Detecting Anomalies in Data Streams using Statecharts

Vasile-Marian Scuturici[1], Dan-Mircea Suciu[2], Romain Vuillemot[1], Aris Ouksel[3], Lionel Brunie[1]

[1] INSA Lyon
[2] Babes-Bolyai University, Cluj-Napoca
[3] Northwestern University, Illinois

Abstract. The environment around us is progressively equipped with various sensors, producing data continuously. The applications using these data face many challenges, such as data stream integration over an attribute (such as time) and knowledge extraction from raw data. In this paper we propose one approach to face those two challenges. First, data streams integration is performed using statecharts which represents a resume of data produced by the corresponding data producer. Second, we detect anomalous events over temporal relations among statecharts. We describe our approach in a demonstration scenario, that is using a visual tool called *Patternator*.

1 Introduction

The physical space surrounding us is progressively equipped with different sensors, giving us a digital view/projection of the real world. These sensors produce raw data streams, such as temperature values, images, and badge readings. Still, it remains difficult for humans to perceive interesting information from those raw data which are increasing in frequency and precision. In this demonstration we propose a mechanism to describe the behavior of data streams based on statecharts. We use the temporal attribute associated with a stream in order to detect relations between statecharts corresponding to non-homogenous datasources. The events non-holding these relations are considered as anomalies.

2 Statecharts modeling datasources

In our vision, a **datasource** is an abstraction for a fountain/fabric of data entities in the environment. Examples of datasources are the values read by a temperature sensor or the events produced by a badge reader. We associate the entities produces by a datasource with the datasource itself. These entities share the same structure, associated with the datasource.

The time plays a role in the description of a datasource *DS*. At every (discrete) moment *t*, *DS* is in a state defined by the produced entity at the instant *t*. The entities produced by a datasource, associated with a timestamp, form a **data stream**.

A *datasource DS* at a time *t* is a sequence of timestamped data entities sharing the same structure (header). Each data entity has values from the cartesian product $R^+ \times A^1 \times A^2 \times ... \times A^n$, where R^+ denotes the set of all positive real numbers (the time