EURANOVA

FACULTIES OF SCIENCE AND ENGINEERING

Master Thesis Offers 2017

For the latest information, please visit http://research.euranova.eu.
3rd April 2017

CONTENTS

EURA NOVA 3

MACHINE LEARNING & DATA SCIENCE 4

Paragraph Vector Representation Applied to Sequential Data 4

Applications of Generative Adversarial Networks in class imbalance case and outlier detection problems 5

Natural Language Sentence Matching 6

DISTRIBUTED DATA PROCESSING 7

Elastic Stream Processing with Latency Guarantees 7

Graph processing-specific optimizations in stream processing Framework . 8

Cost Based Optimization of MapReduce jobs 9

References 10

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

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

APPLICATIONS OF GENERATIVE ADVERSARIAL NETWORKS IN CLASS IMBALANCE CASE AND OUTLIER DETECTION PROBLEMS

Context: Generative classifiers learn a model of the joint probability, $p(x, y)$, of the inputs x and the label y, and make their predictions by using Bayes rules to calculate $p(ylx)$, and then picking the most likely label y. Discriminative classifiers model the posterior $p(ylx)$ directly, or learn a direct map from inputs x to the class labels. A proposed framework combines the two model by training both of them simultaneously. The goal is to estimate generative models via an adversarial process. This framework called Generative Adversarial Networks [GPAM$^+$14]. It have led to significant improvements in image generation [SGZ$^+$16], video prediction [MCL15] and a number of other domains. The basic idea of GAN is to train a discriminator and a generator. The discriminator is trained to distinguish samples of a real dataset from "fake" samples produced by the generator. The generator is trained to produce samples that the discriminator cannot distinguish from real data samples. The two model can be trained simultaneously by means of back-propagation.

In this work, the goal is to explore GAN architectures away from image generation or classification tasks to a more general learning problems. The first problem is class imbalance [BPM04]. It is well known that this problem affects seriously the learning performance. Although, simple techniques such as stratified sub-simpling cloud help. In certain real world applications it still insufficient. The goal is to study the effectiveness of GANs comparing to other efficient techniques. The second problem is outliers detection [HA04]. The particularity of this problem is that it can be seen as both supervised problem and unsupervised problem. Many techniques have been proposed in both views. Our goal will be to compare the GAN performance with existing techniques.

Business Opportunity: Although the works presented in [GPAM$^+$14, SGZ$^+$16, MCL15] focused on image generation and classification, nothing prevents us from applying the same approach to a more general learning problems. In several business machine learning applications we are faced to problems such as class imbalace [BPM04] or even outlier detection [HA04] in noisy datasets. These problems are faced daily by data scientists and which can be benefits of GANs, if they proove to be efficient.

Contribution: The objective of this master thesis is:

- to study the state of the art GAN with therical framework understunding
- to adapt and re-design the model for imbalaced data set probelm
- and/or to adapt and re-design the model for outlier detection probelm
- Perform a benchmark to validate efficiency/unefficiency of the method

NATURAL LANGUAGE SENTENCE MATCHING

Context: In Natural Language Processing (NLP), Sentence Matching (SM) [WHF17] is the task of comparing two sentences and identifying the relationship between two sentences. This NLP task can be useful in various contexts. For example, in question answering and information retrieval SM can be used to evaluate the relevance of a query-answer pair. Recently Quora has released a dataset (in kaggle website) to identify similar questions in forums.

With the rebirth of neural network and Deep Learning [Gol15] as machine learning subfield, a lot of attention have been put to solve NLP task with these techniques. A major advantage here is the ability of creating a meaningful representation of words as numerical vectors called Word Embedding. This work is continuation of Eura Nova R&D department to push further the limits of Word Embedding techniques and adapt it to a difficult tasks such as sentence matching.

Business Opportunity: One of the direct applications of SM is in marketing and customer services. For example such algorithms can be useful to identify similar customer requests and/or to automate answers leading to an effective cost reduction. Another application can be in social media analysis in trend analysis and opinion mining.

Contribution: The objective of this master thesis is:

- To study the state of the art in word representation and sentence matching.
- To design and implement a solution for semantic matching.
- Validate our approach by benchmarking with other known methods.

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[1] 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[2], using Apache Slider[3].
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.

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

[2] https://storm.apache.org/

[3] https://slider.incubator.apache.org/

GRAPH PROCESSING-SPECIFIC OPTIMIZATIONS IN STREAM PROCESSING FRAMEWORK

Context: A Complex Event Processor is a system that correlates events in order to generate further events according to a certain pattern. For instance, a ”fire event” is generated by a CEP when more than 5 rising temperature events above a temperature threshold have been issued within 15 minutes.

Apache Flink is an efficient, general-purpose distributed data processing system.[4] Flink actually proposes a CEP package based on the work of [ADGI08] but it not distributed and can have memory limitations for large patterns, does not partition the events and only considers a single source of events.

The goal of this thesis project is to design and implement a fully distributed CEP on top of Flink. A higher level language will also be implemented in order to express rich and complex temporal patterns. A first internal research project has already highlighted the possible directions, based on the works of [AAB+05], [CM12] and [LIMK12]

Business Opportunity: CEP are widely used in the industry. For instance, in the banking industry for the detection of fraud, or in the telecommunications industry for the detection of CHURN. Being able to monitor large patterns and states of events allows for more complex situtations to be efficiently detected.

Contribution: The student will have to:

1. Review related work and choose a direction based on related work.
2. Implement the chosen methods in Apache Flink.
3. Implement a set of applications in Flink in order to validate and evaluate the work.

Contact: Nam-Luc Tran

[4] http://flink.incubator.apache.org

COST BASED OPTIMIZATION OF MAPREDUCE JOBS

Context: Performance of a particular job on a distributed computation platform (i.e. Spark) can vary significantly depending on the input data type and size, the design and implementation of the algorithm and the computing capability. All these factors makes it extremely difficult to predict the performance metric of a job at execution time. Standard methods such as Rule Based optimization as used in databases became unfit to such problem. To address this challenge, this project aims to predict optimal parameters using Machine Learning techniques. For example, MapReduce (MR) optimization can be defined as: Given a MR program p to be run on input data d and cluster resources r, the goal is to find the optimal setting of parameters $copt = argmin_{C \in S}(F(p, d, r, c)$. A Cost Based Optimization (CBO) addresses this problem [HDB]. It consists of three tasks. First a job profile is generated. Then, using an What-If engine a cost associated to a configuration is predicted. Finally, a cost optimizer finds the best configuration through an enumeration and search over S.

Business Opportunity: In real applications of big data problems, finding of optimal parameters for a job execution turn out to be not an easy task. However, such a solution can help admin in application deployment. In SaaS platform this can lead to an effective cost reduction for companies. In this context CBO seems to be an interesting solution and a good alternative to a Rule based approach.

Contribution: The selected student will have to:

- State-of-the-art Study
- Define strategy to use machine learning to predict parameters :
  - Features definition
  - Learning algorithms tests
- Comparison to existing solutions

REFERENCES

[AAB+05] Daniel J Abadi, Yanif Ahmad, Magdalena Balazinska, Ugur Cetintemel, Mitch Cherniack, Jeong-Hyon Hwang, Wolfgang Lindner, Anurag S Maskey, Alexander Rasin, Esther Ryvkina, Nesime Tatbul, Ying Xing, and Stan Zdonik. The Design of the Borealis Stream Processing Engine. In Second Biennial Conference on Innovative Data Systems Research (CIDR 2005), Asilomar, CA, January 2005.

[ADGI08] Jagrati Agrawal, Yanlei Diao, Daniel Gyllstrom, and Neil Immerman. Efficient pattern matching over event streams. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, pages 147–160, New York, NY, USA, 2008. ACM.

[BPM04] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1):20, jun 2004.

[CM12] Gianpaolo Cugola and Alessandro Margara. Complex event processing with t-rex. J. Syst. Softw., 85(8):1709–1728, August 2012.

[Gol15] Yoav Goldberg. A Primer on Neural Network Models for Natural Language Processing. 2015.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. Advances in Neural Information Processing Systems 27, pages 2672–2680, 2014.

[HA04] Victoria J. Hodge and Jim Austin. A Survey of Outlier Detection Methodoligies. Artificial Intelligence Review, 22(1969):85–126, oct 2004.

[HDB] Herodotos Herodotou, Fei Dong, and Shivnath Babu. MapReduce Programming and Cost-based Optimization? Crossing this Chasm with Starfish.

[LIMK12] Erietta Liarou, Stratos Idreos, Stefan Manegold, and Martin Kersten. Monetdb/datacell: Online analytics in a streaming column-store. Proc. VLDB Endow., 5(12):1910–1913, August 2012.

[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.

[MCL15] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. Iclr, (2015):1–14, 2015.

[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.

[SGZ+16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. NIPS, pages 1–10, 2016.

[WHF17] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral Multi-Perspective Matching for Natural Language Sentences. feb 2017.