Online entropy-based discretization for data streaming classification

S. Ramírez-Gallego, S. García, F. Herrera

PII: S0167-739X(17)32581-5
DOI: https://doi.org/10.1016/j.future.2018.03.008
Reference: FUTURE 4020

To appear in: *Future Generation Computer Systems*

Received date : 8 November 2017
Revised date : 19 January 2018
Accepted date : 3 March 2018

Please cite this article as: S. Ramírez-Gallego, S. García, F. Herrera, Online entropy-based discretization for data streaming classification, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.03.008

Online Entropy-Based Discretization for Data Streaming Classification

S. Ramírez-Gallego[a,*], S. García[a], F. Herrera[a,b]

[a]Department of Computer Science and Artificial Intelligence, CITIC-UGR, University of Granada, 18071 Granada, Spain.
[b]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia.

Abstract

Data quality is deemed as determinant in the knowledge extraction process. Low-quality data normally imply low-quality models and decisions. Discretization, as part of data preprocessing, is considered one of the most relevant techniques for improving data quality.

In static discretization, output intervals are generated at once, and maintained along the whole process. However, many contemporary problems demands rapid approaches capable of self-adapting their discretization schemes to an ever-changing nature. Other major issues for stream-based discretization such as interval definition, labeling or how is implemented the interaction between learning and discretization components are also discussed in this paper.

In order to address all the aforementioned problems, we propose a novel, online and self-adaptive discretization solution for streaming classification which aims at reducing the negative impact of fluctuations in evolving intervals. Experiments with a long list of standard streaming datasets and discretizers have demonstrated that our proposal performs significantly more accurately than the other alternatives. In addition, our scheme is able to leverage from class information without incurring in an overweight cost, being ranked as one of the most rapid supervised options.

Keywords: Data stream, Concept drift, Data preprocessing, Data reduction, Discretization, Online learning

1. Introduction

Learning models and subsequent results are highly dependent on the quality of input data. Incorrect decisions can be taken if raw data are not properly cleaned and structured. The data preprocessing task [1, 2] is an essential step

[*] Corresponding author
Email addresses: sramirez@decsai.ugr.es (S. Ramírez-Gallego), salvagl@decsai.ugr.es (S. García), herrera@decsai.ugr.es (F. Herrera)