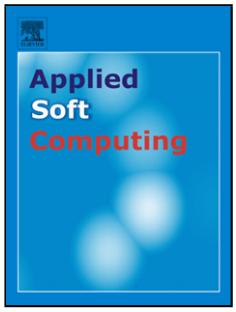

Title: Online Ensemble Learning with Abstaining Classifiers for Drifting and Noisy Data Streams

Author: Bartosz Krawczyk Alberto Cano

PII: S1568-4946(17)30723-8
DOI: https://doi.org/doi:10.1016/j.asoc.2017.12.008
Reference: ASOC 4604

To appear in: *Applied Soft Computing*

Received date: 28-2-2017
Revised date: 2-12-2017
Accepted date: 5-12-2017

Please cite this article as: Bartosz Krawczyk, Alberto Cano, Online Ensemble Learning with Abstaining Classifiers for Drifting and Noisy Data Streams, *<![CDATA[Applied Soft Computing Journal]]>* (2017), https://doi.org/10.1016/j.asoc.2017.12.008

Online Ensemble Learning with Abstaining Classifiers for Drifting and Noisy Data Streams

Bartosz Krawczyk[a,*], Alberto Cano[a]

[a]*Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA*

Abstract

Mining data streams is among most vital contemporary topics in machine learning. Such scenario requires adaptive algorithms that are able to process constantly arriving instances, adapt to potential changes in data, use limited computational resources, as well as be robust to any atypical events that may appear. Ensemble learning has proven itself to be an effective solution, as combining learners leads to an improved predictive power, more flexible drift handling, as well as ease of being implemented in high-performance computing environments. In this paper, we propose an enhancement of popular online ensembles by augmenting them with abstaining option. Instead of relying on a traditional voting, classifiers are allowed to abstain from contributing to the final decision. Their confidence level is being monitored for each incoming instance and only learners that exceed certain threshold are selected. We introduce a dynamic and self-adapting threshold that is able to adapt to changes in the data stream, by monitoring outputs of the ensemble and allowing to exploit underlying diversity in order to efficiently anticipate drifts. Additionally, we show that forcing uncertain classifiers to abstain from making a prediction is especially useful for noisy data streams. Our proposal is a lightweight enhancement that can be applied to any online ensemble method, improving its robustness to drifts and noise. Thorough experimental analysis validated through statistical tests proves the usefulness of the proposed approach.

Keywords: Machine learning, Data stream mining, Concept drift, Ensemble learning, Abstaining classifier, Diversity

1. Introduction

In the context of the big data era, information systems produce a continuous flow of massive collections of data surpassing storage and computation capabilities of traditional knowledge extraction methods. Big data is characterized by its properties which include volume, velocity, variety, veracity, variability, visualization, and value. In recent years, researchers have mainly focused on the scalability of data mining algorithms to address the ever increasing data volume [1]. However, one cannot ignore the importance of the remaining ones, especially velocity and variability. Velocity is critical in real-time decision systems, where new instances are continuously evaluated and fast decision must be outputted under

*Corresponding author

Email addresses: bkrawczyk@vcu.edu (Bartosz Krawczyk), acano@vcu.edu (Alberto Cano)