Investigating the duality of CBR: Performance and Interpretability

Prateek Goel[1,*,†]

[1]Dept. of Computer and Information Sciences, Drexel University, 3675 Market St, Philadelphia, 19104, USA

Abstract

Case-based reasoning (CBR) is an artificial intelligence (AI) methodology with applications across different domains. Its reasoning is rooted in human intuition making this methodology universally accepted as interpretable. In contrast, black box models like neural networks have a widely accepted notion with high-performance accuracy and low interpretability. Despite low interpretability, there is a ready acceptance of black box models in academia and industry for automated decision-making. The focus of this research is investigating the relationship between performance and interpretability for CBR by exploring a variety of ways to note the impact of enhancing performance on CBR interpretability.

Keywords

Interpretability, Case-Based Reasoning, Feature Weighting, Statistical Relevance

1. Introduction

Artificial Intelligence (AI) dominates in automated decision-making applications across a variety of domains. Within the methodologies of AI, the prevalence of black-box models is well-known due to their renowned performance during practical application. A widely accepted notion affiliates the increased complexity of black-box models like Neural Networks (NNs), Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), etc., with high performance. However, the same reason makes these methods non-intuitive. By increasing their design complexity, these methods lack an intrinsic insight into their reasoning making them uninterpretable.

Case-based reasoning (CBR) is an alternative AI methodology to black-box models, that derives its workings from the natural intuition of how humans think. CBR attempts to answer a newly encountered experience using the knowledge from past experiences creating a "natural" sense of interpretability widely accepted across the AI community [1, 2]. However, due to their noted high performance, black-box models are the ones that are predominantly used for high-stakes decision-making despite low interpretability.

This work examines the performance of CBR by investigating its increase with the application of different approaches and noting any influence on interpretability. It is a known notion that using increasingly complex models leads to a decrease in explainability [3, 4, 5]. By this notion, it is worth investigating if that behavior is replicated within CBR as well. Furthermore, exploring

ICCBR DC'23: Doctoral Consortium at ICCBR2023, July 17 – 20, 2023, Aberdeen, Scotland

*Corresponding author.

pg427@drexel.edu (P. Goel)

0000-0003-4644-0610 (P. Goel)

different ways to improve CBR performance (while maintaining its interpretability) would detail impact factors that can serve as a reference to future researchers and provide for the ready acceptance of CBR in high-stakes decision-making as an interpretable alternative to black-box methods.

2. Literature Review

Case-based Reasoning Case-Based Reasoning (CBR) is an AI methodology that uses knowledge from past experiences to solve new problems. What distinguishes it from other reasoning methodologies is its ability to identify the most "similar" experiences for solving a problem. Since the solution of the experience(s) may not be fully applicable to the current problem, it allows the methodology to tweak (or adapt) the past experience to solve the problem [6, 7].

CBR has two foundational perspectives of understanding. The first perspective is as a process cycle (called CBR cycle[6]) where a problem is formulated using details from encountered experience, followed by identification of the similar past problem(s) capable of providing a solution. The solutions to the identified similar problems are adapted to form a new solution. An alternative perspective is with respect to the overall knowledge of the CBR system stored in separate knowledge containers. Each container possesses a different aspect of knowledge contributing to the overall knowledge of the CBR cycle[6, 7]. This study focuses on the knowledge of similarity contained within the similarity container for the CBR cycle[6].

A set of approaches pertaining to the similarity container are reviewed. Wetterscherek and Aha [8] list a variety of similarity-based weight learning approaches that allocate similarity weights per feature of a case problem to obtain the most similar cases to an encountered case problem. Maximini et al. [9] suggest using the concept of 'generalizing' cases to form cases with a feature range for the selection of appropriate cases. Jalali and Leake [10] suggest defining rules guided by the adaptation step within the retrieval step.

Interpretability Interpretability in A.I. stems from different notions inspired to gather a holistic understanding of a machine learning model's operability. The field has gained immense popularity in the past decade primarily due to the increased usage of black box AI models in high-stakes decision-making. Despite the stakes, the field lacks a unifying definition to bind all affiliated notions [2, 11]. Numerous sources have pointed out the subjective nature of interpretability listing it as a domain-specific notion with no unifying definition [4, 2, 5].

Interpretability, as it stands today, is an amalgamation of different notions. Rudin [2] details an interpretable model that has either usability or some structural knowledge of the domain it is being applied to. Arrieta [4] refers to it as a passive character of the model such that the model makes sense to the human observer. Doshi-Velez and Kim [12] relate it to explainability and suggest an interpretable model can explain its reasoning in understandable terms.

An important subject of discussion is the relationship between performance and interpretability for different AI methods. The DARPA XAI Program [3], showcased a notion of the performance-explainability tradeoff. This relation is showcased as inversely proportional where the performance of a model increases at the cost of its explainability. This trade-off has been a subject of discussion where either the trade-off has been rejected completely [2] or

suggested to hold true under specific circumstances [4].

CBR methodology claims to have "natural" transparency (or inherent interpretability) [2, 1]. Due to the methodology's foundations lying in psychological plausibility, there is universal consensus on the interpretability of CBR to the point where the methodology itself is used to provide post-hoc explanations for other black box models. Keane and Kenny [1] use the "twin systems" approach by creating an ANN-CBR twin with shared weights where simultaneous learning of feature weights provides for interpretations using CBR. Caruana et al. [13] generate case-based explanations for artificial neural networks and decision trees.

Rudin [2] and Miller [14] identify the limitation of humans processing information by listing how humans are limited to processing 7 (+-2) cognitive entities at one time. Using this notion, Lage et al. [15] list different factors (called "cognitive chunks") that make models interpretable by building decision sets and conducting human-subject experiments to analyze the impact on performance. The conclusions are used to seed the investigation of CBR interpretability in this research and analyze how various CBR methods accordingly evaluate.

3. Research Plan

This research aims to investigate methods to improve CBR performance. In addition, the impact of these methods on interpretability will be noted. Finally, the structure of the relationship between performance and interpretability will be analyzed and the level of sensitivity is observed in case a relationship is verified.

3.1. Research Objectives

The first objective of this research is to improve CBR performance. The aim is to investigate a variety of methods and approaches to enhance CBR performance, where the approaches are limited to the retrieval step of the CBR cycle (or similarity container). The second objective is to note the impact of CBR approaches considered in the first research objective on interpretability. This objective would entail identifying the factors to govern CBR interpretability objectively and using these factors to witness changes in the approaches considered in the first research objective. Additionally, this objective aims to describe the structure of the relationship between performance and interpretability for the approaches considered in research objectives one and two. The aim is to identify the level of sensitivity between the two characteristics for approaches when an impact is noted.

3.2. Approaches and Methodologies

The entirety of this research is divided into two phases. Phase 1 studies examine the literature on what has been proposed for performance in CBR and note increased performance as a direct consequence of a proposed alternative to similarity and retrieval-based approach.

This notion is explored via three studies. The first study incrementally increases the number of weights per feature to describe the relative relevance of features describing the cases. This includes methods involving clustering and weight-learning approaches. The second study uses the proposed methods in the literature for enhancing CBR performance. This comparison

study would be comparing existing approaches in the literature by identifying datasets used across reviewed literature. The third study compares weight methodologies and their impact on performance. Where the focus is on increasing the number of weights in the first study, this study focuses on how weights are learned while keeping the number of weights constant. An additional aim of this phase is coming up with an approach built using approaches considered within the studies, that mimic the performance behavior of a neural network model.

Phase 2 investigates CBR interpretability notions to obtain factors that influence the interpretability of CBR objectively. Consequently, this phase analyzes the relationships between the performance and interpretability of CBR approaches and notes the sensitivity of the impact one has on the other. Using the suggested concept of "cognitive chunks" discussed in Lage et al. [15] as information units, this phase studies the relationship, with respect to information units, between performance and interpretability by analyzing different approaches from Phase 1. By associating these information units with a number of weights per feature, the goal is to gain some visibility in their relationship. Additionally, this phase hypothesizes the existence of "soft spots" in the performance-interpretability relationship for different methods. These spots are points at which the limited increase in performance comes at a greater cost to interpretability. To draw out the relationships for every method listing their respective soft spots, a quantitative evaluation would be done detailing the performance-interpretability relationship. Post this quantitative evaluation, a user study will be conducted to get validation from humans to confirm the hypothesis and selection of the "soft spots" and note the level of correctness of the identified performance-interpretability relationship.

4. Progress Summary

The first study for Phase 1 is ongoing. The baseline approach of gradient descent-based weight learning (1 weight per feature) is altered by clustering cases according to some distance metric. Once cases are clustered, each cluster has its own sets of weights (using gradient descent) to separate the cases within. This gives every feature within a case represented by two weights instead of one. One weight suggests the cluster the instance belongs to and the other weight corresponds to the gradient descent method for the instance feature. This methodology was evaluated using a 5-fold cross-validation split using the Credit Default Binary classification Dataset [16], comprised of 30,000 instances with 23 features in each instance. The performance was evaluated using accuracy as a metric to compare against baseline retrieval using gradient descent-based weight learning. The noted average accuracy was marginally higher than the baseline model with performance accuracy noted to be 81.4% for this method and 79.8% for the baseline. The optimum number of clusters was 3 and calculated using the elbow method. To confirm the significance of these results McNemar's test[17] was performed. The null hypothesis, suggesting the same performance accuracy across both models, is rejected with a 95% confidence level With a p-value of 2.016e-12.

The design for the second study is ongoing where the methods discussed in the reviewed literature, presented in the Related Works section, will be applied to note enhancements in performance.

5. Conclusion

This research is in its preliminary phases and no conclusions have been drawn yet. For limitations, the scope considers only the classification task. The investigations are, also, limited to the retrieval step and the similarity container.

References

[1] M. T. Keane, E. M. Kenny, How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems, in: K. Bach, C. Marling (Eds.), Case-Based Reasoning Research and Development, Springer International Publishing, Cham, 2019, pp. 155–171.

[2] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. URL: https://doi.org/10.1038/s42256-019-0048-x. doi:10.1038/s42256-019-0048-x.

[3] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI Magazine 40 (2019) 44–58. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850. doi:10.1609/aimag.v40i2.2850.

[4] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103. doi:https://doi.org/10.1016/j.inffus.2019.12.012.

[5] G. K. Dziugaite, S. Ben-David, D. M. Roy, Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability, 2020. arXiv:2010.13764.

[6] M. M. Richter, R. O. Weber, Case-based reasoning: a Textbook, 1 ed., Springer Berlin, Heidelberg, 2013. doi:https://doi.org/10.1007/978-3-642-40167-1.

[7] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, AI Communications 7 (1994) 39–59. URL: https://doi.org/10.3233/AIC-1994-7104. doi:10.3233/AIC-1994-7104, 1.

[8] D. Wettschereck, D. W. Aha, Weighting features, in: Proceedings of the First International Conference on Case-Based Reasoning Research and Development, ICCBR '95, Springer-Verlag, Berlin, Heidelberg, 1995, p. 347–358.

[9] K. Maximini, R. Maximini, R. Bergmann, An investigation of generalized cases, in: K. D. Ashley, D. G. Bridge (Eds.), Case-Based Reasoning Research and Development, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 261–275.

[10] V. Jalali, D. B. Leake, Adaptation-guided case base maintenance, in: AAAI Conference on Artificial Intelligence, 2014.

[11] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue 16 (2018) 31–57. URL: https://doi.org/10.1145/3236386.3241340. doi:10.1145/3236386.3241340.

[12] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017. arXiv:1702.08608.

[13] R. Caruana, H. Kangarloo, J. D. Dionisio, U. Sinha, D. Johnson, Case-based explanation of non-case-based learning methods, Proc AMIA Symp (1999) 212–215.

[14] G. A. Miller, The magical number seven plus or minus two: some limits on our capacity for processing information, Psychol Rev 63 (1956) 81–97.

[15] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, F. Doshi-Velez, Human evaluation of models built for interpretability, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7 (2019) 59–67. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5280. doi:10.1609/hcomp.v7i1.5280.

[16] I.-C. Yeh, C.-h. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Syst. Appl. 36 (2009) 2473–2480. URL: https://doi.org/10.1016/j.eswa.2007.12.020. doi:10.1016/j.eswa.2007.12.020.

[17] Q. McNEMAR, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika 12 (1947) 153–157.