Rule Learning from Data Streams: an overview

Jesús S. Aguilar–Ruiz

School of Engineering
Pablo de Olavide University
Seville, Spain

Abstract. Classification is a very well studied task in data mining. In the last years, important works have been published to scale up classification algorithms in order to handle large datasets. However, due to the high rate of streams of data, a number of emerging applications are demanding new approaches. Rule learning is an efficient alternative to address non–stationary environments. The talk presents an overview of rule–based learning algorithms for data streams and emphasizes some important aspects of these techniques.

Keywords: Data Streams, Rule–based learning.

1 Introduction

The advances in hardware technology have paved the way for the development of algorithms that can process the real–time information at a rapid rate. Streams of data grow at an unlimited rate and traditional data mining algorithms cannot process them efficiently. In spite of the great increase of storage capacity, it is not even enough for hundreds or thousands of instances arriving per second. Nowadays, typical problems such as clustering, classification, frequent pattern mining, change detection or dimensionality reduction are being reconsidered in the realm of data streams. What was initially finite data is now infinite data, thus giving rise to many challenges in machine learning, data mining and statistics.

Classification and rule learning are important, well studied tasks in machine learning and data mining. Classification methods represent the set of supervised learning techniques where a target categorical variable is predicted based on a set of numerical or categorical input variables. A variety of methods such as decision trees, rule based methods, and neural networks are used for the classification problem. Most of these techniques have been designed to build classification models from static data sets, where several passes over the stored data are possible.

In order to classify and model large–scale databases, important works have been recently addressed to scale up inductive classifiers and learning algorithms. In environments where high–rate streams of detailed data are constantly generated, memory and time limitations make multi pass scalable algorithms unfeasible. Also, real–world data streams are not generated in stationary environments, requiring incremental learning approaches to track trends and adapt to changes in the target concept.

Furthermore, the classification process may require simultaneous model construction and testing in an environment which constantly evolves over time. However, within incremental learning, a whole training set is not available a priori as examples arrives over time, normally one at a time t and not time dependent necessarily (e.g., time series). Despite online learning systems continuously review, update, and improve the model, not every online system is based on an incremental approach.

2 Some aspects of learning from data streams

Formally, a data stream D can be defined as a sequence of examples (also called transactions or instances), $D=(e_1,e_2,\ldots,e_i,\ldots)$, where e_i is the i-th arrived example. To process and mine data streams, different window models are often used. A window is a subsequence between the i-th and j-th arrived examples, denoted as $W[i,j]=(e_i,e_{i+1},\ldots,e_j)$, $i<j$. There are three common models: