiments, a weight of $W = 0.85$ provided the optimal balance between labeled and unlabeled data.

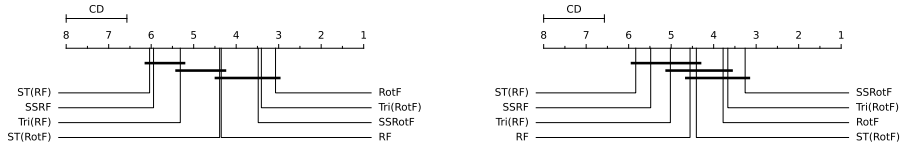
3 Experimental Setup and Results

The proposed SSRotF was evaluated using different supervised and semi-supervised baselines in 54 popular UCI datasets and a 5×2 cross-validation procedure, simulating SSL environments with varying label proportions: from 5% (i.e., only the 5% of data labelled to 99% (i.e., almost SL scenario).

As shown by the Nemenyi post-hoc test in Figure 1, SSRotF consistently ranks among the top-performing methods across evaluation metrics and labeling proportions, outperforming simple wrapper approaches such as ST(RotF).



(a) 10% label proportion



(b) 40% label proportion

Fig. 1: Results of the Nemenyi post-hoc test: F_1 score on the left and MCC on the right for 10% and 40% label proportions.

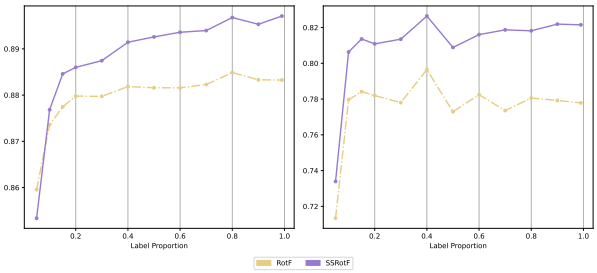
4 Predicting SSL Success via Meta-Learning

A well-known drawback is that SSL only succeeds when specific data assumptions (smoothness, low-density, manifold, and cluster) are met. Unfortunately, to assess whether a dataset fulfills these assumptions is a non-trivial task. Using the *Problexity* library, 22 meta-features were extracted from the 54 datasets. A decision tree classifier revealed that just four characteristics are sufficient to predict when SSL methods will outperform supervised ones: overlap of feature values ($f1$), features' synergy ($f4$), average number of PCA per point ($t3$), and correlation to the target ($c2$).

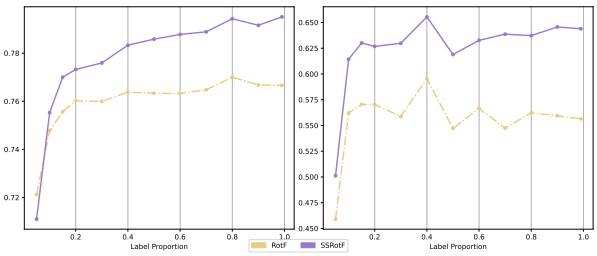
Focusing on the 16 datasets identified as highly suitable for SSL, SSRotF extracts meaningful patterns from unlabeled data, achieving superior results compared to both the purely supervised RotF and other SSL ensembles. Remarkably, in these optimal scenarios, SSRotF trained with only 20%-30% of labeled data achieves performance comparable to a traditional RotF trained with 99% of the labels, as shown in Figure 2.

5 Conclusions

Semi-Supervised Rotation Forest successfully extends the powerful RotF algorithm to deal with both labeled and unlabeled data. By incorporating unlabeled instances into both, the PCA rotation space and the base decision trees, SSRotF achieves robust performance even with scarce labeled data. Furthermore, the accompanying meta-learning analysis provides a data-driven method to identify when the underlying assumptions of SSL are fulfilled.



(a) F_1 score



(b) MCC

Fig. 2: Evolution of method metrics as label proportion increases for `banana` (left) and `mammographic` (right) datasets.

Acknowledgments

This research has been supported by PID2020-119894GB-I00, TED2021-129485B-C43 and PID2023-150694OA-I00 projects funded by MICIU/AEI/10.13039/501100011033 and the FEDER, European Union.

References

1. Ramírez-Sanz, J.M., Martínez-Acha, D., Arnaiz-González, Á., García-Osorio, C., Rodríguez, J.J.: Semi-Supervised Rotation Forest. Journal of Computational Science **94**, 102777 (2026).
2. Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(10), 1619–1630 (2006).
3. Levatić, J., Ceci, M., Kocev, D., Džeroski, S.: Semi-supervised classification trees. Journal of Intelligent Information Systems **49**, 461–486 (2017).
4. van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning **109**(2), 373–440 (2020).