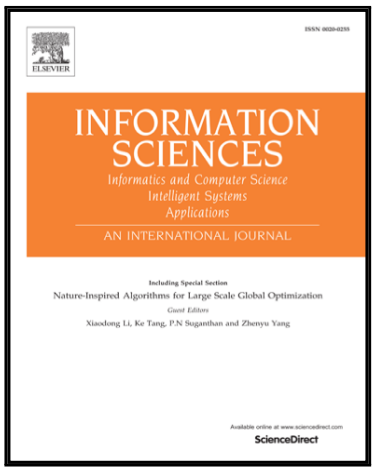

How to Adjust an Ensemble Size in Stream Data Mining?

Lena Pietruczuk, Leszek Rutkowski, Maciej Jaworski, Piotr Duda

PII: S0020-0255(16)31344-5
DOI: 10.1016/j.ins.2016.10.028
Reference: INS 12587

To appear in: *Information Sciences*

Received date: 9 March 2016
Revised date: 22 September 2016
Accepted date: 18 October 2016

Please cite this article as: Lena Pietruczuk, Leszek Rutkowski, Maciej Jaworski, Piotr Duda, How to Adjust an Ensemble Size in Stream Data Mining?, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.10.028

How to Adjust an Ensemble Size in Stream Data Mining?

Lena Pietruczuk[a], Leszek Rutkowski[a], Maciej Jaworski[a], Piotr Duda[a]

[a]*Institute of Computational Intelligence, Czestochowa University of Technology, Al. Armii Krajowej 36, 42-200 Czestochowa, Poland*

Abstract

In this paper we propose a new approach for designing an ensemble applied to stream data classification. Our approach is supported by two theorems showing how to decide whether a new component should be added to the ensemble or not, based on the assumption that such an action should increase the accuracy of the ensemble not only for the current portion of observations but also for the whole (infinite) data stream. The conclusions of these theorems hold with a certain probability (confidence) set by the user. Through computer simulations, among others, we show that decreasing the confidence that decision based on the finite portion of the stream is the same as based on the whole (infinite) data stream only slightly improves the accuracy at the expense of significant memory consumption. Moreover, we will introduce a novel procedure of weighting ensemble components, i.e. decision trees, by assigning a weight to each leaf of the tree. In previous approaches a weight was assigned to the whole ensemble component. The new approach is based on the observation that probability of the correct tree outcome is different in various tree sections.

Keywords: stream data, data mining, classification, ensemble methods

1. Introduction

Data mining is a subject widely investigated by many scientists all over the world. However, most of developed methods were designed to analyze databases of a finite size. Nowadays the technological development allows to obtain and process huge amount of various data. The number of data elements that is available can not be established in advance. Therefore a new field of study emerged called stream data mining [1] [6] [12] [16] [24] [25] [28] [29] [33] [45] [47]. The problem is to obtain methods that can fast and efficiently extract information from constantly incoming data. Because of the high rate and great complexity of incoming data, previously created algorithms are not applicable. Depending on the particular problem tasks can be very different, e.g. classification [23] [31] [32] [37] [42], clustering [4] [19], [46], regression [35], summarization [10], association rule learning [9] and anomaly detection [11]. In this paper we focus on the first task, i.e. the problem of assigning new data element to one of predefined classes. In the literature there exists a wide range of algorithms designed for static data classification. Most commonly known methods are based on decision trees, neural networks, K-NN, Naive Bayes classifier, support vector machines and ensemble methods. These methods are inspirations for new algorithms developed and dedicated to stream data mining, like the McDiarmid Decision Tree and Gaussian Decision Tree [38] [39] [40] [41], the Very Fast Decision Tree (VFDT) [13] and the Concept-adapting Very Fast Decision Tree (CVFDT) [21], On Demand Classification