Silvia Chiusano, Associate Professor, PhD
Politecnico di Torino, Torino, Italy
silvia.chiusano@polito.it

February 10, 2021

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

What is, in your opinion, the most important problem discussed in the dissertation?

The PhD thesis of Alexandre Quemy deals with an interesting and timing topic. Nowadays the expertise required to adopt Machine Learning (ML) based solutions can obstacle the wide adoption of such solutions by companies. Therefore, the huge amount of collected raw data remains unexplored instead of being analysed to mine useful insights. To address this issue, this PhD thesis focuses on the design and development of a ML workflow that does not require a human interaction.

All steps in the ML workflow are covered in the doctoral thesis, including data collection, data pipeline creation and the model selection. The classification problem has been selected as a reference task being a widely studied problem in ML with application in many different situations. To both design and validate the proposed framework, the prediction of judicial decisions has been selected as a reference use case.

In my opinion this dissertation addresses an important problem, and the proposed solutions can have a significant practical impact in different application domains.

Is it a scientific one?

This PhD work is across the areas of Automated Machine Learning (AutoML) and Computational Law. Challenging issues are addressed from both areas, and significant and original contributions are presented.

On the one side the possibility of making ML techniques easier to apply, and reducing the demand for human experts, has recently emerged as a hot topic with both industrial and academic interest. This research field is relatively new, and many research questions are still open.

On the other side, the judicial domain selected as a reference use case is characterized by peculiar characteristics (such as grey areas of interpretation) making this case a complex example. In addition, predicting judicial decisions is a not trivial task, but very important in practical applications.

Personally, I think that the contributions of this dissertation have a significant scientific relevance and a great value.

Does it have a practical meaning?

The contributions presented in this dissertation have different important practical impacts.

So far, a lot of effort has been devoted by the research community to the design of innovative, efficient, and effective ML algorithms. However, applying ML methods to real-world business problems is time-consuming, resource-intensive, and challenging. It requires experts in several disciplines including data scientists. As a consequence, the strong potential of ML techniques is still unexplored in various application contexts. The availability of a ML workflow based on an End-to-End Approach can make ML methods more accessible across the organizations.

The legal field, considered as a reference application case in this dissertation, can definitively benefit from a solution lowering the expertise required to adopt ML solutions. The required expertise in data science is an important barrier for a wider adoption of ML solutions in this field. However, the exploitation of ML techniques in the legal domain can favour new perspectives in legal research and practice. In addition, the considered example case, i.e., predicting the outcome of legal cases, is a challenging task even for the best legal experts.

The approach proposed in this dissertation is validated using data from the judicial domain, but it is not restricted to the legal field. Instead, it is a general solution that can be profitably exploited in other domains.

2. Contribution

What is the main, original contribution of the dissertation?

In my opinion, there are different scientific and original contributions in this PhD thesis. They can be summarized as follows:

ECHR-DB open dataset

For me, the creation of a large and open dataset (named ECHR-DB) of judgment documents related to the European Court of Human Rights (ECHR) is an important contribution of this PhD work. Also, scripts used in the ETL pipeline to generate the benchmark data repository are provided as an open-source software. These scripts are a valuable support to allow the integration of new data from researchers into ECHR, thus incrementally updating the database over time.

Open datasets present useful test cases for the research community because they allow exploring the characteristics of real data collections in a given domain as well as validating the performance of novel ML-based solutions. I suggest identifying the appropriate channels to promote in the scientific community the availability of this dataset and related scripts for the ETL process.

Novel Hypergraph Case-Based Reasoning (HCBR) algorithm

In this PhD thesis, a novel algorithm for binary classification, called Hypergraph Case-Based Reasoning (HCBR), is proposed. The key points of the HCBR algorithm are that it allows to work in unstructured spaces, has few hyperparameters, and does not require to transform the data to work which lowers the expertise required to obtain predictions.

The experimental evaluation of HCBR is conducted both on structured datasets and on unstructured datasets for text classification. HCBR is compared against state-of-the art approaches in two experimental settings: with hyperparameter tuning, and without feature engineering or hyperparameter tuning to evaluate the robustness of the proposed approach.

Experimental results conducted on some well-known structured datasets showed that HCBR provides similar accuracy than the best results from the literature (both with and without hyperparameter tuning). Experimental results conducted on unstructured datasets showed that HCBR is more accurate in most cases compared to reference study.

A two-stage optimization approach to solve the AutoML problem

The approach proposed in this PhD thesis starts from the following considerations. On the one side, the data pre-processing step is time consuming, with a huge impact on the model performance, and it requires the experience of data scientists and the expert knowledge about the data. On the other side, the data pipeline depends both on the data source and the algorithm.

In this PhD thesis, a two-stage optimization approach is proposed, articulated around the data pipeline construction and configuration, and the algorithm selection and configuration. A Bayesian optimization is used to automatically build a data pipeline in order to maximize the performance of the final model. An architecture to allocate the computation time between building the pipeline and tuning the algorithm is proposed and time allocation policies are studied.

If appropriate, you can make a distinction between what the Ph.D. candidate claims to be the main contribution and what you consider as the main contribution. If this is the case, indicate the reason for which you do not agree (e.g. it could be that somebody else has already proposed a given idea or it can be original but not correct due to some flaws described in Sec. 3 of the reviewer's opinion). You can also comment on practicality of the proposed solutions (it could be that the problem is highly practical, but the proposed solution is not).

There is not significant distinction between what the Ph.D. candidate claims to be the main contribution and what I consider as the main contribution.

If applicable, you can refer to other quality indicators you know about (e.g. quality of publications by the candidate, patents authored by the candidate, citations, existing applications of the proposed solutions etc.).

The content of the dissertation has been published as: 3 international journal papers (in journal with Scimago ranking: Quartile Q1 in the area of Computer science since 2005) and 5 good international conferences. Alexandre Quemy's publication record demonstrates that the international research community recognized the scientific value of his research activity.

3. Correctness

Can we trust what is claimed in the dissertation? Are the arguments correct? Indicate the flaws you have noticed, if any. Also point out those aspects concerning correctness that you value most (elegance of proofs, design of experiments, analysis of empirical data, quality of prototype software/hardware etc.).

The manuscript is well organized. The manuscript is well written and the contributions are clearly presented. The literature review is rich and accurate. The theoretical background for the proposed solutions is properly formalized. An extensive experimental evaluation has been conducted to validate the performance of the proposed solutions and demonstrate their effectiveness.

The contributions presented in this PhD thesis seems to be capable to achieve the objectives of this PhD work.

4. Knowledge of the candidate

What are the chapters of the dissertation (or sections in chapters) that resemble a tutorial and thus confirm a general knowledge of the candidate in the discipline of **Information and Communication Technology**.
What areas of that discipline are covered by those chapters/sections?
What do you think about quality of those chapters/sections?

Overall, the thesis presents a clear and accurate literature review on Computational Law, classification problem, and Automated Machine Learning (Chapters 2, 3, and 4, respectively).
This literature review demonstrates the very good knowledge of Alexandre Quemy in the discipline of Information and Communication Technology, in general, but more specifically in machine learning methods and the application of these methods in the legal field. This literature review provides useful insights to properly frame the contribution of this PhD work.

In Chapter 2, the literature review discusses statistical models to predict justice decisions, specific methods to model preferences and ideologies, and expert systems based on rule-based and case-based reasoning systems. I particularly appreciated the table at the end of the section (Table 2.1), that lists the main solutions available in literature in the field of computational law and concisely describes their main characteristics. This table allows the reader to easily compare the different solutions.

In Chapter 3, the classification problem is first introduced; then different classification methods are described and the state of the art in Metric Learning is presented. Also, the main differences between the HCBR algorithm presented in this PhD thesis and previous solutions are discussed at the end of the chapter.

Chapter 4 focuses on Automated Machine Learning by introducing the main problem in the field (i.e., Combined Algorithm Selection and Hyperparameter-tuning (CASH)). Also, the limits of the current approaches are discussed (see Section 4.5).

What is your opinion on the list of references?
The list of references is appropriate and up to date. It demonstrates a good knowledge of Alexandre Quemy on the subjects addressed in this dissertation.

What is the degree of its completeness? Provide any other arguments in favour or against the claim that the candidate has general knowledge and understanding of the Information and Communication Technology discipline.

The characteristics of the technical solutions presented in this PhD thesis, and how these solutions are described, demonstrate that Alexandre Query has a very good knowledge and understating of the Information and Communication Technology discipline, and particularly in the data science area.

5. Other remarks

Even if this PhD thesis is overall well written, clear and complete, I suggest addressing the following points in order to better frame this PhD work, discuss possible applications in other domains and possible future developments.

5.1. Major comments, observations, and questions

- This PhD dissertation can be framed in the context of "Automated Machine Learning" (AutoML) and "Computational Law". In my opinion, both fields should be introduced from the beginning of the manuscript, in Section "Introduction". Instead, in the current manuscript they are mentioned for the first time in Section 1.5 "Thesis Organization".
- Section "Conclusion" could be extended to better discuss how general is the proposed framework, i.e., the possibility to consider other instances of the current framework and the possibility to use the framework in other application domains. The following questions could be addressed:
  - Is it possible to identify other application domains that can benefit from the proposed framework? For example other application domains dealing with document data or other kind of data. In my opinion the medical domain can definitively benefit of a ML workflow that does not require a human interaction, for example for the analysis of patient electronic records.
  - Can the framework be easily adapted to integrate other ML methods? For example, is it possible to target the cluster analysis instead of the classification task? Can the proposed two-step optimization approach work, at least in principle, if the cluster analysis is targeted?
- Based on Alexandre Quemy's publication record, the contributions of this PhD thesis have been mainly presented to the research community in the computer science area. A possible next step for the continuation of this research activity in the future, could be the submission of manuscripts that describe the proposed approach to conferences or journals focused on the area of Computational law. This could be a way to receive useful feedbacks from domain expert on the usability and effectiveness of the proposed solution together with suggestions about possible improvements.

5.2. Minor issues

- Classification metrics ACC, MCC, and F1 are defined multiple times in the PhD thesis, i.e., at pages 61 and 99. I suggest that these metrics are defined only once, for example in a dedicated section, and then used in the different chapters.
- In Chapter 5 - page 52, and in Section 5.4.3 page 66, character "??" should be fixed
- At pag 61 a typo should be fixed: "[…] For **c**the sake of completeness, we include the standard definitions of these metrics.[…]

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

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

particularly the area of **data science, with particular focus on machine learning methods, AutoML, and Computational Law**?

■	□	□	□	□
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*

Moreover, taking into account my comments and observations above, in my opinion the dissertation by Alexandre Quemy addresses important aspects in the adoption of ML methods and provides an interesting and practical solution to research and technological problems in the areas of AutoML and Computational Laws. The very good publication record highlights the quality of the research results. My overall evaluation of Alexandre Quemy's PhD work and PhD thesis is very positive.

Signature