Abstract: An Incremental Optimization Model for Balancing Accuracy, Tree Size and Learning Time of Very Fast Decision Tree

Yang Hang, Simon Fong
Department of Computer and Information Science, University of Macau, Macau SAR
{ya97404, ccfong}@umac.mo

Abstract

A classical model for real-time data stream mining called Very Fast Decision Tree (VFDT) was originally proposed. Its performance however varies under imperfections in data streams, such as noise and imbalanced class distribution. They inflate the tree size enormously; degrade the accuracy and prolong the learning time for the tree induction process. Traditional sampling techniques and post-pruning that operate in batch mode may be impractical for a non-stopping data stream. A corrective method for the adverse effects is proposed in this paper to incrementally optimize the VFDT model for data stream classification even in imperfect data. The improved model is hence called the Incrementally Optimized Very Fast Decision Tree (I-OVFDT) and it balances performance in relation to prediction accuracy, tree size and learning time. The new model diminishes error and tree size dynamically when it runs on the fly. The incremental optimization model is embedded in the tree node-splitting mechanism; therefore the optimization can be flexibly integrated into those existing VFDT-extended algorithms based on Hoeffding bound in node splitting. The experimental results show that I-OVFDT has higher accuracy and more compact tree size than other existing data stream classification methods.