

Leveraging SHAP and CBR for Dimensionality Reduction on the Psychology Prediction Dataset

Zachary Wilkerson, David Leake and David Crandall

Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington IN 47408, USA

Abstract

Effective dimensionality reduction for feature spaces can benefit the accuracy, efficiency, and/or explainability of models using the features. One task of the Explainable AI Challenge of the 2022 International Conference on Case-Based Reasoning is to apply explanation methods for informed feature pruning to refine an artificial neural network model for depression screening. This paper explores how explanations provided by SHAP values can guide feature pruning. It presents and evaluates four approaches developed for this task: 1) iterative feature pruning based on SHAP values, 2) using case-based reasoning to guide search through potential feature prunings, 3) pruning based on using Hamming distance between a reference case and patient cases to partition the case base, and 4) evaluating feature prunings in light of semi-factual and counterfactual cases. Results show that both using case-based reasoning and absolute SHAP values can guide feature pruning to improve model accuracy, with best performance occurring when SHAP values inform feature pruning selection for case-based reasoning-based methods.

1. Introduction

This paper responds to the Explainable AI Challenge at the 2022 International Conference on Case-Based Reasoning, addressing the Psychology Prediction task to apply explanation methods to improve an artificial neural network’s classification accuracy by pruning the feature space of its training dataset [1]. In addition to the accuracy benefits of removing “noisy” features, reducing the number of features could facilitate human assessment of case similarity, making the process potentially beneficial for interpretability of a case-based classifier for this domain. Consequently, our evaluation considers both classification accuracy and the ability to minimize the feature set while maintaining accuracy.

We propose and evaluate four approaches for improving the neural network’s classification accuracy by removing features, testing them both with and without feature importance information from an explanation component based on SHAP [2] and provided by the Challenge. The first approach (Importance-FP) directly applies SHAP values, iteratively pruning the features whose SHAP values suggest that they contribute least to classification. The second approach, CBR-based feature pruning (CBR-FP), uses CBR to focus search of the space of feature sets, favoring exploration of sets that are novel or similar to sets that have performed well in past iterations. This approach is based on Hoffmann and Bergmann’s HypoCBR [3], which uses

ICCBR XCBR’22: 4th Workshop on XCBR: Case-based Reasoning for the Explanation of Intelligent Systems at ICCBR-2022, September, 2022, Nancy, France

✉ zachwilk@indiana.edu (Z. Wilkerson); leake@indiana.edu (D. Leake); djcran@indiana.edu (D. Crandall)

ORCID 0000-0002-8666-3416 (D. Leake)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CBR to evaluate the viability of candidate parameterizations for a neural network. The third approach, Hamming Feature Pruning (Hamming-FP), maps each data point to its distance from a provided set of expected feature values determined by a domain expert using Hamming distance. Based on these distance values, the approach favors feature prunings for which similar distances correlate with similar classifications. The fourth approach, Semi/Counterfactual Feature Pruning (SC-FP), is inspired by CBR research on semi-factual and counterfactual explanations [4]. It selects feature sets by directly testing whether changing the values of particular features generates a case of the same classification (semi-factual) or different classification (counterfactual). This is based on the hypothesis that a good pruning strategy retains features that preserve and are significant to correct classifications (i.e., classifications are more likely to change after inverting feature values). We first test baseline versions of the algorithms and then test their performance when biased using SHAP values. Results show that using SHAP values provides modest accuracy improvements, with CBR-FP yielding particularly promising results.

2. Experimental Setup and Methods

All experiments prune features to improve performance of a provided multilayer perceptron classifier model with ten hidden layers and logistic activation functions, trained on the Psychology Prediction dataset [1]. Baselines are 1) no features removed (0-FP), and 2) feature prunings selected randomly, with the most accurate pruning returned after a set number of iterations (Random-FP). We compare the four methods described below against these baselines. We also explore using SHAP values as relative probabilities weighting the selection of features to prune at each iteration in each approach.

2.1. Importance-FP: Iteratively removing the least-contributing feature based on SHAP values

This method calculates the SHAP values for the model, prunes the least-contributing feature, calculates the model accuracy, and then repeats the process on the resulting feature subset until no more features can be removed. The pruning resulting in the highest accuracy is returned.

2.2. CBR-FP: Using CBR to evaluate potential prunings for uniqueness and/or projected accuracy

This approach begins by selecting a candidate pruning as in Random-FP, but it only applies the pruning if its features are suitably unique from the most similar already-explored pruning or if the retrieved pruning resulted in a suitable classification accuracy. The CBR system is retrieval-only, calculating similarity based on match distance (i.e., if two cases prune the same feature, the distance component is 0, else 1). The pruning resulting in the highest accuracy after a given number of iterations is returned.

2.3. Hamming-FP: Estimating decision boundaries using distances to an extreme case

This method treats patients in the provided dataset as cases and generates a prototypical extreme case by aggregating the values from the Challenge-provided “expected features” document (representing a domain expert’s “textbook example” for a patient at highest depression risk). It generates a random feature pruning for each experimental iteration. Using the features preserved in the current pruning, it calculates distances from each patient case to the prototype using Hamming distance, implicitly defining a spectrum between the extreme case and the patient case farthest from it on which all other cases sit. Finally, it partitions this spectrum into distance-based segments corresponding with output classes such that the number of correctly-classified cases is maximized. As prunings are randomly generated across iterations, the dimensions of the spectrum, cases’ locations on it, and the locations of optimal partitions change accordingly; the algorithm finds the pruning that results in the highest number of correctly-classified cases.

2.4. SC-FP: Using semi/counterfactual explanation behaviors as proxy for model behavior with feature pruning

This approach is based on the hypothesis that observing the effect on classification of “flipping” the value of a binary feature can help predict the effect of removing that feature. This algorithm assesses the value of randomly-generated candidate feature sets by flipping feature values of randomly-generated feature subsets, estimating the result using CBR retrieval. It favors prunings such that flipping feature values retains correctly-classified cases (semi-factual) while changing incorrectly-classified cases to the correct class (counterfactual). The pruning yielding the greatest increase in correctly-classified cases is selected for evaluation.

3. Results and Discussion

Table 1 illustrates the average and maximum accuracy values and corresponding number of features pruned for each approach, with and without sampling bias based on SHAP values. Average and maximum classification accuracy values are calculated using ten-fold cross-validation over 25-30 experimental trials. Broadly, accuracy-based methods (e.g., CBR-FP and Importance-FP) lead to higher accuracy values. Additional findings are discussed in detail below.

3.1. The number of features removed for best accuracy depends on the method used

One surprising result is the wide range of final feature set sizes. Hamming-FP removes few features, suggesting that its simple distance calculation is most stable/performant only for smaller prunings. By contrast, SC-FP removes a significant number of features. Because feature values are flipped rather than removed, this may introduce some information used by SC-FP’s CBR component and not used by the neural network model. CBR-FP usually has only slightly better accuracy than Random-FP, but has the potential benefit of tending toward larger pruning

Table 1

Results for random sampling and sampling biased using SHAP values. Accuracy values are percentages. The number of features removed is listed in the final column. Errors are one standard deviation. The top three average and maximum accuracy values and corresponding pruning sizes are boldfaced (one tie has the top four boldfaced).

Approach	Iters.	Avg. Accuracy	Feats. Pruned	Max Accuracy	Feats. Pruned
0-FP	1	50.8	0	50.8	0
Random-FP	100	53.5 ± 1.7	10.5 ± 7.9	57.6	15
	1000	56.4 ± 1.1	15.6 ± 5.8	58.5	12
Importance-FP	1	53.0 ± 1.1	11.5 ± 7.4	55.9	27
CBR-FP	100	52.4 ± 2.3	11.4 ± 10.2	55.2	9
	1000	55.9 ± 1.1	15.3 ± 5.4	58.5	27
Hamming-FP	100	49.2 ± 4.2	2.2 ± 3.2	53.4	1
	1000	47.0 ± 5.3	3.7 ± 2.8	54.2	7
SC-FP	100	32.4 ± 7.8	37.6 ± 15.4	51.4	17
	1000	32.2 ± 6.8	35.3 ± 12.2	47.3	22

(a) Results for unbiased random sampling.

Approach	Iters.	Avg. Accuracy	Feats. Pruned	Max Accuracy	Feats. Pruned
0-FP	1	50.8	0	50.8	0
Random-FP	100	53.7 ± 1.7	12.7 ± 9.8	57.6	22
	1000	56.4 ± 1.0	14.7 ± 5.9	58.5	18
Importance-FP	1	52.9 ± 1.1	12.6 ± 7.3	56.7	23
CBR-FP	100	54.1 ± 1.5	12.0 ± 8.1	57.6	15
	1000	56.9 ± 1.6	14.2 ± 7.2	61.8	9
Hamming-FP	100	47.8 ± 5.5	2.2 ± 2.3	53.5	1
	1000	45.2 ± 6.2	3.6 ± 2.2	54.2	2
SC-FP	100	32.9 ± 8.6	47.3 ± 17.4	51.5	18
	1000	33.7 ± 9.2	40.9 ± 16.4	57.5	27

(b) Results using SHAP values to bias sampling.

sizes (as does Importance-FP). This suggests that if small feature sets facilitate explanation, both feature set size and accuracy should be considered in choosing a pruning method.

3.2. SHAP biases enable subtle accuracy improvements

Using SHAP values with CBR is challenging, because SHAP values imply ordered pruning, while CBR favors comparing whole prunings. However, using SHAP values to assign relative probabilities to features for pruning selection appears to be effective and to act as a stabilizing force—increasing overall and/or maximum accuracy values in several methods (esp. CBR-FP).

3.3. Hamming-FP and SC-FP can offer interpretability for the pruning process for developers, though they underperform in terms of accuracy

Compared with Random-FP, CBR-FP and Importance-FP lead to similar/slightly better accuracy improvements. By contrast, Hamming-FP and SC-FP appear much less effective. Both methods

estimate model performance using proxy classification accuracy values, making them potentially valuable to developers, but for these experiments, they are less successful on average than accuracy-based approaches. Further research is necessary to help clarify ways in which Hamming-FP and SC-FP might be more useful/applicable.

4. Conclusions and Future Directions

We have presented and evaluated multiple potential methods for integrating explanatory information and CBR for feature pruning to improve performance of an artificial neural network, testing those methods on the provided Psychology Prediction dataset. Of these, using CBR to mediate random feature pruning selection weighted using absolute SHAP values appears to yield the highest model accuracy, with Random-FP providing strong performance as well. This provides some support for the generality of the HypoCBR [3] approach on which our approach was based. Both methods provide the potential benefit of removing the greatest number of features, which could facilitate explaining similarity between cases. Future work could investigate using weighted feature values for CBR and/or for weighted Hamming distance calculations, along with more detailed analysis of the experimental algorithms (esp. SC-FP).

Acknowledgments

This work was funded by the US Department of Defense (Contract W52P1J2093009), and by the Department of the Navy, Office of Naval Research (Award N00014-19-1-2655). We thank Karan Acharya and Lawrence Gates for very helpful discussions.

References

- [1] M. G. Orozco-del Castillo, E. C. Orozco-del Castillo, E. Brito-Borges, C. Bermejo-Sabbagh, N. Cuevas-Cuevas, An artificial neural network for depression screening and questionnaire refinement in undergraduate students, in: M. F. Mata-Rivera, R. Zagal-Flores (Eds.), *Telematics and Computing*, Springer, Cham, 2021, pp. 1–13.
- [2] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, Curran, 2017, pp. 4765–4774.
- [3] M. Hoffmann, R. Bergmann, Improving automated hyperparameter optimization with case-based reasoning, in: *Case-Based Reasoning Research and Development*, ICCBR 2022, Springer, 2022, pp. 273–288. In press.
- [4] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, in: *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, AAAI, 2021, pp. 11575–11585.