EURA NOVA

FACULTIES OF SCIENCE AND ENGINEERING

Master Thesis Offers 2016

For the latest information, please visit http://research.euranova.eu.
14th April 2016

CONTENTS

EURA NOVA

INTRODUCTION

EURA NOVA is a Belgian company founded in September 2008. Our mission is simple: ``Being a technological incubator focusing on the pragmatic use of knowledge''. Our research and engineering activities are linked to technological directions and industrial opportunities.

Please visit our website http://euranova.eu for more information on our activities.

OUR MASTER THESIS OFFERS

This document presents Master Thesis offers supervised by our Research and Development department. Each of these projects represents a concrete opportunity to generate impact for EURA NOVA.

The student will work in close cooperation with the engineering team and will communicate the advance through the in-house EURA NOVA knowledge management tool.

For more information on our R&D activities, please visit our departement's website at http://research.euranova.eu.

HOW TO APPLY

If you are interested in one of our offers, please contact us at career@euranova.eu.

MACHINE LEARNING & DATA SCIENCE

METRIC LEARNING FOR EFFICIENT SEARCH WITH HIGH-DIMENSIONAL HETEROGENEOUS DATA

Context: Heterogeneous/multi-modal data arise naturally in many real world use cases. For example, a blog article might be made of one or multiple photos, texts, author informations. The Million Song Dataset [1] is another example of multi-modal data. It consists of many audio features and meta-data for a million popular tracks which can be enriched with lyrics [2], genre and many other modalities.

On the other hand, information retrieval has become a challenging task within the big data setting leading to the quest for efficient yet accurate strategies to compute similarity/distance between entities.

To deal with such a context, a usual way consists in learning a similarity function (see [BHS13] for a survey), most of the time from training data composed of pairs of similar/dissimilar data. Despite a few attempts, those strategies have mainly focused on uni-modal data (e.g. text only). The objective of this internship is to propose a strategy to tackle tasks, especially search, involving heterogeneous/multi-modal data.

This internship will be jointly supervised by EURA NOVA and the Machine Learning team of the Hubert Curien laboratory in Saint-Etienne. It will take place at EURA NOVA

Business Opportunity: Having the capacity to find similar objects/products or to detect duplicates is a key feature for many e-commerce websites as well as for many on-line newspapers to prevent plagiarism for example.

Contribution:

1. Explore the state-of-the-art techniques.
2. Propose a technique to address the high-dimensional heterogeneous data metric learning task.
3. Use the Million Song Dataset to assess the efficiency of the proposed technique in term of accuracy and computation time.

[1] http://labrosa.ee.columbia.edu/millionsong/
[2] https://www.musixmatch.com/fr

DISTRIBUTED FRANK-WOLFE OPTIMIZATION FOR NUCLEAR NORM REGULARISED PROBLEMS : APPLICATION TO MATRIX COMPLETION

Context: Regularisation is a key factor of the ability for machine learning algorithms to generalise to unseen data. When the model parameters are expressed as matrices, a commonly used regulariser is the nuclear norm (also called trace norm). This regularisation technique promotes low-rank structure among the parameter matrices (see for example [JS10]).

On the other hand, the Frank-Wolfe algorithm has re-gained interest in the recent years partially due to its good scalability properties. In the case of convex nuclear norm regularised optimisation problems, the Frank-Wolfe step involves finding the top eigenvalue when projected gradient algorithms (such as SGD) involves a complete SVD decomposition. In the case of large-scale matrices this can lead to a substantial gain over the training time needed by machine learning techniques.

Business Opportunity: Recommendation systems are nowadays a well-know example of machine learning models used by companies (movie or music streaming platforms, online shopping, ...) to improve their revenue, but also the management of their knowledge base. In its simplest form, a recommendation system acts between two type of entities (namely recommending something to someone). But those problems are becoming more and more complex, involving many types of entities and leading to potentially huge-scale optimisation problems [BYG13], sometimes modeled with an underlying graph structure [KBBV14].

Contribution:

1. A state of the art of techniques to solve nuclear norm regularised problems, putting emphasis on matrix completion/factorisation techniques.
2. An implementation of some of them and a distributed Frank-Wolfe approach on a distributed processing framework like Apache Flink [3].
3. A benchmark of the implemented techniques of a large-scale dataset and use case to be defined during the internship.

[3] https://flink.apache.org/

TIME SERIES PREDICTION WITH DEEP LEARNING AND BAYESIAN METHODS

Context: Time series data represent a high volume of collected data. This type of data is particularly valuable when capturing variations of observed properties over time. It can be collected for example from digitized industrial sensors or from usage log.

Once this data is captured, it can be used to highlight correlations between the evolution of parameters over time. Furthermore, machine learning models can be trained to predict the variation of the parameters but also the variation of target variables in response to real-time variation of correlated parameters.

Although these types of analysis are widely required in numerous industries, there is still a need for an integrated tool-suite allowing to prototype and explore in this domain on large datasets.

Business Opportunity: Many critical industries (such as healthcare, electricity providers, telecom providers, metallurgy, ...) have collected production data over the past years. Once used for control and regulation, the availability and volume of the data are now sufficient to perform predictive and real time analysis, directly impacting production costs and revenue.

Contribution: The goal of the project is to

1. Provide a state-of-the art of time series analysis, including recent techniques in deep learning and bayesian methods.
2. Propose a model suitable for the analysis of the datasets to be defined in the scope of the project, including both descriptive analysis on historical data and real-time prediction over monitoring data.
3. Propose and implement the state-of-the art technique for analysis and prediction into an existing data science tool such as pandas [4].

[4] http://pandas.pydata.org/

METRIC LEARNING FOR EFFICIENT SEARCH WITH HIGH-DIMENSIONAL HETEROGENEOUS DATA

Context: Heterogeneous/multi-modal data arise naturally in many real world use cases. For example, a blog article might be made of one or multiple photos, texts, author informations. The Million Song Dataset [5] is another example of multi-modal data. It consists of many audio features and meta-data for a million popular tracks which can be enriched with lyrics [6], genre and many other modalities.

On the other hand, information retrieval has become a challenging task within the big data setting leading to the quest for efficient yet accurate strategies to compute similarity/distance between entities.

To deal with such a context, a usual way consists in learning a similarity function (see [BHS13] for a survey), most of the time from training data composed of pairs of similar/dissimilar data. Despite a few attempts, those strategies have mainly focused on uni-modal data (e.g. text only). The objective of this internship is to propose a strategy to tackle tasks, especially search, involving heterogeneous/multi-modal data.

This internship will be jointly supervised by EURA NOVA and the Machine Learning team of the Hubert Curien laboratory in Saint-Etienne. It will take place at EURA NOVA

Business Opportunity: Having the capacity to find similar objects/products or to detect duplicates is a key feature for many e-commerce websites as well as for many on-line newspapers to prevent plagiarism for example.

Contribution:

1. Explore the state-of-the-art techniques.
2. Propose a technique to address the high-dimensional heterogeneous data metric learning task.
3. Use the Million Song Dataset to assess the efficiency of the proposed technique in term of accuracy and computation time.

[5] http://labrosa.ee.columbia.edu/millionsong/
[6] https://www.musixmatch.com/fr

PARAGRAPH VECTOR REPRESENTATION APPLIED TO SEQUENTIAL DATA

Context: In these past years, the Natural Language Processing community has seen a significant interest in learning meaningful representation of words. This approach has been formalized as word vector [MH09] and more recently by the paragraph vector [LM14]. The main idea is to find a mathematical representation, as a vector, in which word like "Strong" and "Powerful" are much closer than "Paris". The method Paragraph Vector is an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. The extension to paragraphs enables to take into account all the paragraph in context of analysis and opens new doors in terms of sentiment analysis, text classification and information retrieval. Although the work presented in [LM14] focused on text representation, nothing prevents us from applying the same approach on Sequential data such as mobile positions or business process activity logs.

Business Opportunity: There are multiple applications of this approach:

1. Trajectory analysis: the trajectory mining is still an important topic in telecommunications and marketing. If we use this approach, we could potentially cluster trajectories based on their "similarity" in their new representation, but also asking for the most similar trajectory to an example. Going further, we could use this model to predict the next position of a mobile object according to its context. In this case, its context is the set of previous positions.
2. Business Process mining: We can model a business process as sequence of activity. In this way, if we can represent all business process in this learned-representation, we could cluster them, and even recommend to merge existing ones.

Contribution: The objective of this master thesis is:

1. to study the state of the art in word and vector representation
2. to adapt and re-design the model for sequential data
3. to validate the approach on a sequential data set

DISTRIBUTED DATA PROCESSING

ELASTIC STREAM PROCESSING WITH LATENCY GUARANTEES

Context: Stream processing is the next evolution in the processing of Big Data. Indeed, the last iteration of stream processing frameworks such as Apache Flink[7] has seen the ability of handling both batch and streaming data use cases on large volumes of data, fulfilling the kappa architecture.

In this context Flink is able to distribute a processing job defined by a graph of parallel operation on a cluster of computing ressources. The scheduling is however static. A dynamic adaptation of the number of parallel instances of an operator might be valuable for instance in the cases where gurantees on latency, throughput or quality of service is required. This dynamic adaptation of ressources in response to a change of a key performance indicator in distributed processing is called ``elasticity''.

Business Opportunity: The success of a service is entirely dependent on its quality. As such, it is important to determine and monitor a series of key performance indicators (such as latency and thoughput) and react accordingly in order to keep the quality of the service at an acceptable level.

Contribution: The project covers the following tasks:

1. Implement the algorithm defined in [LJK15] on Apache Flink or Apache Storm[8], using Apache Slider[9].
2. Implement a use case that will be used to benchmark the algorithm.
3. Complete the elasticity model by taking into account the scaling time to convergence.

[7] https://flink.apache.org/
[8] https://storm.apache.org/
[9] https://slider.incubator.apache.org/

REFERENCES

[BHS13] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. CoRR, abs/1306.6709, 2013.

[BYG13] Guillaume Bouchard, Dawei Yin, and Shengbo Guo. Convex collective matrix factorization. In Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, pages 144--152, 2013.

[JS10] Martin Jaggi and Marek Sulovsky. A simple algorithm for nuclear norm regularized problems. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 471--478, 2010.

[KBBV14] Vassilis Kalofolias, Xavier Bresson, Michael Bronstein, and Pierre Vandergheynst. Matrix completion on graphs. arXiv preprint arXiv:1408.1717, 2014.

[LJK15] Björn Lohrmann, Peter Janacik, and Odej Kao. Elastic stream processing with latency guarantees. In 35th IEEE International Conference on Distributed Computing Systems, ICDCS 2015, Columbus, OH, USA, June 29 - July 2, 2015, pages 399--410, 2015.

[LM14] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. CoRR, abs/1405.4053, 2014.

[MH09] Andriy Mnih and Geoffrey E. Hinton. A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Advances in Neural Information Processing Systems 21, pages 1081--1088. Curran Associates, Inc., 2009.