Non-Deterministic Solvers and Explainable AI through Trajectory Mining

Martin Fyvie, John A.W. McCall and Lee A. Christie

The Robert Gordon University, Garthdee Road, Aberdeen, UK

Abstract. Traditional methods of creating explanations from complex systems involving the use of AI have resulted in a wide variety of tools available to users to generate explanations regarding algorithm and network designs. This however has traditionally been aimed at systems that mimic the structure of human thought such as neural networks. The growing adoption of AI systems in industries has led to research and roundtables regarding the ability to extract explanations from other systems such as Non-Deterministic algorithms. This family of algorithms can be analysed but the explanation of events can often be difficult for non-experts to understand. Mentioned is a potential path to the generation of explanations that would not require expert-level knowledge to be correctly understood.[1]

Keywords: XAI · Non-Deterministic · Trajectories · Data Mining

1 Introduction

Explainable AI (XAI) Adoption has been an area of growing interest for several years and as the adoption of AI decision making systems continues to increase, the need for explanations of a suitable quality by the end user has also grown. This growth in adoption of AI decision making processes in industries in which explanations are critical, such as the medical field, has lead to increased awareness of the need for high quality explanations regarding the decisions and recommendations that these systems provide. As seen in the collection of recommendations and conclusions from PHG [1], the issue of explainability regarding black box systems in industries has become a matter of much concern. Some of the themes from the series of talks and round tables closely match the three pillars of Accountability, Responsibility and Transparency outlined as the "ART" principals for AI from the comprehensive survey of explanation methods for AI in [2]:

– Accountability – refers to the need to explain and justify one's decisions and actions to its partners, users and others with whom the system interacts

- Responsibility refers to the role of people themselves and to the capability of AI systems to answer for one's decision and identify errors or unexpected results
- Transparency refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to its environment, and to the governance of the data used and created.

As methods of explanation generation continue to develop over time it is key that, as industry adoption and regulation grows, so does the ability to explain the solutions provided by such AI systems. This in turn may aid in the acceptance of a set of solutions generated by the end user and an overall increase in understanding of the search methods and how the solutions were arrived at by the end user. Historically, this area of research has focused on optimisation systems that mimic the thought processes of human reasoning. These processes can be traced, and the path taken by these systems, such as Artificial Neural Networks, can be somewhat more readily derived and understood by the end users of these systems.

2 Non-Deterministric Solvers

Non-Deterministic Solvers or Non-Deterministic Algorithms are a metaheuristic search method that search for a solution or set of solutions to a given problem using stochastic processes. These stochastic processes, such as those in population-based algorithms like Genetic algorithms, capture in a series of steps the learnings of the algorithm as they solve the optimisation problem rather than through the use of prior knowledge. This method allows the algorithm to avoid an exhaustive search of the problem space with the trade-off that solutions are often near-ideal. The stochastic search processes utilised mean that the creation of explanations based on a technical description of the decision points of an NDA tend to be difficult for non-experts and most end users to understand due to their complexity and need for prior knowledge of the system itself.
As an example, consider a delivery scheduling system for packages. It may fall on a company representative to attempt to explain the order that delivery sites were placed in or why a customer's site is the last to be visited. The exact path taken by an NDA cannot be predicted in advance and so with each run, a slightly different schedule may be created. A significant barrier to explanation in this scenario would be the end users lack of detailed information on how the scheduling system algorithm operates and how each decision was arrived at in order to form the solution provided.

3 Current Solutions

Existing approaches to the explanation of decision points within AI systems include the use of surrogate models. This approach involves the creation of an external and more easily understood model that is trained on the results of the

NDA. The extraction of explanations is then performed on this surrogate model in order to discover how solutions were arrived at. Examples of this approach can be seen the **Local Interpretable Model-Agnostic Explanations** system (LIME)[3]. Another example is the **SHapley Additive exPlanations** [4] (SHAP) game-theory based approach using its internal metrics of Local Accuracy, Missingness and Consistency to calculate the influence if individual features in a variety of AI systems. The output from these methods can often be used in conjunction with Natural Language Generation [5] techniques to create content-specific explanations using a rules-based approach. This can be seen in [6] in which the output from the LIME system is entered into a set of pre-generated sentences to give the end user a better understanding of the algorithm solution, in this case credit decisions, to produce sentences such as"... The single greatest contribution to the decision is from the variable 'current account' with the value of 'in debit' this produced 40% of the whole decision, influencing the algorithm to refuse credit..."

A novel approach to this problem that may provide a new avenue of explanation generation is the post-hoc analysis of the changes between populations generated by NDAs. This has the benefit of avoiding further runs of an algorithm as the populations of solutions has already been created and chart the trajectory through the solution space that the algorithm has taken to arrive at the solution or set of solutions offered to the end user.

4 Algorithm Trajectories

Algorithm Trajectories represents the collective implicit knowledge gained by a population-based algorithm as it traverses the solution space for high fitness solutions to a provided problem and all decision steps taken. This knowledge is comprised of the populations created throughout the search process. Each population contains a set of solutions selected by the algorithm for their fitness and throughout the search the solutions are altered with the intent to increase their fitness. Over time as higher fitness patterns are found within any given population, these may propagate throughout, leading to an overall decrease in population diversity as the search converges in a set of similar near-optimal solutions. Collectively, these populations over the course of one or more optimisation runs can be considered the Algorithm Trajectory.

This may prove to be a rich source of features and statistical evidence that can be mined and visualised with the aim of generating an explanation with suitable detail as to increase the end users ability to understand the problem and how the decision was arrived at. The results of mining this algorithm trajectory, in the form of visualisations and statistical figures, can then be used to develop an explanation that is more easily understood through the inclusion of Natural Language Generation (NLG) [5] in a similar manor to the examples mentioned earlier in this paper.

One example of using the post-hoc method of algorithm trajectory mining to generate visualisations for the purpose of algorithm comparison can be seen in

[7]. In this paper the authors present the structure "Search Trajectory Networks" building on local optima networks [8] values determined in such a way that it is possible to visualise the exact path taken across the solution landscape by the tested algorithms. This has the potential to add support to explanations generated in a similar fashion to the earlier examples that use NLG in combination with LIME coefficient outputs and SHAP values.

5 Conclusion

While it is possible to generate explanations that cover the decision points of NDAs and their path through the search space, they often require expert-level knowledge of the model and concepts to understand. In this position paper, we hypothesise that it is possible to develop a method following a similar path to the previously mentioned surrogate model and LIME / SHAP example. This hypothesised method would draw valuable statistical evidence and information on problem structure from the algorithm trajectories generated during an optimisation run. This post-hoc approach would involve using NLG to convert numerical and scientific data into human-understandable sentences and paragraphs while being supported by visualisation techniques such as those outlined in [7]. This approach has the potential to be a valuable source of human-understandable explanations for complex NDA based systems with the aim of increasing end user trust in the solutions provided.

References

1. J. Ordish, T. Brigden, and A. Hall, "Black box medicine and transparency — PHG Foundation," p. 34, 2020.
2. A. Adadi and M. Berrada, "Peeking Inside the BlackBox: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.
3. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?" Explaining the predictions of any classifier," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 1317August2016, pp. 1135–1144, 2016.
4. S. M. Lundberg, G. G. Erion, and S. I. Lee, "Consistent individualized feature attribution for tree ensembles," arXiv, no. 2, 2018.
5. A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," Journal of Artificial Intelligence Research, vol. 61, p. 65–170, 2018.
6. J. Forrest, S. Sripada, W. Pang, and G. M. Coghill, "Towards making NLG a voice for interpretable Machine Learning," INLG 2018 11th International Natural Language Generation Conference, Proceedings of the Conference, pp. 177–182, 2018.
7. Gabriela Ochoa, Katherine M. Malan, Christian Blum, Search trajectory networks: A tool for analysing and visualising the behaviour of metaheuristics, Applied Soft Computing, Volume 109, 2021, 107492, ISSN 1568-4946
8. Ochoa G., Tomassini M., Verel S., Darabos C. A study of nk landscapes' basins and local optima networks Genetic and Evolutionary Computation Conference, GECCO, ACM (2008), pp. 555-562