Perspectives: Improving Interpretability of Algorithmic Errors and Algorithm-In-The-Loop Decision-Making

Kim de Bie

November 26, 2019

1 Motivation

Machine learning algorithms are used in increasingly high-stakes domains. In many cases, algorithmic predictions are used as support tools, and the responsibility for the decision remains with a human. In this way, humans serve as a safety mechanism to check whether the predictions of the algorithm make sense for an individual case (Cummings, 2006). In addition, humans can safeguard a multitude of objectives beyond accuracy that the algorithm cannot optimize for, such as fairness (Lipton, 2017). As such, the differences between human and algorithm can supposedly be leveraged such that better outcomes can be achieved than when either operates alone (Kamar, 2016).

It is unclear in which way decision-makers incorporate algorithmic predictions into their decisions. Lai and Tan (2019) propose that there exists a spectrum between fully-human and fully-automated decision-making, and show that certain types of algorithmic outputs affect human behavior more than others. This shows that algorithms should not be perceived as neutral support tools, and that the way in which algorithmic outputs are understood and perceived by humans matters. Indeed, algorithms should not only be evaluated as "statistical tools of prediction, [but] we must consider them as sociotechnical tools that take shape only as they are integrated into social contexts" (Green & Chen, 2019b).

To achieve an optimal cooperation between human and algorithm in 'algorithm in the loop' systems (Green & Chen, 2019a), a first requirement seems to be that the human develops an accurate understanding of how the algorithm performs. Indeed, "designing effective human-machine collaborations hinges as much (if not more) on presenting guidance that is tailored to human trust and understanding as it does on providing the technically optimal advice" (Green & Chen, 2019b).

The development of tools that increase the interpretability of black-box models has been proposed as a tool to assess the functioning of an algorithm, and as such to allow humans to better make use of algorithmic predictions (Doshi-Velez

& Kim, 2017). However, the model itself is essentially the only part of the machine learning pipeline that has been studied through the lens of explainability (Yin, Wortman Vaughan, & Wallach, 2019). In this context, interpretability of model evaluation metrics has largely been ignored, with several exceptions (Green & Chen, 2019b; Lai & Tan, 2019; Lucic, Haned, & de Rijke, 2019). Nonetheless, it is crucial that a user understands the uncertainty in algorithmic predictions and can make an appropriate assessment of a model's usefulness in terms of its performance (Wortman Vaughan & Wallach, 2017).

Research in numerous domains has shown that people, including experts, have difficulties understanding statistical measures such as accuracy, false positive rate, or recall (Berwick, Fineberg, & Weinstein, 1981; Handmer & Proudley, 2007), and have difficulty making correct risk assessments on the basis of such statistics (Peters et al., 2006; Reyna, 2008; Slovic & Peters, 2006). Nonetheless, the performance of algorithms is usually only reported in such terms. This raises doubts about the extent to which traditional model evaluation metrics are helpful to aid decision-making in algorithm-in-the-loop systems.

Perspectives have been proposed as a tool that can help humans makes sense of otherwise abstract quantities (Barrio, Goldstein, & Hofman, 2016; Chevalier & Vuillemot, 2013). For example, areas can be expressed in multiples of areas that an individual is familiar with: "695,000 square kilometers is about 3,000 times the size of the city of Amsterdam". We propose that perspectives that describe prediction errors can be a tool for improving the user's grasp of an algorithm's performance and errors. Eventually, perspectives may help users lead to make better decisions about when (not) to rely on the algorithm, and thus lead to better overall system performance.

2 Research objectives

2.1 Automatic generation of perspectives

2.1.1 Explaining global and local errors with perspectives

Firstly, this project will attempt to automatically generate perspectives for describing model errors. It seems that perspectives could be particularly helpful when an algorithm is used to predict an abstract amount, such as in revenue forecasting. Total revenue is most likely not a relatable unit for many users exposed to such a prediction. In this case, it might be more helpful to express metrics such as accuracy in terms of weekly sales of a particular product, or in terms of salary costs.

Perspectives may be generated for explaining both global and local errors. A global perspective would then frame the overall performance of a model (e.g. its overall accuracy as measured on some test set), while a local perspective can be used to explain the magnitude of individual errors that the algorithm has made.

The specific task for which perspectives are developed is yet to be decided upon. Discussions with AholdDelhaize staff are planned to determine where perspectives could bring the most value. The perspectives are generated based on a dataset that is most relevant to the chosen task. It can be imagined that this dataset contains e.g. data on sales, salaries, revenue by shop, over specific time periods etcetera.

2.1.2 Quality framework for perspectives

As a core objective of this work, a framework to describe the requirements for good perspectives is developed based on literature from a wide set of domains, including cognitive science, human-computer interaction and others. Two preliminary dimensions of this framework are domain-specificity and personalization.

Past research has shown that perspectives are most effective when they make use of domain-relevant objects (Hullman, Kim, Nguyen, Speers, & Agrawala, 2018; Kim, Hullman, & Agrawala, 2016; Riederer, Hofman, & Goldstein, 2018). For example, for a liquor store, we could aim to generate a perspective of the following format: "Over the past week, the model overestimated sales by 3%. To put this into perspective, this is equivalent to the value of 6 times the bottles of wine sold on a typical Saturday evening in the Diemen store." Another example: "To put this into perspective, this is equivalent to the cost of employing five 16-year-olds full-time over a month". Or: "This is equivalent to the total revenue of the store in Eindhoven in a week".

In addition to domain-specificity, it has been shown that it is beneficial to express perspectives in terms that are familiar to the individual. For example, it may be imagined that a financial manager derives more value from perspectives that express error as percentage of monthly revenue, whereas a staff planner for the store in Apeldoorn may prefer perspectives framing errors as a fraction of her employees.

2.1.3 Generating perspectives from a knowledge graph

To create perspectives, we aim to develop a model that extracts appropriate multipliers ('3 times') and objects ('the value of a banana') from a knowledge base. This knowledge base should contain data related to e.g. sales, prices of products, staff etcetera, that are relatable to the consumers of the perspectives.

The knowledge base is stored as a graph, following the approach of Chaganty and Liang (2016). From this graph, formulas can be extracted that express the desired quantity in the form of a perspective. For example:

$$5\% \text{ error} = 3 \times \text{value of a banana} \times \text{weekly banana sales} \tag{1}$$

Next, a sequence-to-sequence NLP model is applied to turn this formula into a natural sentence, such as "This is equivalent to 3 times the value of all bananas sold in a week". As there is no existing training data for this task, this will have to be created by hand. Recently, generator-pointer models have been proposed that could be particular helpful for the generation of perspectives from formulas, e.g. the work by Malmi, Krause, Rothe, Mirylenka, and Severyn (2019).

2.1.4 Personalization of perspectives

As we aim to create perspectives that concretize abstract numbers in terms that are relatable to the individual, personalization is an important aspect of this work. As stated, different units may make sense for an HR employee than for a category manager. Because the expected number of users of the system is very low, content-based recommendation strategies are more applicable than approaches based on collaborative filtering (Vuurens, Larson, & de Vries, 2016). Even within the content-based domain, only methods requiring relatively little training data are applicable. This excludes many recent neural approaches like Musto, Franza, Semeraro, de Gemmis, and Lops (2018).

For creating personalized perspectives, a personalized knowledge graph is constructed for each user as a subset of the full knowledge graph. The task of personalization of knowledge graphs is not very widely studied, and to my knowledge, no fully suitable method is available in existing literature. The only work performing a very similar task is Safavi et al. (2019), which proposes the GLIMPSE algorithm. It seems feasible to adapt this algorithm in such a way that is applicable to the current task.

GLIMPSE requires a set of expressed preferences over (a subset of) the nodes in the graph, which are initially not available. Again, preference elicitation over knowledge graphs is a very little-studied problem, and to my knowledge, there are no existing works in this domain. Moreover, the structure of the current knowledge graph lacks the semantic meaning normally present in a knowledge graphs. Therefore, a two-step approach is proposed. In the first step, items are represented individually (i.e. not on the knowledge graph), and are categorized as relevant or not. For this, given the constraints of limited existing and expected user interactions, we leverage pre-trained word embeddings to represent items (Musto, Semeraro, de Gemmis, & Lops, 2016). Next, the relevant items are extracted from the knowledge graph using GLIMPSE, and a personalized graph is constructed from these nodes.

After an initial preference model is created, this can be updated by proposing new items to users so that a wider set of relevant perspectives can be generated. Then, two diverging tasks exist: on the one hand, we want to recommend items from the knowledge base that we think users like, and on the other, we want to gather more information about items about which we are uncertain. Together, these tasks can be framed as an exploration/exploitation problem, and contextual bandit algorithms seem applicable (e.g. (Li, Chu, Langford, & Schapire, 2010)).

For gathering a user's preference on the items about which we are uncertain (i.e. exploration), we may use concepts from active learning. In the active learning paradigm, a sample with the highest uncertainty is selected for labelling in each iteration (Goudjil, Koudil, Bedda, & Ghoggali, 2018). Conversely, for proposing items that we think are relevant to the user, but that are not yet part of the knowledge graph (i.e. exploitation), we can propose items that are not labelled yet, but for which the model expresses a high certainty.

To summarize, the research questions that this part of the study aims to answer are the following:

1. What are the dimensions along which the quality of a perspective should be assessed? (What makes a good perspective?)

2. How can perspectives automatically be generated from a knowledge graph?

3. How can these perspectives be made domain-specific and personal, and compliant with other requirements as set out in the quality framework?

2.2 Assessment of usefulness of perspectives

2.2.1 Improving comprehension of algorithmic errors

To assess the usefulness of perspectives, we assess whether they lead to a better assessment of the errors that are made by an algorithm. In the context of journalism, perspectives have been shown to improve people's ability to recall measurements they have read, and estimate ones they have not. These tasks are equally applicable and relevant for assessing algorithmic error. Therefore, the first part of the user study will be to create a task that assesses the extent to which people can reproduce algorithmic error with vs. without having access to a perspective. Other tasks can be developed to assess understanding of model error.

2.2.2 Improving algorithm-in-the-loop decision-making

An important motivation for the development of perspectives is that they do not only improve understanding of the model, but also help users to make better decisions using the model. This is related to a core assumption in the field of interpretable AI: if only users understood our algorithms better, they would be more willing to use them. Here, this assumption is slightly refined. We assess not only whether trust increases, but whether trust is of an appropriate level, in the sense that users know when to rely on the system and when to disregard it (see Yin et al. (2019)). In this way, we test whether integrating perspectives into a human-in-the-loop system leads to a better overall performance of the system.

Specifically, the research questions this part of the study addresses are the following:

1. Can perspectives be used as assistance in algorithm-in-the-loop systems to improve human performance?

2. How do perspectives affect a human's understanding of a model's performance?

3. How do perspectives affect a human's trust in a model?

4. How do perspectives compare to other types of explanations, such as explanations using LIME?

References

Barrio, P. J., Goldstein, D. G., & Hofman, J. M. (2016). Improving Comprehension of Numbers in the News. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2729–2739). New York, NY, USA: ACM. Retrieved 2019-11-11, from http://doi.acm.org/10.1145/2858036.2858510 (event-place: San Jose, California, USA) doi: 10.1145/2858036.2858510

Berwick, D. M., Fineberg, H. V., & Weinstein, M. C. (1981, December). When doctors meet numbers. *The American Journal of Medicine*, *71*(6), 991–998. Retrieved 2019-11-12, from http://www.sciencedirect.com/science/article/pii/0002934381903259 doi: 10.1016/0002-9343(81)90325-9

Chaganty, A. T., & Liang, P. (2016). How Much is 131 Million Dollars? Putting Numbers in Perspective with Compositional Descriptions. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 578–587. Retrieved 2018-09-26, from http://arxiv.org/abs/1609.00070 (arXiv: 1609.00070) doi: 10.18653/v1/P16-1055

Chevalier, F., & Vuillemot, R. (2013). *Using Concrete Scales: A Practical Framework for Effective Visual Depiction of Complex Measures - IEEE Journals & Magazine.* Retrieved 2019-11-11, from https://ieeexplore.ieee.org/abstract/document/6634143

Cummings, M. L. (2006). Automation and Accountability in Decision Support System Interface Design. *Journal of Technology Studies*, *32*(1), 23–31. Retrieved 2019-11-18, from https://eric.ed.gov/?id=EJ847567

Doshi-Velez, F., & Kim, B. (2017, March). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*. Retrieved 2019-11-12, from http://arxiv.org/abs/1702.08608 (arXiv: 1702.08608)

Goudjil, M., Koudil, M., Bedda, M., & Ghoggali, N. (2018, June). *A Novel Active Learning Method Using SVM for Text Classification.* Retrieved 2019-11-25, from http://html.rhhz.net/GJZDHYJSJZZ/20180304.htm doi: 10.1007/s11633-015-0912-z

Green, B., & Chen, Y. (2019a). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 90–99). New York, NY, USA: ACM. Retrieved 2019-11-18, from http://doi.acm.org/10.1145/3287560.3287563 (event-place: Atlanta, GA, USA) doi: 10.1145/3287560.3287563

Green, B., & Chen, Y. (2019b, November). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 50:1–50:24. Retrieved 2019-11-15, from http://doi.acm.org/10.1145/3359152 doi: 10.1145/3359152

Handmer, J., & Proudley, B. (2007, January). Communicating uncertainty via probabilities: The case of weather forecasts. *Environmental Hazards*, *7*(2), 79–87. Retrieved 2019-11-12, from https://www.tandfonline.com/doi/

abs/10.1016/j.envhaz.2007.05.002 doi: 10.1016/j.envhaz.2007.05.002

Hullman, J., Kim, Y.-S., Nguyen, F., Speers, L., & Agrawala, M. (2018). Improving Comprehension of Measurements Using Concrete Re-expression Strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (pp. 1–12). Montreal QC, Canada: ACM Press. Retrieved 2018-09-25, from http://dl.acm.org/citation.cfm?doid=3173574.3173608 doi: 10.1145/3173574.3173608

Kamar, E. (2016). Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *IJCAI*.

Kim, Y.-S., Hullman, J., & Agrawala, M. (2016). Generating Personalized Spatial Analogies for Distances and Areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 38–48). New York, NY, USA: ACM. Retrieved 2019-11-11, from http://doi.acm.org/10.1145/2858036.2858440 (event-place: San Jose, California, USA) doi: 10.1145/2858036.2858440

Lai, V., & Tan, C. (2019). On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 29–38). New York, NY, USA: ACM. Retrieved 2019-11-12, from http://doi.acm.org/10.1145/3287560.3287590 (event-place: Atlanta, GA, USA) doi: 10.1145/3287560.3287590

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A Contextual-Bandit Approach to Personalized News Article Recommendation. *Proceedings of the 19th international conference on World wide web - WWW '10*, 661. Retrieved 2019-11-26, from http://arxiv.org/abs/1003.0146 (arXiv: 1003.0146) doi: 10.1145/1772690.1772758

Lipton, Z. C. (2017, March). The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]*. Retrieved 2019-11-13, from http://arxiv.org/abs/1606.03490 (arXiv: 1606.03490)

Lucic, A., Haned, H., & de Rijke, M. (2019, July). Contrastive Explanations for Large Errors in Retail Forecasting Predictions through Monte Carlo Simulations. *arXiv:1908.00085 [cs]*. Retrieved 2019-11-19, from http://arxiv.org/abs/1908.00085 (arXiv: 1908.00085)

Malmi, E., Krause, S., Rothe, S., Mirylenka, D., & Severyn, A. (2019, September). Encode, Tag, Realize: High-Precision Text Editing. *arXiv:1909.01187 [cs]*. Retrieved 2019-11-14, from http://arxiv.org/abs/1909.01187 (arXiv: 1909.01187)

Musto, C., Franza, T., Semeraro, G., de Gemmis, M., & Lops, P. (2018). Deep Content-based Recommender Systems Exploiting Recurrent Neural Networks and Linked Open Data. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (pp. 239–244). New York, NY, USA: ACM. Retrieved 2019-11-26, from http://doi.acm.org/10.1145/3213586.3225230 (event-place: Singapore, Singapore) doi: 10.1145/3213586.3225230

Musto, C., Semeraro, G., de Gemmis, M., & Lops, P. (2016). Learning Word

Embeddings from Wikipedia for Content-Based Recommender Systems. In N. Ferro et al. (Eds.), *Advances in Information Retrieval* (Vol. 9626, pp. 729–734). Cham: Springer International Publishing. Retrieved 2019-11-25, from http://link.springer.com/10.1007/978-3-319-30671-1_60 doi: 10.1007/978-3-319-30671-1_60

Peters, E., Västfjäll, D., Slovic, P., Mertz, C., Mazzocco, K., & Dickert, S. (2006, May). Numeracy and Decision Making. *Psychological Science*, *17*(5), 407–413. Retrieved 2019-11-12, from https://doi.org/10.1111/j.1467-9280.2006.01720.x doi: 10.1111/j.1467-9280.2006.01720.x

Reyna. (2008, January). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*(1), 89–107. Retrieved 2019-11-12, from https://www-sciencedirect-com.proxy.uba.uva.nl:2443/science/article/pii/S1041608007000428 doi: 10.1016/j.lindif.2007.03.011

Riederer, C., Hofman, J. M., & Goldstein, D. G. (2018). To Put That in Perspective: Generating Analogies that Make Numbers Easier to Understand. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (pp. 1–10). Montreal QC, Canada: ACM Press. Retrieved 2018-09-26, from http://dl.acm.org/citation.cfm?doid=3173574.3174122 doi: 10.1145/3173574.3174122

Safavi, T., Mottin, D., Belth, C., Muller, E., Faber, L., & Koutra, D. (2019). Personalized Knowledge Graph Summarization: From the Cloud to Your Pocket. , 10.

Slovic, P., & Peters, E. (2006, December). Risk Perception and Affect. *Current Directions in Psychological Science*, *15*(6), 322–325. Retrieved 2019-11-12, from https://doi.org/10.1111/j.1467-8721.2006.00461.x doi: 10.1111/j.1467-8721.2006.00461.x

Vuurens, J. B. P., Larson, M., & de Vries, A. P. (2016). Exploring Deep Space: Learning Personalized Ranking in a Semantic Space. *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems - DLRS 2016*, 23–28. Retrieved 2019-11-25, from http://arxiv.org/abs/1608.00276 (arXiv: 1608.00276) doi: 10.1145/2988450.2988457

Wortman Vaughan, J., & Wallach, H. (2017). The inescapability of Uncertainty..

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 279:1–279:12). New York, NY, USA: ACM. Retrieved 2019-11-12, from http://doi.acm.org/10.1145/3290605.3300509 (event-place: Glasgow, Scotland Uk) doi: 10.1145/3290605.3300509