Counterfactual, Contrastive, and Hierarchical Explanations with Contextual Importance and Utility

Kary Främling[1,2(✉)]

[1] Computing Science, Umeå University, 901 87 Umeå, Sweden
kary.framling@cs.umu.se
[2] Department of Industrial Engineering and Management, Aalto University, Maarintie 8, 00076 Aalto, Finland
kary.framling@aalto.fi

Abstract. Contextual Importance and Utility (CIU) is a model-agnostic method for post-hoc explanation of prediction outcomes. In this paper we describe and show new functionality in the R implementation of CIU for tabular data. Much of that functionality is specific to CIU and goes beyond the current state of the art.

Keywords: Contextual Importance and Utility · Explainable AI · Open source · Counterfactual · Contrastive

1 Introduction

Contextual Importance and Utility (CIU) was presented by Kary Främling in 1992 [1] for explaining recommendations or outcomes of decision support systems (DSS) in a model-agnostic way. CIU was presented formally in [2,3] and more recent developments have been presented *e.g.* in [5]. This paper presents new functionality of CIU that is implemented in the R package for tabular data, available at https://github.com/KaryFramling/ciu. An earlier version of the package was presented at the Explainable Agency in Artificial Intelligence Workshop of the AAAI conference in 2021 [4].

CIU has a different mathematical foundation than the state-of-the-art XAI methods SHAP and LIME. CIU is not limited to "feature influence" and therefore offers richer explanation possibilities than the state-of-the-art methods.

After this Introduction, Sect. 2 resumes the core theory of CIU. Section 3 shows the new functionality, followed by Conclusions that include a brief discussion about CIU versus comparable state-of-the-art XAI methods.

2 Contextual Importance and Utility

Contextual Importance (CI) expresses to what extent modifying the value of one or more **feature(s)** $x_{\{i\}}$ can affect the output value y_j (or rather the *output utility* $u_j(y_j)$). CI is expressed formally as:

The work is partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

D. Calvaresi et al. (Eds.): EXTRAAMAS 2023, LNAI 14127, pp. 180–184, 2023.
https://doi.org/10.1007/978-3-031-40878-6_16

$$CI_j(c,\{i\},\{I\}) = \frac{umax_j(c,\{i\}) - umin_j(c,\{i\})}{umax_j(c,\{I\}) - umin_j(c,\{I\})}, \tag{1}$$

where c is the studied context/instance, $\{i\} \subseteq \{I\}$ and $\{I\} \subseteq \{1,\ldots,n\}$ and n is the number of features. $umin_j$ and $umax_j$ are the minimal and maximal output utility values that can be achieved by varying the value(s) of feature(s) $x_{\{i\}}$ while keeping all other feature values at those of c.

In classification tasks we have $u_j(y_j) = y_j \in [0,1]$ and for regression tasks where $u_j(y_j) = Ay_j + b$ (which applies to most regression tasks) we can write:

$$CI_j(c,\{i\},\{I\}) = \frac{ymax_j(c,\{i\}) - ymin_j(c,\{i\})}{ymax_j(c,\{I\}) - ymin_j(c,\{I\})}, \tag{2}$$

Contextual Utility (CU) expresses to what extent the current **value(s)** of given feature(s) contribute to obtaining a high output utility u_j. CU is expressed formally as:

$$CU_j(c,\{i\}) = \frac{u_j(c) - umin_j(c,\{i\})}{umax_j(c,\{i\}) - umin_j(c,\{i\})} \tag{3}$$

When $u_j(y_j) = Ay_j + b$, this can again be written as:

$$CU_j(c,\{i\}) = \left| \frac{y_j(c) - yumin_j(c,\{i\})}{ymax_j(c,\{i\}) - ymin_j(c,\{i\})} \right|, \tag{4}$$

where $yumin = ymin$ if A is positive and $yumin = ymax$ if A is negative.

Contextual influence expresses how much feature(s) influence the output value (utility) relative to a *reference value* or *baseline*, here denoted $neutral.CU \in [0,1]$. Contextual influence is conceptually similar to Shapley value and other additive feature attribution methods. Formally, Contextual influence is:

$$\phi = CI \times (CU - neutral.CU) \tag{5}$$

where "$_j(c,\{i\},\{I\})$" has been omitted for easier readability.

It is worth noting that CI and CU are values in the range $[0,1]$ by definition, which makes it possible to assess whether a value is high or low. Contextual influence also has a maximal amplitude of one, where the range is $[-neutral.CU, 1 - neutral.CU]$. CIU calculations require identifying $ymin_j$ and $ymax_j$ values, which can be done in many ways. The approach used for the moment is described in [4] and is omitted here due to space constraints.

All CIU equations apply to each feature separately as well as to coalitions of features $\{i\}$ versus other coalitions of features $\{I\}$, where $\{i\} \subseteq \{I\}$ and $\{I\} \subseteq \{1,\ldots,n\}$. Such coalitions can be used to form *Intermediate Concepts*, which name a given set of inputs $\{i\}$ or $\{I\}$. Such Intermediate Concepts make it possible to define arbitrary explanation vocabularies with abstraction levels that can be adapted to the target user.

3 New Explanation Functionality with CIU

The source code for producing the results shown here is published at https://github.com/KaryFramling/EXTRAAMAS2023. The R package source code is available at https://github.com/KaryFramling/ciu and on CRAN.

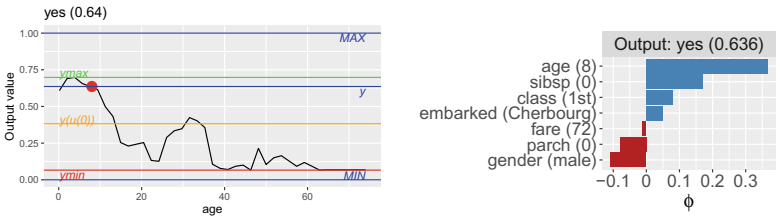

Fig. 1. Left: Generated illustration of CIU calculations. Right: Contextual influence barplot explanation for "Johnny D".

To begin, we use the Titanic data set, a Random Forest model and an instance "Johnny D", as in https://ema.drwhy.ai. "Johnny D" is an 8-year old boy that travels alone. The model predicts a survival probability of 63.6%. 63.6% is good compared to the average 32.5%, which is what we want to explain. Figure 1 illustrates how CI, CU and Contextual influence is calculated for the feature "age" and a Contextual influence plot for "Johnny D" with $neutral.CU = 0.325$.

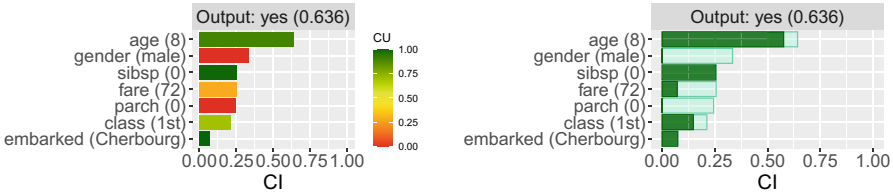

Fig. 2. Left: "Old" visualisation of CI and CU, with CU illustrated using colors. Right: New visualisation that answers more exactly the counterfactual "what-if" question.

Counterfactual Explanations answer the question "What if?". Figure 2 shows an older visualisation with CI as the bar length and CU illustrated with a color. The new visualisation illustrates CI as a transparent bar and CU as a solid bar. When $CU = 0$ (worst possible value), the solid bar has length zero. When $CU = 1$ (best possible value), the solid bar covers the transparent bar. This is called "counterfactual" because it indicates what feature(s) have the greatest potential to improve the result. In Fig. 2 we can see that being accompanied by at least one parent (feature "parch") could increase the probability of survival.

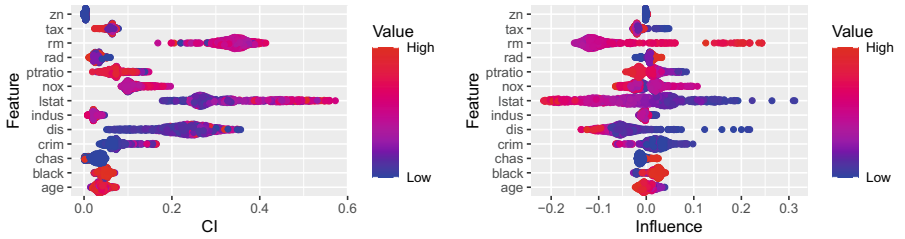

Fig. 3. Beeswarm visualisation of CI and Contextual influence for Boston data set.

Beeswarm Visualisation. Beeswarms give an overview of an entire data set by showing CI/CU/influence values of every feature and every instance. As in https://github.com/slundberg/shap, we use the Boston data set and a Gradient Boosting model. The dot color in Fig. 3 represents the feature value. The CI beeswarm in Fig. 3 reveals for example that the higher the value of "lstat" (% lower status of the population), the higher is the CI (contextual/instance-specific importance) of "lstat". The influence plot reveals that a high "lstat" value lowers the predicted home price and is nearly identical to the one produced for Shapley values. We use $neutral.CU = 0.390$, which corresponds to the average price so the reference value is the same as for the Shapley value.

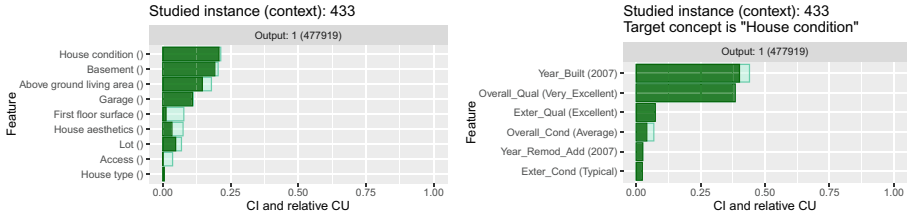

Fig. 4. Left: Top-level explanation for why Ames instance 433 is expensive. Right: Detailed explanation for Intermediate Concept "House condition".

Intermediate Concepts. Ames housing is a data set with 2930 houses described by 81 features. A gradient boosting model was trained to predict the sale price based on the 80 other features. With 80 features a "classical" bar plot explanation becomes unreadable. Furthermore, many features are strongly correlated, which causes misleading explanations because individual features have a small importance, whereas the joint importance can be significant. Intermediate Concepts solve these challenges, as illustrated in Fig. 4 that shows the top-level explanation and an explanation for one of the Intermediate Concepts for an expensive house. Here, the vocabulary has been constructed based on common-sense knowledge about houses but it could even be provided by the explainee.

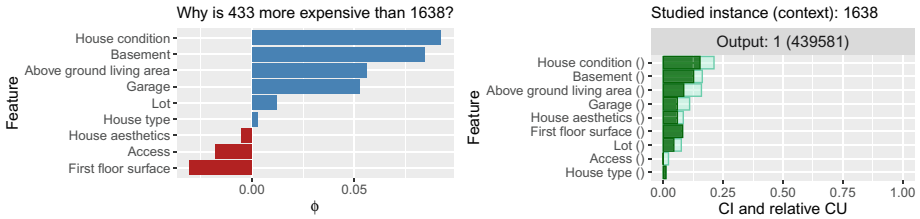

Fig. 5. Left: Contrastive "Why?" explanation for two expensive Ames houses. Right: Top-level counterfactual explanation for Ames instance 1638.

Contrastive Explanations. Contrastive explanations answer questions such as "Why alternative A rather than B" or "Why not alternative B rather than A". Any value in the range $[0, 1]$ can be used for *neutral.CU* in Eq. 5, including CU values of an instance to compare with. Figure 5 shows a contrastive explanation for why Ames instance #433 ($477919, see Fig. 4) is predicted to be more expensive than instance #1638 ($439581). Contrastive values are in the range $[-1, 1]$ by definition, so the differences between the compared instances in Fig. 5 are small.

4 Conclusion

CIU enables explanations that are not possible or available with current state-of-the-art methods. Notably, Shapley value and LIME are limited to "influence" values only. Even for influence values, Contextual influence offers multiple advantages such as a known maximal range and adjustable reference value. However, the emphasis of the paper is to show how CI together with CU can provide counterfactual explanations and give a deeper understanding of the model behaviour in general, including the possibility to produce contrastive explanations.

References

1. Främling, K.: Les réseaux de neurones comme outils d'aide à la décision floue. D.E.A. thesis, INSA de Lyon (1992)
2. Främling, K.: Explaining results of neural networks by contextual importance and utility. In: Andrews, R., Diederich, J. (eds.) Rules and networks: Proceedings of Rule Extraction from Trained Artificial Neural Networks Workshop, AISB'96 Conference. Brighton, UK (1–2 April 1996)
3. Främling, K.: Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère. Phd thesis, INSA de Lyon (Mar 1996)
4. Främling, K.: Contextual Importance and Utility in R: the 'CIU' Package. In: Madumal, P., Tulli, S., Weber, R., Aha, D. (eds.) Proceedings of 1st Workshop on Explainable Agency in Artificial Intelligence Workshop, 35th AAAI Conference on Artificial Intelligence, pp. 110–114 (2021)
5. Främling, K.: Contextual importance and utility: a theoretical foundation. In: Long, G., Yu, X., Wang, S. (eds.) AI 2022. LNCS (LNAI), vol. 13151, pp. 117–128. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97546-3_10