End-to-end approach to classification in unstructured spaces with application to judicial decisions

Alexandre Quemy
IBM, Data and AI
Faculty of Computing, Poznan University of Technology, Poznań, Poland
aquemy@pl.ibm.com

1 Context and problem importance

Classification consists in predicting whether a given element belongs to a particular class. It is one of the most common problems in machine learning due to the large amount of situations that can be modeled as a classification problem. For instance, classification techniques have been successfully applied to medecine to make diagnostic, finance to assess credit attribution, food indutstry to classify products by quality. Among those fields, the legal domain is probably the least researched while having a considerable impact on every citizen. Helping the justice system to make better decisions would reduce cost and bias, thus giving a better access to a fair justice to every citizen. On top of that, the industry recently gained interest in providing modern tools for the legal domain. Those firms, known as LegalTech, grow at an impressive pace (25% growth a year for a market estimated at over 1 billion USD per year[1]).

Predicting justice decisions is challenging by itself, even for the best legal experts: for the Supreme Court of the United States (SCOTUS), 58% accuracy has been reached, as reported in [29]. Using crowds, the Fantasy Scotus[2] project reached 84.85% correct predictions. No similar result exists in Europe at the exception of a study on a small dataset [1].

The legal environment is a *messy concept* [28] that intrinsically poses a certain number of difficulties to analyze: grey areas of interpretation, many exceptions, non-stationarity, deductive and inductive reasoning, non-classical logic, to name few. Statistical models often act as black-boxes which is redhibitory for practical applications. In other words, the legal domain combines some of the most challenging elements of today's machine learning research.

Beside the scope of classification, the main obstacle for a wider adoption of machine learning solutions by companies and institutions is the expertise required to obtain satisfying results: many companies have a large amounts of data but they lack employees with adequate knowledge of data science and machine learning. A machine learning application can be summarized by the workflow described by Figure 1:

1. **Data Collection:** some raw data must be collected, possibly from different sources. The quality of machine learning models is bounded by the quality of the data.
2. **Data Pipeline:** the data scientist creates a *data pipeline*, i. e., a sequence of operations to transform and preprocess data such that a machine learning algorithm will perform well with regards to the problem to solve (for instance, accuracy for classification).
3. **Model Selection:** the data scientist selects an algorithm to solve his problem. Each algorithm comes with *hyperparameters*, i. e., parameters that do not belong to the model but guide the learning process. Setting proper values for hyperparameters is thus of primary importance for a machine learning algorithm to correctly learn from the data.

[1] https://prismlegal.com/legal-tech-market-sizing-and-opportunities/
[2] https://fantasyscotus.lexpredict.com/

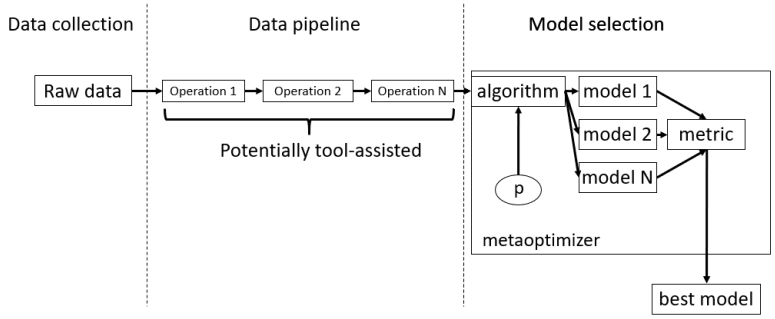

Figure 1: The machine learning workflow. Automation effort is made on the model step.

Usually, most of computational time is spent on selecting and tuning the algorithm while data scientists spend up to 80% of their time on setting up the data pipeline [7]. The state-of-the-art provides data scientists with semi-automated tools to help them to setup a *good* data pipeline. For the model selection phase, meta-optimizers are capable to automatically select an algorithm and tune the model without human intervention but at the price of a large computational time overhead. Notice that those state-of-the-art techniques are far from being widely adopted due to the computational and expertise requirement: hyperparameter tuning is sometimes not even done [8]!

It is then necessary to develop techniques to lower the expertise and human intervention required to setup a full end-to-end machine learning workflow without increasing drastically the computational time. Otherwise, the cost of adopting machine learning solutions will remain higher than maintaining less efficient processes already in place, especially in domains traditionally far from technical environments such as the legal domain.

To summarize, this doctoral project focuses on creating a **fully automated** (aka *end-to-end*) **approach to classification** in order to widespread the adoption of machine learning solutions. While the domain of applications of classification is potentially unlimited, we chose to focus on the **prediction of justice decisions** for the challenges the field offers and the little research interest the field received so far.

2 Solution proposed

In this doctoral project, we propose an alternative approach to the machine learning workflow as illustrated by Figure 2.

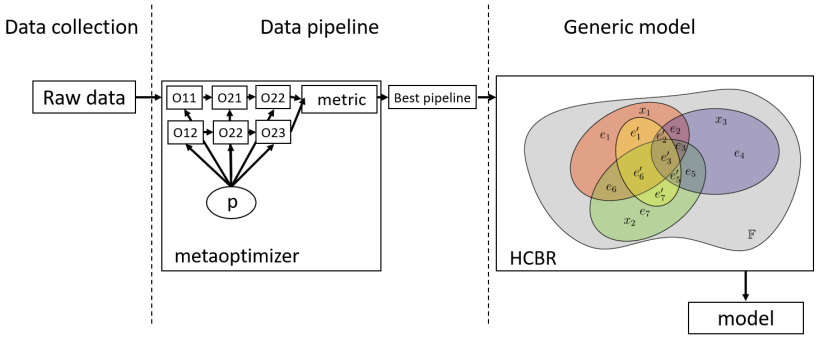

Figure 2: The modified machine learning workflow proposed as a solution to the end-to-end classification problem.

As said before, most of computational time is spent on selecting and tuning the algorithm while most of data scientists time is spent on setting up data pipelines. There is a growing consensus on the fact that data are more important than algorithms since a *good* machine learning algorithm is capable of learning almost any function [10]. Those two remarks lead us to the following solution for an end-to-end machine learning:

- **Model Selection:** if an algorithm is generic enough to work with any type of data and learn fairly complex functions while having few or no hyperparameters, we could reduce the computational time needed during the Model Selection phase without increasing the need for human expertise. This time could be spent on the Data pipelines contruction phase.
- **Data Pipeline:** as Data pipeline construction can be formulated as a (black-box) optimization problem, the meta-optimizers usually used for Model Selection can be reused to solve the problem automatically and with minimal human expertise.

As a result, the human expertise needed to transform the data and create a classification model would be minimal while the required computational time would not change or even decrease.

The main problem with the proposed workflow is that the *shape* of the data processed by the data pipeline is not know a priori and might evolve during the process. This is a problem because not all machine learning algorithms can handle every type of data. In particular, some algorithms work only with numerical data or continuous values, some cannot work with missing values or are sensitive to outliers.

Therefore, we developed a novel method for classification that works for any unstructured spaces. In other words, the algorithms accept any type of data representation. It accepts an arbitrary combination of different sources of data. The concrete *representation* of features does not influence the model. The algorithm is not sensitive to outliers and can deal with missing values. On top of that, the algorithm has few hyperparameters with little sensitivity and provide easy-to-understand explanations for each decision such that it requires minimal expertise to use and understand.

3 Contributions

State-of-the-art on data sciences techniques for the legal domain has been developped [23] and they guided the development of all the contributions of this project. The contributions of this doctoral project concern each of the three steps described on the machine learning workflow shown in Figures 1 and 2:

- **Data Collection:** due to the lack of data science resource about the legal domain, especially in Europe, a part of this doctoral project has been dedicated to collect and sanitize exhaustive data from various sources on the European Court of Human Rights to create quality datasets for the data science et machine learning community [25]. This contribution is presented in Section 3.1.
- **Data Pipeline:** a generic approach to automatically construct and optimize data pipelines using standard metaoptimization techniques is proposed and validated experimentally in [27]. This contribution is presented in Section 3.2.
- **Model Selection:** a new generic mathematical framework for classification in unstructured space called Hypergraph Case-Based Reasoning is developed in [24, 26]. This contribution is presented in Section 3.3.

3.1 European Court of Human Rights Open Data

To be able to apply our contributions to the legal domain, we needed to create datasets using real-life data. The European Court of Human Rights publishes all documents related to cases in natural language. This court is very important for all Europeans and provides over 50k decisions. From the available judgement documents, we extracted standard descriptive features (very structured database with several columns like dates, parties, court members, article in discussion) and complex bag-of-words representation from the court judgments (structured by paragraphs), including entity matching using IBM Watson Services (semi-structured representation). The data transformation, integration, and cleaning processes are described in Figure 3.

For now, the project provides 13 datasets for three versions of the classification problem: binary classification, multiclass and multilabel classification. The datasets are exhaustive (for english language) with over 11 thousands entries. Each dataset comes in three flavors: descriptive features only, textual features only and a combination of both. The datasets are available at:

- **main website:** `https://echr-opendata.eu/`

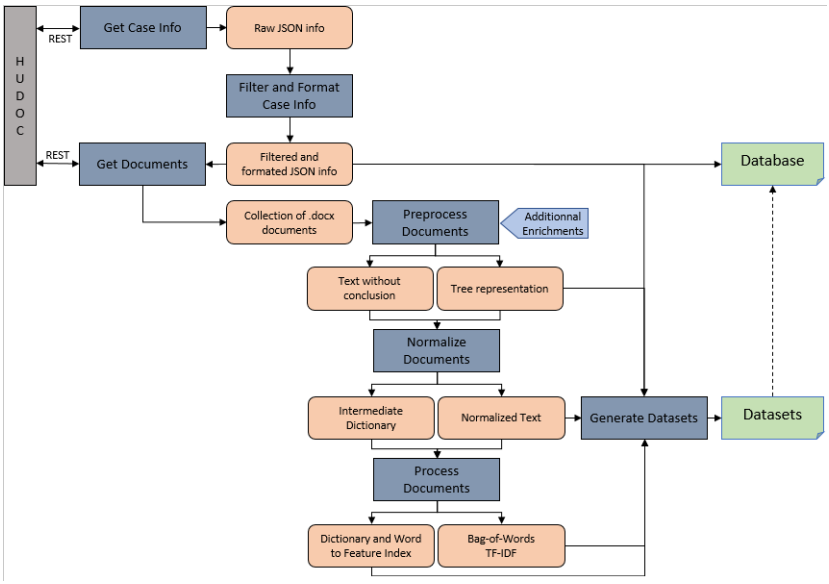

Figure 3: The ETL process to retrieve and sanitize the data, from HUDOC database to the final curated datasets.

- **download mirror:** `https://osf.io/52rhg/`

The purpose of this work is twofold. First, we expect it to draw the attention of researchers on a very important subject for society that offers new challenges to the Machine Learning community. Second, large and open datasets base on real-life data is relatively rare. Indeed, people tend not to share their data and/or keep reusing small synthetic datasets that are not reflecting the real-life difficulties.

The European Court of Human Rights Open Data project is guided by three core values: **reusability**, **quality**, and **availability**. To reach these objectives,

- each version of the datasets is carefully versioned and publicly available, including the intermediate files,
- the integrality of the process and files produced are carefully documented,
- the scripts to retrieve the raw documents and to build the datasets from scratch are open-source and carefully versioned to maximize reproducibility and trust,
- no data is manipulated by hand at any stage of the creation process.

To test the predictive power of those datasets, we performed the first experiments on all datasets and for each flavor. We compared 13 standard machine learning algorithms for classification with regards to several performance metrics. The results are consistently good across the binary datasets with an accuracy comprised between 75.86% and 98.32% for **an average accuracy of 96.45%**.

Those results provide a baseline for future studies and provide some insights about the interest of some types of features to predict justice decisions. Notably, as for previous studies, we found that textual features contain interesting elements to predict the (binary) outcome. However, for the first time, we showed that they are not as good as purely descriptive features to determine what article a given case is about, such that, for real-life predictive systems, the methodology of previous studies might not be the best.

3.2 Automated Data Pipeline Construction and Optimization

In this doctoral project, we proposed, to the best of our knowledge, the first generic way to automatically build and configure data pipeline to prepare data for any machine learning algorithm. We proposed a modified workflow is shown in Figure 4.

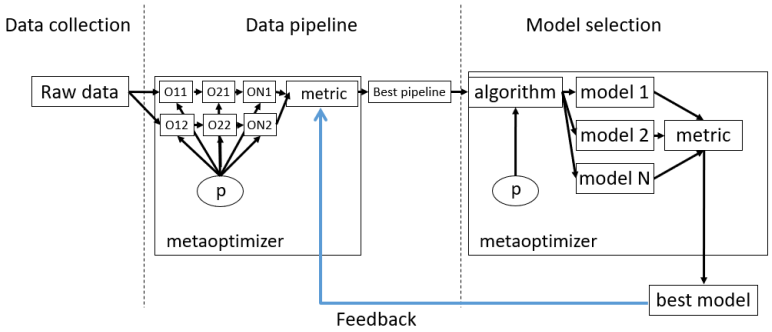

Figure 4: The workflow to automate the data pipeline construction. The main idea is to reuse standard metaoptimizer with a feedback provided by the model performance.

1. We showed that the impact of data pipeline configuration on the classification accuracy is huge in comparison to the impact of hyperparameter and model selection.
2. We defined the Data Pipeline Selection and Optimization (DPSO) problem by analogy with the CASH problem (see Section 5) for Algorithm selection and Hyperparameter tuning.
3. We showed that the data pipelines can be built and configured automatically using existing meta-optimizers, even with a restricted computational or time budget.

In order to show the potential of the approach, we defined a light grammar to define high level pipelines (called **pipeline prototype**) represented as graphs. Each node can be instantiated with several operators (e. g., . PCA), each of them having its own set of parameters (e. g., number of components in a PCA). The end user does not need to know anything about these operators.

To evaluate this approach, we designed an experiment where we defined a 3 steps pipeline prototype, as illustrated in Figure 5, for a total of 4750 possible pipeline configurations. The first step consisted in rebalancing the dataset, the second step in normalizing data, and the third step in performing a feature selection. The interations of the meta-optimizer were limited to 100 different pipelines. The experiments were performed on three benchmark datasets, i.e., Wine[3], Iris[4], and Breast[5]. To these datasets we applied the four following classification algorithms, namely: SVM, Random Forest, Neural Network, and Decision Tree. A 10-fold cross-validation was used to assess the performance of the constructed pipelines.

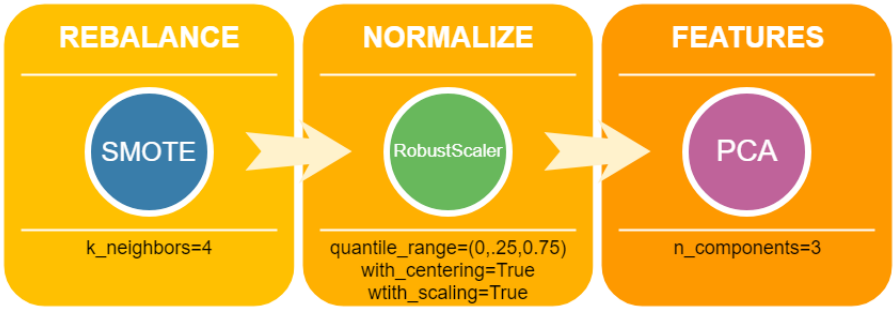

Figure 5: The pipeline prototype used in the experiments: reblance the dataset, normalize the input vectors, and then perform a feature selection.

The results for Random Forest on Breast data are shown in Figure 6. On the left, the yellow distribution of configurations explored by the algorithm is skewed towards higher accuracy, indicating our approach statistically creates good pipelines. On the right, it shows that the algorithm is really fast at finding a pipeline that performs almost the best within the search space.

[3]https://archive.ics.uci.edu/ml/datasets/wine
[4]http://archive.ics.uci.edu/ml/datasets/Iris
[5]http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

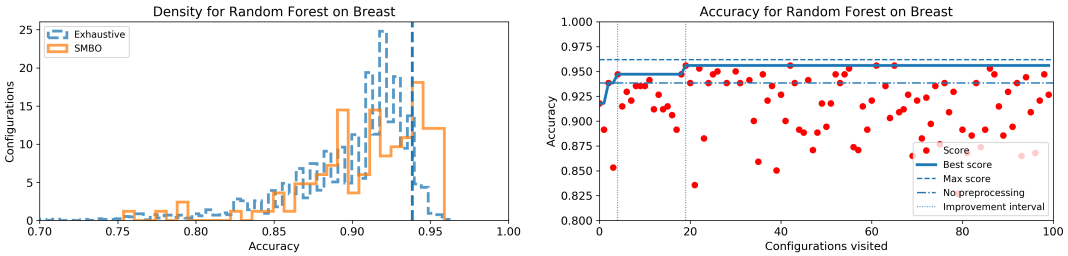

Figure 6: Example of results for Random Forest on Breast. The meta-optimizer is more likely to sample high-accuracy configurations.

In average, for all datasets and methods, with only **20 pipelines constructed (0.42% of the search space)**, the automated process was able to **decrease the error by 58.16%** compared to the baseline and reached 98.92% score in the normalized score space.

3.3 Hypergraph Case-Based Reasoning

As the third component of the proposed workflow (cf. Figure 4) we developed a new supervised algorithm for classification called Hypergraph Case-Based Reasoning (HCBR) [24,26]. As suggested by the name, it represents a training set as a hypergraph and uses the partition induced by the sub-hypergraphs to estimate, for a given subset of features, the support toward a specific class.

- It has been shown to perform as good as the state-of-art methods on some well-known datasets and unstructured datasets.
- It performs in general better than the reference study [1] on predicting the European Court of Human Rights decisions.
- It has been shown to be the most robust algorithm, i. e., performing well on average without expertise about data science, in comparison with 9 other methods.

The method offers several interesting properties, not only useful for the application to the legal domain. In particular, the model space and data representation as hypergraph provide a convenient way to explain each decision separately based on the interactions with past decisions (e.g. seen as "counter-examples" or "analogies" in case of a trial, like in Case-Based Reasoning systems). Additionally, the sensitivity to hyperparameters is negligable s.t. time-consumming tuning is not mandatory for the end-user. The hyperparameters can also be used to control the risk associated to the prediction, which is more aligned with business needs (e.g. a judge prefers to take less risk to have false positives i. e., sending an innocent to jail, while a doctor prefers to take less risk to have false negatives i. e., not detecting a cancer). Last but not least, HCBR does not assume any metric on the feature space, is not impacted by the feature representation and can work with incomplete or unstructured datasets.

For the experiments, a fast, scalable and open-source[6] modern C++ implementation of the different versions of HCBR has been developed.

4 Work-in-progress

The main remaining step for this doctoral study to be complete is a paper using the three contributions described in the previous section. This work will test the end-to-end approach that consists in the automated data pipeline construction followed by HCBR as classification algorithm, on the European Court of Human Rights datasets. The results will be compared to the baseline study from [25] and some variations using other machine learning algorithms.

The rest of the project will focus on consolidating existing work by investigating the axis of developement as follows:

[6] https://github.com/aquemy/HCBR

- **Model:** in [24, 26] we showed model space limitations (i. e., how complex are the functions the model can learn) and some model space extensions were proposed for further work. Also, we mostly studied binary classification but we would like to extend the method to multiclass classification.

- **Data Pipeline:** additional experiments on larger datasets with larger pipelines needs to be performed. In addition, we plan to explore the possibility to define a metric between pipelines, which would enable to leverage gradient-based algorithms, often more performant than black-box algorithms.

- **Datasets:** the primary results obtained on the datasets were done without spending much time on feature engineering and hyperparameter tuning. We plan to investigate how to improve the prediction performances.

5 Related work

5.1 AutoML and end-to-end ML

What is traditionally called AutoML or end-to-end machine learning focuses in reality on the Combined Algorithm Selection and Hyperparameter optimization (CASH) problem [13, 18]. It totally neglects the importance of data pipelines on the final results [9]. Among the techniques to solve CASH, the bayesian black-box optimization is the most widely used techniques. Called Sequential Model-Based Optimization [15], this framework can be instantiated in many ways, using Random Forest [15], Tree-Parzen Estimator [5] or Gaussian regression [21] to name few.

Concerning data pipeline and preprocessing, most approaches are semi-automated tools to guide the data scientists. However, recently some different (almost) fully automated approaches have been proposed.

In [22], guidelines are used to verify the quality of preprocessed data in *continuous machine learning*, i. e., machine learning models in production and receiving continuously new training data. Recently, a method using meta-features to estimate the impact of preprocessing operators on model accuracy has been proposed [6]. This approach constructs a latent space using meta-features (e. g., number of classes or attributes, entropy, noise to signal ratio) in which any dataset can be represented. A meta-learner is trained over several different datasets obtained from different raw data and data pipeline. The meta-model is thus able to predict the influence of data pipeline operators on new datasets without training the model and evaluating it using for example cross-validation. Finally, another approach consists in asking user to provide feedback on data quality to optimize pipelines [17].

Deep Learning is a family of techniques based on stacking layers of neurons whose weights are adjusted using gradient back-propagation [19]. It represents the state-of-the-art in classification in multiple domains [30]: vision, audio and natural language processing to name few. When input data can be expressed in metric space, Deep Learning is an end-to-end approach because it learns the representation itself, without any data pipeline. However, if the data are non-numerical or cannot easily be expressed as numerical, Deep Learning cannot be applied directly, and thus is not end-to-end. The core problem is the need for a metric in the space of features. On top of that, finding the proper neural architecture requires a lot of expertise, the results are highly dependant on hyperparameters and is expensive with regards to computational time.

Metric learning is concerned with learning a proper metric such that data can be correctly compared or classified [4, 31]. Choosing an appropriate pairwise metric to measure distance between two points is crucial in the success of classification algorithms [11]. Learning a metric consists in finding a projection f from an initial space to a Euclidian space, s.t. for any elements $\mathbf{x}$ and $\mathbf{x}'$, $d(\mathbf{x}, \mathbf{x}') = ||f(\mathbf{x}) - f(\mathbf{x}')||$. The metric should reflect a semantic difference between objects. Suprisingly enough, most metric learning methods assume that the data are initially represented in a vector space, which, once again, might not be suitable for plenty of problems where data can take several forms. HCBR, proposed in this doctoral project, can be seen as a metric learning method for non-metric input spaces, allowing fully end-to-end machine learning workflow.

5.2 Justice domain

Predicting justice decision is challenging by itself, even for the best legal experts: for the Supreme Court of the United States (SCOTUS), 58% accuracy has been reached in [29]. Using crowds, the Fantasy Scotus[7] project reached 84,85% correct predictions. No similar results exist in Europe at the exception of a study on a small dataset [1].

In general, the previous approaches can be broken down into three groups, namely: the statistical models, the case-based reasoning (CBR) and the abstract argumentation (AA). If the statistical methods provide interesting results for the prediction problem [14, 16, 20, 29], they cannot handle the justification problem. On the opposite, CBRs [2] do not integrate non-legal factors and thus are unable to handle the prediction problem while they provide justification to their decisions. In AA, two kinds of opposed approaches emerged: a *positive* one that intends to model real-life decision processes [3], and a *normative* one that tries to elaborate methods to select among the best alternatives and discuss arguments [12]. The first approach may handle the prediction problem and the second one the justification problem. They both heavily rely on expert knowledge to construct the so-called "arguments", which limit the applicability of AA. For a more comprehensive view on the state-of-art, we refer the reader to [23], and in particular to Table 1. The Hypergraph Case-Based Reasoning method developed in this doctoral project tries to take the advantages of statistical methods, CBR and Argumentation Frameworks, without the drawbacks.

References

[1] Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the european court of human rights: a natural language processing perspective. PeerJ Comput. Sci. 2, e93 (Oct 2016)

[2] Aleven, V., Ashley, K.D.: Evaluating a learning environment for case-based argumentation skills. In: Proceedings of the 6th international conference on Artificial intelligence and law. pp. 170–179. ACM (1997)

[3] Baroni, P., Romano, M., Toni, F., Aurisicchio, M., Bertanza, G.: Automatic evaluation of design alternatives with quantitative argumentation. Argument & Computation 6(1), 24–49 (2015)

[4] Bellet, A., Habrard, A., Sebban, M.: A survey on metric learning for feature vectors and structured data. CoRR abs/1306.6709 (2013)

[5] Bergstra, J., Yamins, D., Cox, D.D.: Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In: Proceedings of the 12th Python in science conference. pp. 13–20. Citeseer (2013)

[6] Bilalli, B., Abelló, A., Aluja-Banet, T.: On the predictive power of meta-features in openml. Int. J. Appl. Math. Comput. Sci. 27(4), 697–712 (2017)

[7] Chessell, M., Scheepers, F., Nguyen, N., van Kessel, R., van der Starre, R.: Governing and managing big data for analytics and decision makers

[8] Couronné, R., Probst, P., Boulesteix, A.L.: Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinf. 19(1), 270 (2018)

[9] Crone, S.F., Lessmann, S., Stahlbock, R.: The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. Eur. J. Oper. Res. 173(3), 781 – 800 (2006)

[10] Csáji, B.C.: Approximation with artificial neural networks

[11] Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proc. Int. Conf. Mach. Learn. pp. 209–216. ACM (2007)

[12] Dung, P.M., Kowalski, R.A., Toni, F.: Dialectic proof procedures for assumption-based, admissible argumentation. Artificial Intelligence 170(2), 114–159 (2006)

[13] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: Advances in neural information processing systems. pp. 2962–2970 (2015)

[7] https://fantasyscotus.lexpredict.com/

[14] Guimerà, R., Sales-Pardo, M.: Justice blocks and predictability of u.s. supreme court votes. PLOS ONE 6(11), 1–8 (2011)

[15] Hutter, F., Hoos, H.H., Leyton-Brown, K.: Sequential model-based optimization for general algorithm configuration. In: International Conference on Learning and Intelligent Optimization. pp. 507–523. Springer (2011)

[16] Katz, D.M., Bommarito, M.J., Blackman, J.: Predicting the behavior of the supreme court of the united states: A general approach. SSRN Electronic Journal (2014)

[17] Konstantinou, N., Paton, N.W.: Feedback driven improvement of data preparation pipelines (2019)

[18] Kotthoff, L., Thornton, C., Hoos, H.H., Hutter, F., Leyton-Brown, K.: Auto-weka 2.0: Automatic model selection and hyperparameter optimization in weka. The Journal of Machine Learning Research 18(1), 826–830 (2017)

[19] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436 (2015)

[20] Martin, A.D., Quinn, K.M., Kim, P.T., Ruger, T.W.: Competing approaches to predicting supreme court decision making. Perspectives on Politics pp. 761–767 (2004)

[21] Martinez-Cantin, R.: Bayesopt: A bayesian optimization library for nonlinear optimization, experimental design and bandits. The Journal of Machine Learning Research 15(1), 3735–3739 (2014)

[22] Polyzotis, N., Roy, S., Whang, S.E., Zinkevich, M.: Data management challenges in production machine learning. In: Proc. ACM Int. Conf. Manage. Data. pp. 1723–1726. ACM (2017)

[23] Quemy, A.: Data science techniques for law and justice: Current state of research and open problems. In: New Trends in Databases and Information Systems - ADBIS 2017 Short Papers and Workshops, AMSD, BigNovelTI, DAS, SW4CH, DC, Nicosia, Cyprus, September 24-27, 2017, Proceedings. pp. 302–312 (2017), https://doi.org/10.1007/978-3-319-67162-8_30

[24] Quemy, A.: Binary classification with hypergraph case-based reasoning. In: Proceedings of the 20th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data co-located with 10th EDBT/ICDT Joint Conference (EDBT/ICDT 2018), Vienna, Austria, March 26-29, 2018. (2018), http://ceur-ws.org/Vol-2062/paper06.pdf

[25] Quemy, A.: European court of human right open data project. CoRR abs/1810.03115 (2018), http://arxiv.org/abs/1810.03115

[26] Quemy, A.: Binary classification in unstructured space with hypergraph case-based reasoning. Information Systems (2019), http://www.sciencedirect.com/science/article/pii/S030643791830303X

[27] Quemy, A.: Data pipeline selection and optimization. In: Proceedings of the 21st International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, co-located with EDBT/ICDT Joint Conference, DOLAP@EDBT/ICDT 2019, Lisbon, Portugal, March 26, 2019. (2019), http://ceur-ws.org/Vol-2324/Paper19-AQuemy.pdf

[28] Rissland, E.L.: Ai and similarity. IEEE Intelligent Systems 21(3), 39–49 (2006)

[29] Ruger, T.W., Kim, P.T., Martin, A.D., Quinn, K.M.: The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. Columbia Law Review 104(4), 1150–1210 (2004)

[30] Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Net. 61, 85–117 (2015)

[31] Wang, F., Sun, J.: Survey on distance metric learning and dimensionality reduction in data mining. Data Min. Knowl. Discov. 29(2), 534–564 (Mar 2015)