

Real World Issues in Stream Classification

Vincent Lemaire¹

Orange Labs, France
vincent.lemaire@orange.com

Abstract. Statistical learning provides numerous algorithms to build predictive models on past observations. These techniques proved their ability to deal with large scale realistic problems. However, new domains generate more and more data. This large amount of data (the buzz “big data”) can be dealt with using batch algorithms (parallelized . . .) if the paradigm to store the data is realistic. But sometimes data are only visible once and need to be processed sequentially. These volatile data, known as data streams, come from telecommunication network management, social network, web mining, to name a few. The challenge is to build new algorithms able to learn under these constraints. The aim of this presentation will be to present several studies and research topics at Orange¹ focusing on “supervised classification in data streams”, with the idea to stimulate a discussion on “the real issues”.

¹ See <http://www.orange.com/en/home>