POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXX cycle

Supervisor

Prof. Elena Baralis

Classification algorithms for big data, with applications on urban safety

PhD Candidate: *Luca Venturini*

1. Introduction

Big data frameworks offer scalable algorithms to extract information from raw data, but i) strive to deal with large-domain categorical features and ii) often require a sophisticated fine-tuning and a detailed knowledge of machine learning algorithms.

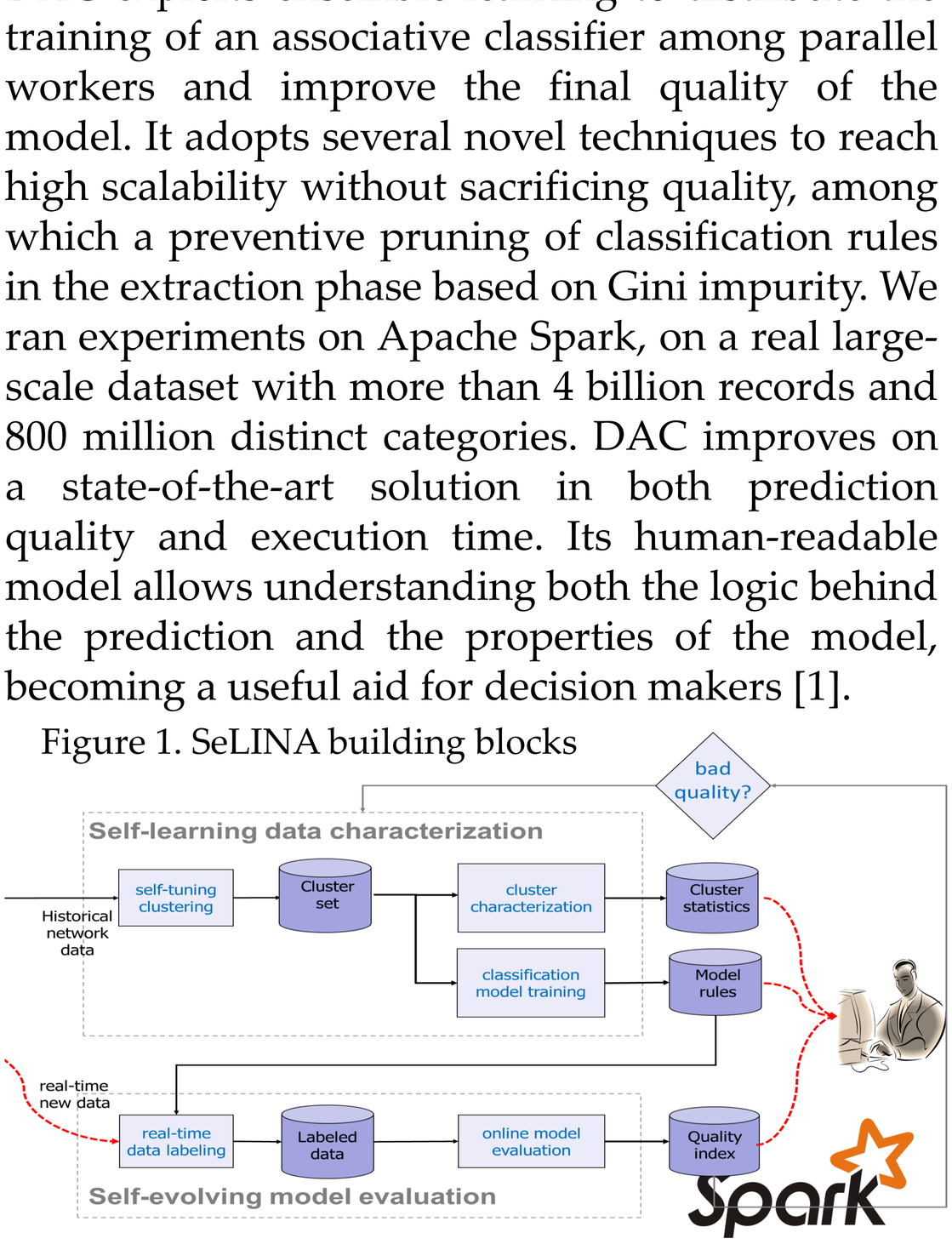
2. Aim

The aim of this thesis is to scale machine learning algorithms beyond current limits, make a reality check of Big Data machine learning frameworks and validate the effectiveness of a data science process in improving urban safety.

3. Scaling associative classification to very large datasets

DAC is a Distributed Associative Classifier. DAC exploits ensemble learning to distribute the training of an associative classifier among parallel workers and improve the final quality of the model. It adopts several novel techniques to reach high scalability without sacrificing quality, among which a preventive pruning of classification rules in the extraction phase based on Gini impurity. We ran experiments on Apache Spark, on a real large-scale dataset with more than 4 billion records and 800 million distinct categories. DAC improves on a state-of-the-art solution in both prediction quality and execution time. Its human-readable model allows understanding both the logic behind the prediction and the properties of the model, becoming a useful aid for decision makers [1].

Figure 1. SeLINA building blocks



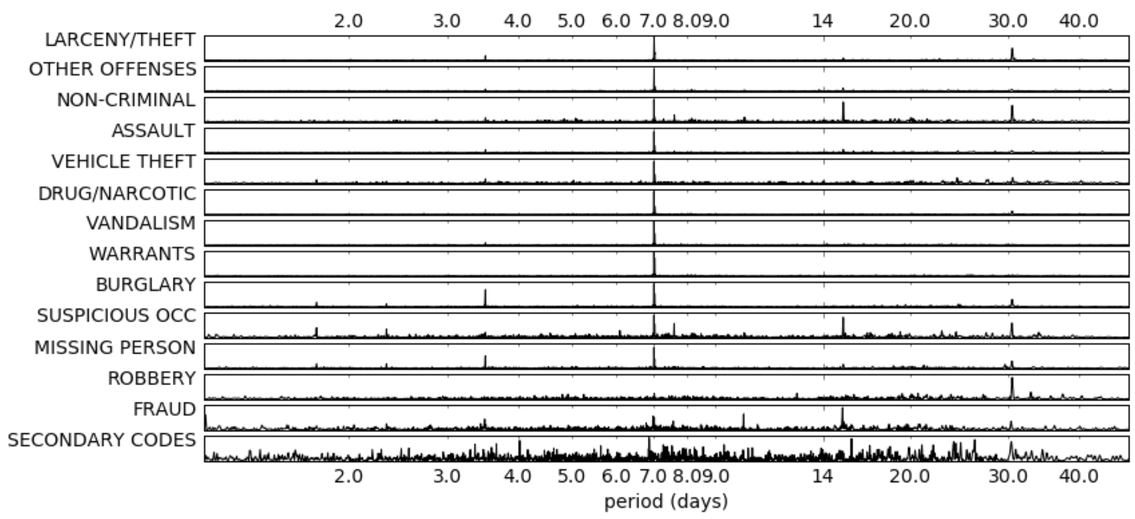
4. Real life use cases of a Big Data machine learning pipeline

SeLINA (Self-Learning Insightful Network Analyzer) is a generic, self-tuning, simple tool to extract knowledge from network traffic measurements. SeLINA provides self-learning capabilities to state-of-the-art scalable approaches and off-loads the network expert from parameter tuning. With minimal user intervention, it supports domain experts in extracting actionable knowledge and detect changes in the data [2].

5. Data Science for urban safety

The research of this PhD assessed the application of data analytics process to urban safety. Spatio-temporal patterns were found in the crimes of San Francisco and Stockholm, by applying spectral analysis and spatial clustering [3]. An integrated data mining and Business Intelligence architecture for the analysis of non-emergency open data has been made to generate informative dashboards and alert the municipality actors of a Smart City. This methodology can improve our understanding of the dynamics of crime and be exploited to design effective policing policies.

Figure 2. Periodograms of crimes in San Francisco



6. References

[1] Venturini, Luca; Garza, Paolo; Apiletti, Daniele (2016) BAC: A bagged associative classifier for big data frameworks. In: BigDap 2016, Prague, Czech Republic, 28-8-2016

[2] Apiletti, Daniele; Baralis, Elena; Cerquitelli, Tania; Garza, Paolo; Giordano, Danilo; Mellia, Marco; Venturini, Luca (2016) SeLINA: a Self-Learning Insightful Network Analyzer. In: IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, vol. 13 n. 3

[3] Venturini, Luca; Baralis, Elena (2016) A spectral analysis of crimes in San Francisco. In: UrbanGIS 16, Burlingame, California (USA), 31-10-2016.