Density estimation on data streams : an application to Change Detection

Alexis Bondu*, Benoît Grossin*,
Marie-Luce Picard*

*EDF R&D ICAME/SOAD, 1 avenue du Général de Gaulle, 92140 Clamart.
firstname.name@edf.fr

Abstract. In recent years, the amount of data to process has increased in many application areas such as network monitoring, web click and sensor data analysis. Data stream mining answers to the challenge of massive data processing, this paradigm allows for treating pieces of data on the fly and overcoming data storage. The detection of changes in a data stream distribution is an important issue. This article proposes a new schema of change detection : i) the summarization of the input data stream by a set of micro-clusters; ii) the estimate of the data stream distribution exploiting micro-clusters; iii) the estimate of the divergence between the current estimated distribution and a reference distribution; iv) diagnostic step through the contribution of each predictive variable to the overall divergence between both distributions. Our schema of change detection is applied and evaluated on artificial data streams.

1 Introduction

In recent years, the amount of data to process has increased in many application areas such as network flows, web click and sensor data analysis. Data stream mining indicates algorithms which process tuples [1] on the fly : when they are emitted, without storing them. The processing of tuples should be as fast as possible which allows for managing high rate data streams. An important issue in processing data streams is detecting changes in underlying distribution that is generated by tuples. The designing of change detection schemes which are general, scalable and statistically relevant is a great challenge.

A change in the underlying distribution can be interpreted into different ways : i) the observed phenomenon is naturally drifting due to a change in some hidden *context* (Widmer and Kubat, 1996) which is not explicitly given by predictive features; ii) an abnormal change is taking place in the observed system. Distinguish the two cases is a very difficult issue which requires expertise on the application. In this article we assume an expert, who well knows the observed data stream, may rule on the interpretation of detected changes.

An overview of the main change detection approaches is given by A. Dries (Dries and Rückert, 2009) : *Change detection in the distribution of tuples can be considered as a statistical hypothesis test which involves two samples of multidimensional tuples. Such problems are studied in the statistical literature. The Wald-Wolfowitz and Smirnov tests was generalized*

1. The term “tuple” refers to a piece of data which is emitted from the input stream.