INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Nidhi Ahlawat
Roll Number	:	186101005
Programme of Study	:	Ph.D.

Thesis Title:
Isolation Forest Based Efficient Unsupervised Machine Learning Algorithms

Name of Thesis Supervisor(s)	:	Dr. Amit Awekar
Thesis Submitted to the Department/ Center	:	CSE
Date of completion of Thesis Viva-Voce Exam	:	11-03-2025
Key words for description of Thesis Work	:	Isolation Forest, Efficient Machine Learning, Unsupervised Learning, Incremental Algorithms, Scalable Algorithms, Clustering, Anomaly Detection

SHORT ABSTRACT

Many ML algorithms have common redundancies that make them impractical for large datasets. The overarching goal of this thesis is to prune the redundant computations with minimal loss in the quality of the downstream tasks. This dissertation focuses on three unsupervised machine-learning tasks: clustering, anomaly detection, and model update. We utilize the Isolation Forest data structure as a tool to improve efficiency for all three tasks. This data structure was initially developed to perform anomaly detection task in an unsupervised manner. Specifically, we focus on the following three scenarios: 1. When an application needs all-pair distances: How to compute all-pair distances faster by optimizing the order of distance computation? 2. When an application needs only a subset of all-pair distances: How do we quickly identify the required subset of all pairs? 3. When new data causes concept drift: How to update the model quickly? For the first scenario, we develop an algorithm: fast MBD (*fMBD*) that computes all-pair distances with up to 5X speed-up. Our *fMBD* algorithm has no approximation or heuristic, and it computes the exact distance for each data point pair. We demonstrate the effectiveness of the *fMBD* algorithm with clustering and anomaly detection applications. For the second scenario, we develop a scalable MBScan (*sMBScan*) clustering algorithm that selectively computes distances between data point pairs. Our algorithm achieves up to 53X speed up with up to 96% reduction in the memory footprint and no loss in the clustering quality. For the third scenario, we develop an Incremental Isolation Forest (*I²Forest*) that quickly updates the Isolation Forest data structure in response to the arrival of new data. *I²Forest* is particularly effective when the new data causes concept drift. *I²Forest* has significantly lower training time than retraining the model from scratch. *I²Forest* also performs better than other incremental approaches for model update.