IREX: A reusable process for the iterative refinement and explanation of classification models.

Cristian E. Sosa-Espadas[1,2], Manuel Cetina-Aguilar[1,2], Jose A. Soladrero[1,2],
Jesus M. Darias[3], Esteban E. Brito-Borges[1,2], Nora L. Cuevas-Cuevas[1,2] and
Mauricio G. Orozco-del-Castillo[1,2,*]

[1] *Tecnológico Nacional de México/IT de Mérida, Department of Systems and Computing, Merida, Mexico*
[2] *AAAIMX Student Chapter at Yucatan, Mexico, Association for the Advancement of Artificial Intelligence, Mexico*
[3] *Department of Software Engineering and Artificial Intelligence, Universidad Complutense de Madrid, Spain*

Abstract

This paper presents **IREX**: a reusable method for the **I**terative **R**efinement and **EX**planation of classification models. It has been designed for domain-expert users –without machine learning skills– that need to understand and improve classification models. This way, it only requires the expected classification outcomes given by a domain expert. IREX proposes a smart combination of XAI methods that identifies potential inconsistencies in the model, explaining the causes to the user. Following a cycle analogous to CBR, a set of candidate anomalous variables are identified (retrieved) and proposed for its revision by an expert user. Once revision is confirmed, the model is purged and retrained for further optimization.

This is a novel process where explanation methods are not only applied to explore a black-box model, but also to detect which input variables led to misclassifications, proposing and explaining their negative impact to the domain-expert user. We propose an automatic evaluation approach based on computing the number of anomalous input variables that the expert was able to identify and its comparison to the evolution of the classification model's performance. Then we apply this evaluation method to demonstrate the performance of the proposal on the given dataset. Finally, we provide a reusable implementation that can be directly applied to other classification models and domains.

1. Method Description

The goal of this development is to provide an iterative explanation and refinement strategy based on the combination of several XAI methods. Although this method –named IREX– is generic, reusable and can be applied to other classification models, we illustrate its functionality through its application to the provided dataset [1]. This way, we instantiate it on a screening tool that allows the early identification of students with high risk of depression using an Artificial Neural Network (ANN) model applied over that dataset.

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

[*] Corresponding author.

mauricio.orozco@itmerida.edu.mx (M. G. Orozco-del-Castillo)
0000-0002-0933-5152 (C. E. Sosa-Espadas); 0000-0002-2423-0711 (M. Cetina-Aguilar); 0000-0001-5864-275X (J. A. Soladrero); 0000-0002-2173-1422 (J. M. Darias); 0000-0002-4257-944X (E. E. Brito-Borges); 0000-0002-5477-6176 (N. L. Cuevas-Cuevas); 0000-0001-5793-6449 (M. G. Orozco-del-Castillo)

CEUR Workshop Proceedings (CEUR-WS.org)

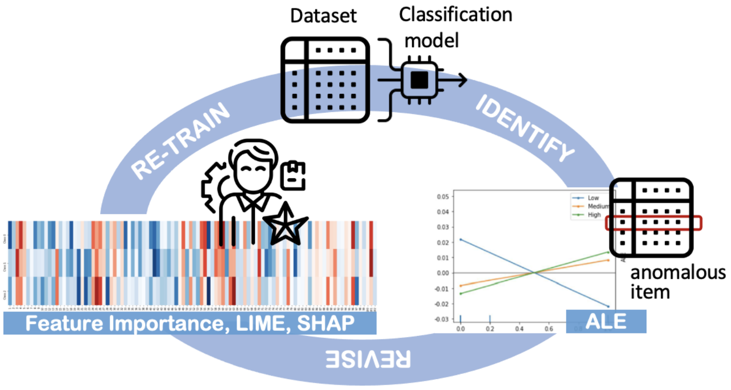

Figure 1: Global schema of the proposed iterative explanation and refinement method

1.1. Personas

The approach being presented is mainly addressed to the domain expert –without machine learning (ML) skills– that needs to improve and understand a classification model. This way, the IREX method allows the domain expert to identify which input variables led to anomalous classifications, understand its impact, and refine the model.

1.2. Explanation Strategy

The proposed method is illustrated to the concrete use case of the refinement of a depression screening tool based on an ANN. This tool is based on a questionnaire with several items (the input variables) where the corresponding answers feed a classification model. Thus, this use case exemplifies the IREX method by identifying the anomalous items in the questionnaire through explanation methods that provide insights of the prediction model.

The IREX process follows a cycle analogous to CBR, where a set of candidate input variables are identified (retrieved) and proposed for its revision by an expert user. Finally, once anomalous variables are confirmed, they are used to modify the prediction model. This way, we propose an iterative process consisting on the following steps:

1. First, the classification model is introspected with explanation methods to identify those variables that may lead to a lower performance according to the behaviour expected by the expert user.
2. The expert will confirm or discard the abnormal behaviour of these variables through a deeper analysis supported by additional introspective XAI methods.
3. Those variables confirmed by the expert are removed from the dataset, and a new model is trained. Continuing to the next iteration from step 1.

This process is illustrated in Figure 1. The dataset used on this project corresponds to the one described in [1]. It consists of the answers from 157 users who answered a screening questionnaire of 102 items. The output variable estimated by the model –an ANN– is the "potential depression score" (PDS) codified as three classes: Low, Medium, High.

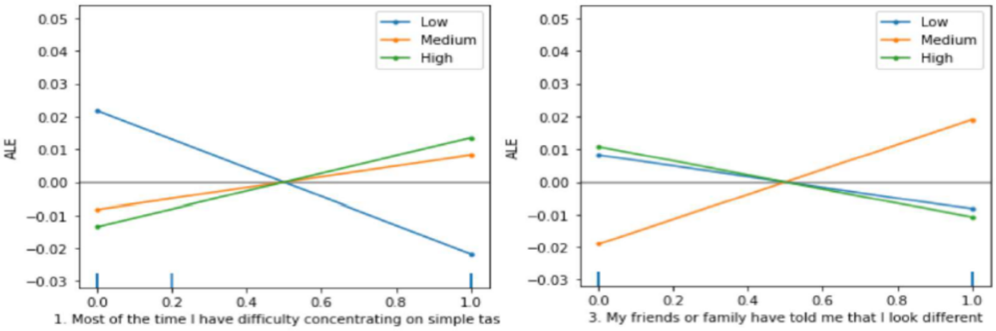

Figure 2: Example using ALE to identify anomalous input variables.

The first step for each iteration is the identification of the anomalous items of the questionnaire through the Accumulated Local Effects (ALE) explanation method. When applied to our model, ALE returns a plot for each questionnaire item. Each plot contains three lines, representing –on average– the influence of that item for the classification as High, Medium, and Low classes. The x-axis represents the possible answers (0-false, and 1-true) and the y-axis corresponds to the influence of the probability to be classified as each class. Figure 2 (left) shows one of these plots, where we can observe, for example, that answering true (1 in x-axis) to questionnaire item 1 decreases significantly the possibility to be classified as low PDS, whereas it increases (with less influence) the classification as medium or high PDS. If we read the question under the plot, we can intuitively confirm that the ANN model is working properly, since it corresponds to the expected answer defined by the expert. To detect anomalous input variables we only need to identify the opposite pattern. The expected answer to item 3 *"My friends have told me that I look different"* to diagnose high PDS according to the psychologist is 1. However, the ALE plot (Figure 2, right) shows that the behavior of the ANN model is the other way around: answer 1 contributes to be classified as medium PDS but decreases the possibility of belonging to classes low and high, that is a very anomalous pattern.

At this point, these items (model's input variables) should be considered as "potentially anomalous variables" until the expert confirms that the behavior of the model is correct or not. This revision process (similar to the one performed by the CBR cycle) can be really complex for users that may not have any kind of skill on ML. Therefore, we propose the use of additional introspective XAI methods to support this task. Concretely, we provide support to this decision-making process through several explanation methods such as Feature Importance, LIME and SHAP. These methods are able to measure the impact of each item in the accuracy of the model. Here, we propose a novel approach to visualize this effect using heatmaps, as presented in Figure 3 where we use a blue-color scale to indicate the impact of variables behaving correctly according to the "expected answers datasheet" and a red-color scale for inconsistent variables. These heatmaps can be used by the expert to decide whether a variable is actually an anomaly of the ANN model or it describes a certain pattern that makes sense. Once the domain expert confirms the anomalous input variables, they are removed from the model that is retrained, and the cycle starts over. Then, the whole process finishes when no potentially anomalous variables are identified by the ALE method.

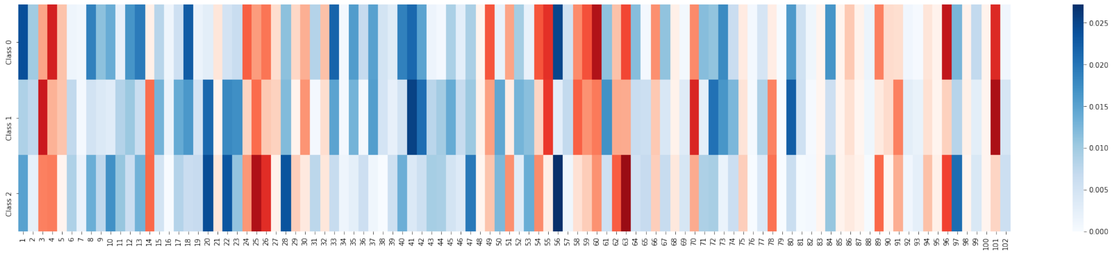

Figure 3: Heatmap reflecting the impact of every item in the model according to SHAP. Red color scale is used for inconsistent items.

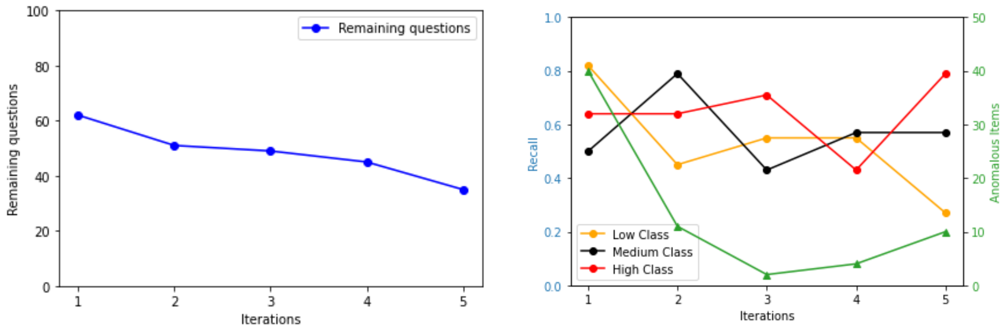

Figure 4: (Left) Global accuracy and number of anomalous questions identified on each iteration. (Right) Number of questions used as inputs on each iteration.

1.3. Evaluation method and Performance

The evaluation of our explanation and model refinement method can be performed automatically through the comparison of the number of anomalous input variables identified by the expert user and the evolution of the model's performance. This way, we can measure if the explanation methods allowed the expert to identify the anomalous variables and improve the model.

For the concrete dataset used provided by the challenge, we present the evaluation results in Figure 4. The number of inputs used to train the classification model on each iteration is presented in Figure 4 (left). From 102 original items, our refinement process reduced that number down to approximately 40 questions.

Since the main objective of removing questionnaire items was to improve the identification of people in the High PDS class, it is also useful to analyse the recall metrics. Corresponding results are presented in Figure 4 (right). Although recall decreases for the Medium PDS and Low PDS classes, the High PDS presents an important raise, which indicates that the ANN has achieved a better ability to detect individuals with a high presence of depression symptoms, which is the goal of this concrete use case.

```
# Load required external libraries
loadImports()
# Load dataset
data   = loadData()
# Load (or train) classification model
model  = loadModel()
# Load expected classifications given by domain expert
expertKnowledge = loadExpectedOutputs();
# Configuration parameters (with default values)
config = configExplanationParameters()
# Main explanation method
explain(data, model, expertKnowledge, config)
```

Figure 5: IREX reusable source code. Main methods to load data, model and expert knowledge; then configure and execute the explanation process.

2. Benefits and Impact

We propose a *novel process* where XAI methods are not only applied to explore a black-box model, but also to detect which input variables led to errors in the classification model, explaining their negative impact to the expert user. Our explanation and refinement process is based on a smart combination of explanation methods such as ALE, Feature Importance, LIME or SHARP, following an iterative cycle analogous to CBR where a set of candidate variables are identified (retrieved) and proposed for its revision by an expert user. Finally, once anomalous variables are confirmed (retained), the prediction model is improved.

3. Reusability and source code

The provided explanation and refinement method is completely reusable in other models and domains, as it only requires the expected model's classifications from an expert user. The provided source code, has been designed to support its reusability and integration into explanation libraries or APIs. As illustrated in Listing 5, it isolates the domain dependent data, model and expected classifications. Then it provides a configuration method with default values. And, finally, the explanation process itself, encapsulated as an only executable method.

Github (full source code):	https://github.com/Manuel080800/IREX.git
Colab (online execution):	https://colab.research.google.com/drive/1IhmBtRnstL8SthhkECfydPOFuS5jxgLi
Jupyter Notebook (pdf):	https://github.com/Manuel080800/IREX/raw/master/Jupyter%20Notebook
Full execution video	https://aaaimx.org/irex/

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: Telematics and Computing. WITCOM 2021., volume 2, Springer Nature Switzerland AG 2021, 2021, pp. 1–13. doi:10.1007/978-3-030-89586-0_1.