Improved feature selection and stream traffic classification based on machine learning in software-defined networks

ABSTRACT

Traffic classification (TC) in software-defined networks (SDN) using machine learning (ML) appears to be a viable option for improving network management. TC improves SDN operability, while SDN speeds up the feature selection (FS) process, especially when ML is used as a classification mechanism to extract measurements and related information from incoming data to the SDN controller. Despite these advantages, there is still a lack of adequate support for TC and FS tasks due to the frequent similarity of traffic profiles, making classification difficult. Furthermore, when combined with TC, stream learning (SL) poses numerous challenges. As a result, robust statistical flow features are needed to reduce the overhead of the SDN control plane. As a result, these statistical flow features could extract online features, handle concept drift, and process an infinite data stream using limited resources (time and memory). This paper aims to improve the overall performance of TC using the SL technique to select relevant FS to alleviate load from the SDN control plane by doing the following. First, an FS mechanism called Boruta is proposed. Second, we propose three streaming-based TC methods for SDN: Hoeffding adaptive trees (HAT), adaptive random forest (ARF), and k-nearest neighbour with adaptive sliding window detector (KNN-ADWIN). These techniques can dynamically handle the concept drift and solve the problem of memory and time consumption, lowering the overhead of the SDN controller. Third, real and synthetic traffic traces are used to evaluate the proposed FS and streaming TC performance. According to simulation results, the Boruta FS technique can achieve up to 95% average accuracy and up to 87% average per application in terms of precision, recall, and f-score, outperforming other works in the literature. Furthermore, results for SL techniques show that the proposed methods can maintain up to 85% average accuracy, 78% kappa, and average rates of 62-88% in precision, recall, and f-score. In addition, when compared to ART and KNN-ADWIN, the HAT consumes less time and memory (15s and 105KB, respectively).