Violation-based Feature Selection for Isolation Forest

D. M. Dissanayake[1], R. Navarathna[2], S. Viswakula[1]

[1]*Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka*
[2]*OCTAVE, John Keells Group, Sri Lanka*

Anomaly detection is crucial in sectors like finance and healthcare to identify deviations from normal behavior. The Isolation Forest (ISF) algorithm, introduced by Liu et al. (2008), is effective but has limitations, such as bias towards correlated variables and suboptimal results with irrelevant features. This study introduces the Violation Features Based Isolation Forest (VFIF) algorithm, which improves on ISF by evaluating the features used to build each tree and prioritizing those trees with features that violate strong patterns in the data. Unlike ISF, which aggregates all trees, VFIF strategically selects a subset of trees for anomaly scoring. Our empirical evaluation, using five real-world datasets from the UCI Machine Learning Repository, involved selecting subsets of trees at α = 30%, 50%, and 75%. Here, α represents the subset of trees from the full set of trees; for example, α = 30% means selecting 30% of trees from the total tree count. α can take any value greater than 0%, up to 100%, allowing for flexible selection of the subset size. Results show that VFIF outperforms ISF in most cases regarding Area Under the Curve (AUC), True Positive Rate (TPR), and False Positive Rate (FPR). Users can tune the α parameter to optimize their desired metric, highlighting VFIF's flexibility and enhanced performance over ISF. Tuning α involves experimenting with different subset proportions to balance computational efficiency and detection accuracy, enabling users to find the optimal trade-off for their specific application. Experimental results demonstrate that VFIF offers a novel and effective approach to anomaly detection by strategically selecting and prioritizing trees, and improving metrics across diverse datasets. Future work will focus on refining the selection criteria and incorporating additional parameters to enhance VFIF's performance through extensive simulation studies.

Keywords: *Isolation Forest, Anomaly Detection, Rule Violation*