

Reviewer's opinion
on Ph.D. dissertation authored by
Alexandre QUEMY
entitled:

*End-to-End Approach to Classification in Unstructured Spaces
with Application to Judicial Decisions*

1. Problem and its impact

Alexandre Quemy addresses machine learning issues, and especially classification from heterogeneous data sources providing data of various types. His global objective is to automate the classification workflow, including the time-consuming data preparation tasks required for the learning *per se*, in business contexts where machine learning expertise is little available.

In such a context, the most important problem pertains to cope with data variety (in the sense of big data). Classification is indeed achieved by processing data attributes or features, and requires a metric (usually a distance or similarity) in the attribute space. Yet, working with data of heterogeneous types yields spaces without any *a priori* metric. Thus, Alexandre Quemy hypothesizes that the metric should be learned from the data available, with respect to training examples. Metric learning is a subfield of machine learning that is currently booming and that addresses current and relevant scientific issues.

Eventually, Alexandre Quemy has chosen as a field of application the prediction of judicial decisions, which poses many problems including interpretation and exceptions. This choice is particularly relevant, since it leverages interesting problems both in computer science and law.

2. Contribution

Alexandre Quemy's main contribution is a binary classification algorithm named Hypergraph Case-Based Reasoning (HCBR). In an automated framework, HCBR exhibits relevant features that, as a whole, significantly enhance the state of the art.

More precisely, HCBR operates in non-metric spaces, thus allowing to combine data from multiple sources without complex transformations, unlike neural networks and deep learning that work with vector spaces. Moreover, the metric learning strategy used in HCBR also helps handle any data representation and cope with missing values, unlike many machine learning techniques.

In terms of efficiency, HCBR achieves a classification accuracy that is similar to several state-of-the-art methods on several structured datasets. Moreover, HCBR also improves accuracy on the unstructured (textual) judicial dataset used as an application. In summary, HCBR is thus robust both in terms of input data and classification efficiency.

Finally, different aspects of HCBR were published, with Alexandre Quemy as the only or first author, in two international conferences of very good standing (DOLAP 2018 and DEXA 2020; both ranked B

in the CORE Conference Portal) and the *Information Systems* top-ranked journal (Q1 in the Scimago Journal & Country Rank) in 2019.

3. Correctness

Alexandre Quemy follows a sound methodology in his dissertation. He indeed takes great care to discuss and justify any choice. The definitions he states are both properly formalized and illustrated by schemas. Whenever applicable, lemmas and properties are adequately proven. Algorithms are correct and lie at the right level of detail, allowing good readability.

I noticed a small number of undefined notations throughout the dissertation, and a mix of conceptual (relationships) and logical (tables) concepts in Section 3.5 (providing a conceptual data model would have been better). Yet, these minor imperfections did not hinder my comprehension of the manuscript, overall.

Finally, Alexandre Quemy performed experiments as rigorously as he formalized the problems he addresses and the solutions he proposes, by systematically establishing experiment objectives, protocols and thoroughly discussing the experimental results, thus demonstrating an in-depth comprehension of the behaviour of the algorithms he proposes. Moreover, all his code is available on github.

4. Knowledge of the candidate

The first part of Alexandre Quemy's dissertation (three chapters) is dedicated to a thorough review of the state of the art. It not only covers the core of his work, i.e., classification and metric learning, on one hand, and the automation of machine learning processes, on the other hand, but also computational law.

All three chapters are well documented, including adequate bibliographical references, and present different aspects of the addressed topics that even span out the scope of the dissertation. Furthermore, Alexandre Quemy exhibits great written pedagogical skills and systematically draws the pros and cons of the approaches he describes, hence demonstrating his mastery of machine learning issues and approaches, his openness to other disciplines (law, here) and his general knowledge and understanding of the Information and Communication Technology discipline.

5. Other remarks

Beyond Alexandre Quemy's main contribution, I would like to underline two other aspects of his research work. The first one relates to automating machine learning workflows, which was addressed up to now as a global task scheduling problem mixing data preparation and model learning. Hypothesizing that these two steps are independent and acknowledging that data quality is preponderant in machine learning results, Alexandre Quemy cleverly cut the optimization process in two distinct parts. He showed that focusing on data preparation can be much more efficient than tuning machine learning methods, and proposed suitable time allocation policies between the two steps. This work was published in the DOLAP 2019 conference and the *Information Systems* journal in 2020.

Second, even though it is not research *per se*, Alexandre Quemy built an open database from a European Court of Human Rights' corpus of documents. After cleansing and structuring the data, he made the dataset available in various formats, with reusability, quality and availability in mind. I have but one small regret: the approach could have been pushed a little further to comply with the FAIR

(findability, accessibility, interoperability and reusability) principles. Yet, this is an important achievement, for this database is usable by both lawyers (as a knowledge base) and computer scientists (as a real-life test dataset) for scientific purposes.

6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 13 of the *Act of 14 March 2003 of the Polish Parliament on the Academic Degrees and the Academic Title* (with amendments)¹, my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with X)

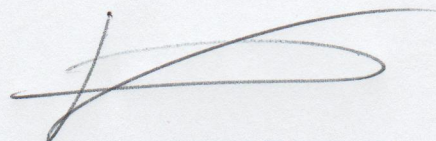
Definitely YES *Rather yes* *Hard to say* *Rather no* *Definitely NO*

B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**, and particularly the area of **machine learning**?

Definitely YES *Rather yes* *Hard to say* *Rather no* *Definitely NO*

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

Definitely YES *Rather yes* *Hard to say* *Rather no* *Definitely NO*



Signature

¹ http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf