Explaining your Neighbourhood: A CBR Approach for Explaining Black-Box Models

Betül Bayrak*, Paola Marin-Veites and Kerstin Bach

Norwegian University of Science and Technology (NTNU), Høgskoleringen 1, Trondheim, 7034, Norway

Abstract

In this work, we propose a multi-agent Case-Based Reasoning system (MA-CBR system) approach for explaining a Black-box model that provides a 3-class depression classification. We show how to develop a CBR system that can accompany a Black-box model to provide similar supportive or contrastive cases and how domain knowledge influences the performance of a CBR system. Eventually we provide a concept and showcase for a MA-CBR system that can help to explain the results of a Black-box model.

1. Introduction

As part of the International Conference on Case-Based Reasoning of 2022 (ICCBR-2022), a set of XCBR Challenges were proposed. In this paper, we chose to work with the Psychology prediction dataset. The task is to provide an explanation for a Blackbox model that provides a 3-class depression classification, with class 1 being low risk of depression score and class 3 being high risk. The provided data, derived from [1], consists of a dataset with 104 answered questionnaires as well as expected answers questionnaire designed by a medical health professional that assess if the patient is at risk of depression.

In our proposed solution, we developed a multiagent CBR system (MA-CBR system) to create explanations for a Black-box model's classification. We suggest the MA-CBR system that supports or contrasts the classification of the Black-box model. To develop the CBR system we use Shapley values to obtain relevant questions to define the weights of our global similarity measure. We also extend the CBR case representation with an attribute representing domain knowledge. The core idea can be compared to a Twin CBR system as introduced by [2].

To test our hypothesis, we conduct three experiements. In the first experiment, we compare the classification results for the CBR system and Black-box model. In the second experiment, we add domain knowledge to the case representation and assess its influence in the CBR system. And in the third experiment, we show how the CBR system can provide supportive explanations on the class classification from the Black-box model or contrast it providing the differences explanation.

ICCBR XAI'22: Explainable AI Challenge at ICCBR-2022, September, 2022, Nancy, France

*Corresponding author.

betul.bayrak@ntnu.no (B. Bayrak); paola.m.veites@ntnu.no (P. Marin-Veites); kerstin.bach@ntnu.no (K. Bach)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Explanation Methodology

2.1. Pre-Processing

We used the black-box classification model, a multi-layer-perceptron provided by the organizers of the challenge, to calculate Shapley values. To work with stable Shapley values we used the training data (83 examples) and calculated the Shapley value for each fold. Based on these folds, we calculated the mean for each attribute in the respective 3 classes. The resulting mean Shapley values are the weights of our global similarity functions.

2.2. CBR System for Explaining Black-box models

The concept of how we created the CBR system is shown in Figure 1. We modeled the attributes in the myCBR workbench. All questions were included as integer attributes with a value range of [0, 1]. Every attribute has the same, binary local similairy measure where a matching answer is 1 and a mismatch is 0. In this dataset we only had binary attributes, but for more complex attributes the local similarity measures can be modeled as introduced by [3]. The weights for the global similarity measure are based on the Shapley values derived from the Black-box model.

Further, we introduced a new attibute called *Matched* as an integer with a value range from [0, 103]. That attribute describes the number of matches between a user's answer and an experts evaluation on how questions from a depressed person would be answered. The similarity measure was created using the methods described by [3, 4]. In our approach we will use the *Matched* attribute to show how domain knowledge can be incorporated in the CBR system.

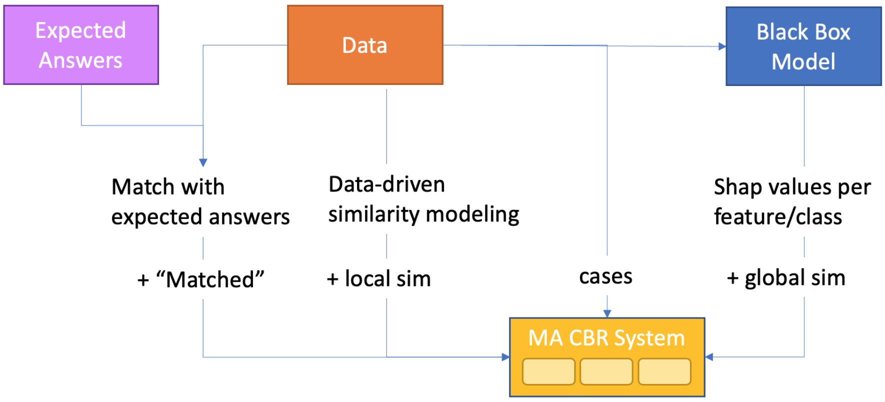
Expected Answers
Data
Black Box Model
Match with expected answers
Data-driven similarity modeling
Shap values per feature/class
+ "Matched"
+ local sim
cases
+ global sim
MA CBR System

Figure 1: Modeling approach for the CBR Explainer

The myCBR project file contains three casebases and three amalgamation functions, each representing one class of the Black-box model prediction range. Each of these configuration is referred to as a CBR agent. Once the MA-CBR system is created, we use it to explain the results of the Black-box model as shown in Figure 2. First a query is created from the incoming data. This query is sent to the Black-box model which provides the classification for the dataset. Next, the query is extended with the *Matched* attrtribute and sent to the MA-CBR system. The MA-CBR system will retrieve the most similar cases from each CBR agent and provide them as explanation candidates. We propose that the retrieval results for the CBR agent that matches the Black-box model class prediction are handeled as *Supportive* examples while the others are *Contrastive* examples.

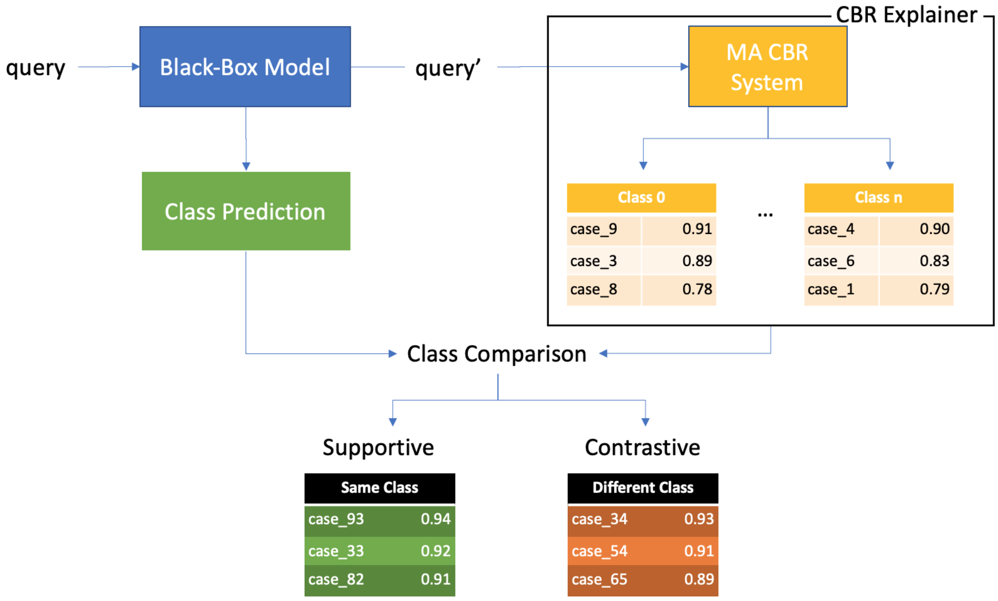

Figure 2: Explanation system using the CBR system

Supportive Examples are provided if both systems agree on the classification while contrastives are suggested when the CBR Explainer suggests a different class.

2.3. Implementation

The presented work is implemented using the myCBR workbench for modeling the skeleton CBR system. Further we use the mcbr Rest API to deploy the global similarity measure, cases and deploy the CBR agents. All interaction with the CBR system is implemented in Python using the *requests* librabry.

While working on developing the CBR system we extended the myCBR Rest API allowing us to add Integer attributes through the API, changing the amalgamation functions, and adding and deleting instances from a concept. These extensions allow us to deploy a skeleton CBR system and change the system as needed for experimentation through the myCBR Rest API via Python. The updated API has been made available on GitHub[1].

3. Experiements and Results

In the provided experiments we will show how we developed the Black-box model and the MA-CBR system using the given dataset. We will provide experiments that support the decisions

[1] https://github.com/ntnu-ai-lab/mycbr-rest

we made to develop the CBR explainer.

3.1. Experiment I

The goal of the first experiment is to test the CBR system's and the blackbox model's performance using the given dataset. Therefore we build each system using the raw and oversampled data. The reason for exploring both approaches is that in the given Black-box model oversampling is used to increase the quality of the model, while for a CBR system, we prefer real-world rather than simulated cases.

In this experiements we ran a set of experiments that do 100 different 80/20 splits of the training and test data. The dataset is very small and provides huge variation in the performance of the Black-box model and CBR systems depending on the splits. To have comparable results, we used this experiment to find the most suitable setup for experiment II and III.

3.1.1. Results

Based on Figure 3 we decided that a Black-box model using oversampled data performs best, while oversampling has little influence on the CBR system.

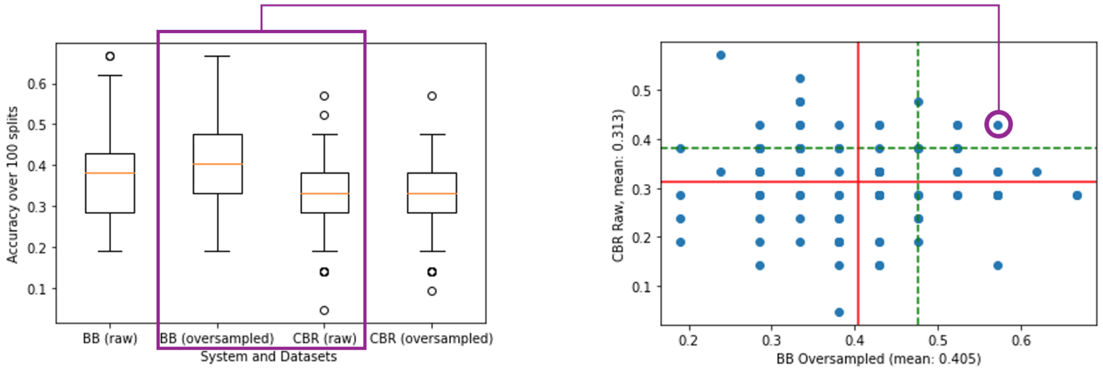

Figure 3: Experiement 1 results: Boxplot of the accuracy (y-axis) over 100 different splits for the Black-box model and the CBR system for raw and oversampled (os) data.

To select a combination, we picked a split where the CBR system and BB system performed best (see the plot over all runs in the right hand plot of Figure 3).

3.2. Experiment II

In this set of experiments, we aim at improving the performance of the CBR system. From Figure 3, we see that the overall accuracy is lower that the Black-box model's one. To provide good similar cases to support or contrast the Black-box model's classification, we suggest to incorporate domain knowledge into the CBR system. Given the nature of the challenge, no domain expert was available. Therefore we used the expected answer file from the provided documents and included an attribute, *Matched*, that represents the number of questions in which the query case matches the expert's opinion on whether a test subject has depression.

We integrated the *Matched* in the case representation. For the next experiments, we compared the similarity scores of the CBR system with and without the *Matched* attribute. To find the best weight in the global similarity measure, we picked the maximum shap-based weight and then ”boosted” the weight to increase it's importance in the classification. During the experiments we tested different ”boosters” to find the best weight for the global similarity measure. We used the overall accuracy of the CBR system to decide on the booster.

3.2.1. Results

The accuracy of both CBR systems differs greatly with the added domain knowledge indicating the importance of such information that is not included in the pure dataset. The CBR system can easily include such knowledge from experts. In Figure 4, we show the confusion matrix for three systems. The first matrix are the results of the Black-box model and the second of the CBR system developed in this experiment.

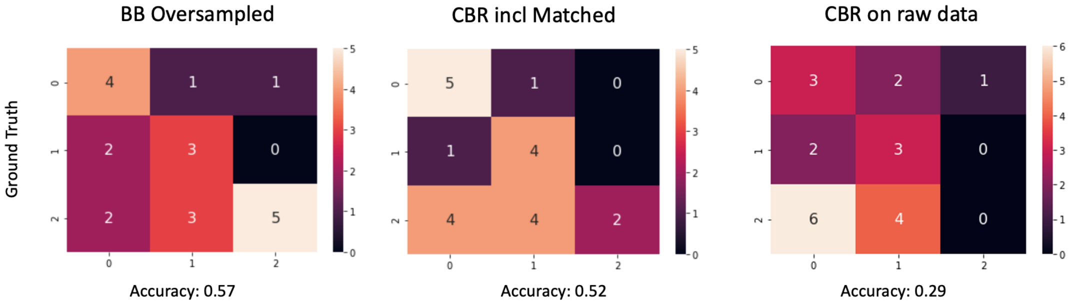

Figure 4: Experiement 2 results: Confusion Matrices for the Black-box model and the two CBR systems. The CBR incl matched in the results of the experiment II optimization and the CBR raw shows the performance of the CBR system on the raw data only.

Comparing the Black-box model and the CBR system that includes domain knowledge, we can see that two classes (class 0 and class 1) the CBR system performs better while for class 2 it performs worse. We assume that such information would be available to the CBR Explainer and can therefore be included when decisions are explained.

3.3. Experimental Setup III

In the final experiment we show how the MA-CBR system can be used in conjunction with the Black-box model. We therefore implemented the concept described in Figure 2 to provide supportive and contrastive cases to explain a decision by the Black-box model.

3.3.1. Results

In our implementation, we use the test data and provide the information whether the comparsion leads to a supportive or a contrastive page. Further we show the matching and differing attributes along with the case ids of the most similar cases. We believe this information can be used to explain results to the user.

4. Conclusion

With our work, we show that a CBR system or a MA-CBR system can be developed using the meta information coming from a Black-box model. In particular, we use Shapley values and add domain knowledge to create a CBR explainer that provides similar cases to the query to either support or contrast a Black-box model's decision.

We hope to show that the myCBR Rest API is a useful tool to develop such application and promote CBR as methodology for building model-agnostic explainers.

The provided dataset, however, has some limitations, first and foremost are relativly small datapoints for the large number of features available. Further, the creation of the dataset and concept behind is not provided, which makes it hard to find suitable domain knowledge to include. During our work we aimed at reducing the number of features using feature importance and shap methods. However, none of them provided a smaller set of features without losing model performance.

Having no access to experts made it challenging for us to understand *what* should be explained. We showed how nearest neighbor cases can be found. To create an XAI system, we would now need to work with users of a system to understand the needs of an explanation before a user interface can be developed.

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 International Publishing, Cham, 2021, pp. 1–13.

[2] M. T. Keane, E. M. Kenny, The twin-system approach as one generic solution for xai: An overview of ann-cbr twins for explaining deep learning, arXiv preprint arXiv:1905.08069 (2019).

[3] D. Verma, K. Bach, P. J. Mork, Similarity measure development for case-based reasoning–a data-driven approach, in: Symposium of the Norwegian AI Society, Springer, 2019, pp. 143–148.

[4] P. Marín-Veites, K. Bach, Explaining cbr systems through retrieval and similarity measure visualizations: A case study, in: M. T. Keane, N. Wiratunga (Eds.), Case-Based Reasoning Research and Development, Springer, 2022, pp. 111–124.