Incremental Random Forest for Unsupervised Learning

Li-Chiao Wang
Department of Industrial Engineering and Engineering Management
National Tsing Hua University
Hsinchu, Taiwan
bonny@gapp.nthu.edu.tw

Wei Liu
School of Computer Science
University of Technology Sydney
Sydney, Australia
wei.liu@uts.edu.au

Chung-Shou Liao
Department of Industrial Engineering and Engineering Management
National Tsing Hua University
Hsinchu, Taiwan
csliao@ie.nthu.edu.tw

***Abstract*—Incremental learning aims to develop models capable of continuously learning knowledge, consistent with many real-world scenarios where data evolves over time. While most of the existing research work focuses on supervised incremental learning, it is a challenging task to obtain all the required data labels in some cases. This study proposes a novel incremental learning approach, Incremental Random Forest for Unsupervised Learning with Prototypes (IRFUP), for incremental learning using a modified Random Forest (RF) model in unsupervised scenarios. Importantly, by enhancing the RF dissimilarity and customizing a clustering algorithm, this research explores the potential utility of our proposed models in the unwell-studied field of unsupervised incremental learning.**

***Keywords*—Incremental learning, Unsupervised learning, Random forest**

I. Introduction

Incremental learning mainly discusses developing models that can continuously acquire new knowledge without completely forgetting past tasks. To ensure that models can address new incoming data, one may retrain models on both previous data and newly collected data. However, the model retraining process is much more time-consuming than updating an original model, and significant space may be needed to store all the previous training data. On the other hand, obtaining all the required labels can be challenging in some real-world scenarios. As a result, we aim to develop unsupervised incremental learning models to address these concerns.

Random Forest (RF) is a well-known machine learning technique where random sampling and random selection of features contribute to good results in classification. It is generally accepted that RF [1] was initially used for supervised problems. However, T. Shi and S. Horvath [5] first proposed the use of RF to construct data dissimilarity and introduced the concept of using RF for unsupervised clustering problems. Clustering with RF has numerous benefits, such as its ability to efficiently handle mixed variable types of data and its robustness to outlier observations. Despite the fact that there are many merits of using RF for unsupervised learning, the number of studies applying unsupervised RF models in incremental settings is limited. Therefore, we propose an Incremental Random Forest for Unsupervised Learning with Prototypes (IRFUP) to explore the potential utility of RF models in an incremental unsupervised setting.

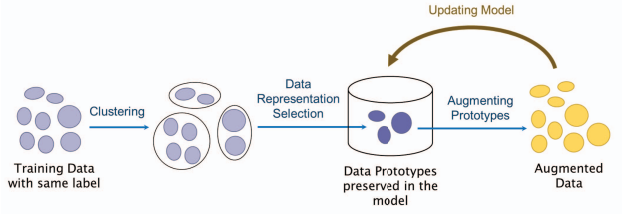

Fig. 1. An illustration of prototype construction techniques.

II. Proposed Incremental RF Model

In order to exploit the RF model in an incremental learning scenario, we utilize the RF model proposed by [2]. The specific RF model permits the decision trees based on the axis-aligned minimum bounding box to dynamically adjust with new data. However, this RF model retains all training data for decision tree updates. To tackle this issue, we propose a new approach for constructing data prototypes to preserve knowledge from training data instead of storing all training data in RF.

More precisely, we generate prototypes for data that share the same labels and then augment them to imitate authentic data. Our approach involves: (1) Clustering Data: Initially, we employ a hierarchical algorithm to partition the data into k clusters. (2) Prototype construction: we calculate the average attribute values and record the boundary conditions (B_{max} and B_{min}), representing the maximum and minimum attribute values. These values contribute to enhance the performance of data simulation. (3) Augmentation: Utilizing the prototype information, we introduce noise to the attribute means. This allows us to create simulated data closely resembling authentic samples. When there is a need to update the model with old data, Gaussian noise is generated based on the range determined by B_{max} and B_{min}, ensuring model accuracy and adaptability over time. This process is illustrated in Figure 1.

We propose extending incremental RF models to unsupervised scenarios by maintaining the data dissimilarity matrix constructed by RF. This eliminates the need to rebuild the

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00137

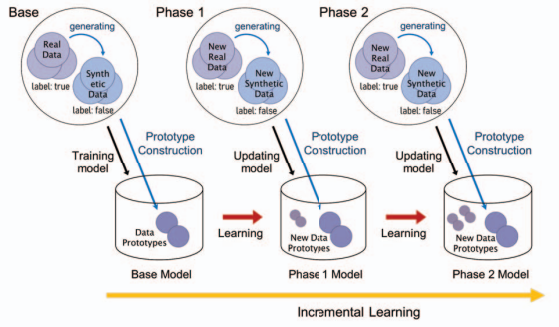

Fig. 2. An illustration of a 2-phase unsupervised incremental learning model.

TABLE I
PERFORMANCE COMPARISON: WITH VERSUS WITHOUT PROTOTYPE CONSTRUCTION TECHNIQUES.

dataset	algorithm	base	phase1	phase2	phase3
s3	*All training data*	*0.886*	*0.904*	*0.868*	*0.755*
	IRFUP	*0.886*	*0.904*	*0.870*	*0.763*
ecoli	*All Training data*	*0.960*	*0.762*	*0.612*	*0.584*
	IRFUP	*0.960*	*0.759*	*0.670*	*0.615*
aggrega-tion	*All training data*	*1*	*0.868*	*0.808*	*0.816*
	IRFUP	*1*	*0.868*	*0.841*	*0.801*

similarity matrix from scratch or retrain the RF model. Following the RF model construction method introduced by [5], we generate synthetic data by randomly sampling from the univariate distributions of the real data, disrupting attribute dependencies. We label synthetic data as false and real data as true, then train a standard RF model. Using the probability that two data points fall on the same node, we construct a similarity matrix. The dissimilarity between data points is then computed as $\sqrt{1 - similarity}$.

In the incremental phase, we generate new synthetic data from new real data and k prototypes for both synthetic and real data separately. This results in a total of $k \cdot (i+1) \cdot 2$ preserved prototypes in the i-th incremental phase. We start at the root node. If all the data falls within the defined boundaries, set by the maximum N_{max} and minimum N_{min} values, we split it into two groups according to the node's decision value. However, if the data falls outside the boundaries, we add a parent decision node and a child leaf node for this case. Additionally, if the data reaches a leaf node that does not contain split information and all attributes fall outside the boundaries, we train a new decision tree using the aforementioned augmented prototype data and new data. This incremental RF enables us to expand the similarity matrix using only the RF model from the previous phase without relying on any old training data. Lastly, we proceed with a clustering algorithm, such as k-means or k-medoid, by using the RF-based distance. The whole process is illustrated in Figure 2.

III. Experimental Results and Analysis

We acquired three datasets (s3, ecoli, and aggregation) from the UCI Machine Learning Repository, splitting them

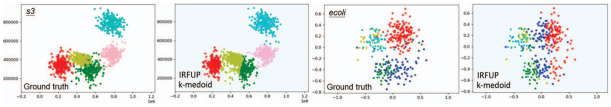

Fig. 3. The visualization results from a 3-phase unsupervised incremental model on datasets s3 and ecoli

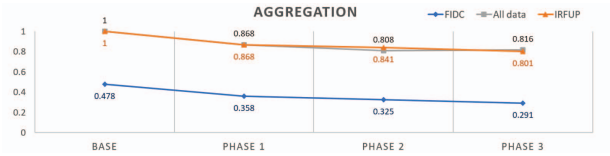

Fig. 4. The plot involves three outcomes: two unsupervised incremental models (with and without prototypes) and an incremental clustering algorithm.

for incremental learning. The base model was trained on a larger set than the incremental phase. Using the f-measure, we evaluated models with and without prototype construction techniques. Table I displays competitive clustering results, showing our models perform well even without relying on past training data. We adopt a three-phase setting based on class numbers, with the flexibility to add more phases as required.

We constructed a similarity matrix through an incremental RF model and applied it to the k-medoid algorithm [3]. The visualization results are shown in Figure 3. Our IRFUP models incrementally identify clusters without relying on past data. While some boundaries might not align precisely with the truth, they consistently identify most of the data.

We compared our models to the incremental density-based clustering algorithm (FIDC) [4] using a time window of 40, equivalent to an incremental data volume of 40. Our models surpassed FIDC (refer to Figure 4) in performance. Despite longer training times required for prototype modeling, our proposed prototype-based methods consistently achieve higher f-measure performance.

IV. Conclusion and Future Work

This study introduces IRFUP, an RF model for incremental unsupervised learning. It extends the traditional RF model to incrementally construct data similarity incrementally, integrating unsupervised learning into incremental scenarios. Experiments demonstrate the effectiveness of the proposed IRFUP models. In future, we plan to extend the IRFUP method to more unsupervised incremental learning models.

References

[1] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, October 2001.
[2] L. Hu C. Hu, Y. Chen and X. Peng, "A novel random forests based class incremental learning method for activity recognition," Pattern Recognit., vol. 78, pp. 277–290, June 2018.
[3] L. Kaufman and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, John Wiley & Sons, 2009.
[4] S. Laohakiat, V. Sa-Ing, "An incremental density-based clustering framework using fuzzy local clustering," Information Sciences, vol. 547, pp. 404–426, 2021.
[5] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," Comput. Graph. Stat., vol. 15, no. 1, pp. 118–138, 2006.