Feedback - Study and Improvement of the Random Forest of the Mahout library in the context of marketing data of Orange

C. Thao[*,**], N. Voisine[*], V. Lemaire[*], R. Trinquart[*]

[*] Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion, France
[**]Predicsis, 5 rue de Broglie, 22300 Lannion, France

Abstract. In the realm of Big Data systems, Hadoop has emerged as one of the most popular systems and a very diverse ecosystem has grown around it, meeting all kinds of functional and technical needs. One niche that should have been a place of choice in this ecosystem is data analytics: first because getting value out of large datasets requires efficient Machine Learning (ML) algorithms, second because large clusters with abundant CPUs resources seem like appropriate playfields for ML algorithms which are often very resource-intensive computing tasks. Unfortunately among the myriad of open source projects, there are very few data analytics tools that have been ported to the Hadoop framework. Apache Mahout stands out among those rare initiatives: this project is mainly known for its recommendation application, but it also offers a warehouse of ML algorithms, advertised to run on Map/Reduce. We did investigate the twenty algorithms proposed within Mahout and in this report we focus on the most promising one: the Random Forest implementation. Relying on extensive tests, including specific marketing data from Orange, we provide an in-depth feedback on the use of this tool, both from a practical and theoretical perspective, and we suggest several improvements.

1 Introduction

The decreasing cost of data storage has led to the accumulation of large and complex datasets, which are widely seen as new opportunities for business. Orange - a multinational telecommunications corporation - has to analyze the data of its network to improve profitability and create new services. To give a sense of scale, in order to increase customer satisfaction on services, Orange has to analyze Quality of Services (QoS) and Quality of Experience (QoE) indicators for its 150 million of mobile customers. Those QoS and QoE indicators result from the combination of different data sources (network probe, SI). The main purpose consists in real time detection or prediction of QoS or QoE. This would allow Orange either to improve quality of network or to provide new services based on QoE. Therefore applying data mining techniques to these vast amounts of data is crucial. This raises numerous issues such as scalability of data mining algorithms, automation of the data mining process and control of over-fitting.