Towards Linking Local and Global Explanations for AI Assessments with Concept Explanation Clusters

Elena Haedecke[1,2], **Maram Akila**[2, 3], **Laura von Rueden**[2, 4]
[1]University of Bonn, Germany,
[2]Fraunhofer IAIS, Sankt Augustin, Germany
[3]Lamarr Institute, Sankt Augustin, Germany,
[4]Hochschule für Technik Stuttgart, Germany
elena.haedecke@fraunhofer.iais.de

Abstract

Understanding the inner workings of artificial intelligence (AI) systems is important both in the light of regulation (e.g., the EU AI Act), but also to uncover hidden weaknesses. Although local and global explanation methods can support this, a scalable and human-centered combination is required to combine the detail of the former with the latter's efficiency. Therefore, we present our method *concept explanation clusters* as a step towards explaining (sub-)strategies of the model through human-understandable concepts by identifying clusters in the input data while accounting for model predictions by local explanations. In this way, all the benefits of local explanations can be retained while allowing contextualisation on a larger (i.e., data-global) scale.

Introduction

Deep neural networks (DNN) will be used in many areas of application in the future, such as the detection of diseases (Shoeibi et. al 2024). However, the non-transparent nature of these systems poses a hurdle to their widespread use, especially in safety-critical areas. In general, regulations (e.g., the EU AI Act), standards and ethical aspects must be taken into account for the trustworthy use of AI systems. Both internal and external AI assessments are therefore necessary for organisations. This makes it particularly important for developers and assessors to understand the inner workings of the AI models used in order to uncover hidden vulnerabilities, such as learned spurious correlations, mitigate them and thus validate their safe application (Poretschkin et. al 2023).

Explainable AI (XAI) methods are commonly used to mitigate the opaqueness of the highly complex DNN models. A distinction is usually made between local methods for explaining individual decisions and global methods for explaining the entire model behavior. However, inherent drawbacks are: (1) local XAI methods are prone to misinterpretations due to the lack of context, and (2) overwhelm human analysts by impractical case-by-case explanations (Nguyen et. al 2021), and (3) global XAI methods have lower explanatory power due to their human-oriented simplicity.

Currently, a human-centered linkage between local and global explanations that allows an efficient yet in-depth and understandable model evaluation is still missing. Promising approaches tackling this challenge can be found in the area of concept-based explanations, i.e., human-understandable combinations of features representing model prediction strategies. However, these mainly use latent space concepts, while we identify concepts based on a combination of local explanations with input space features. Therefore, we introduce *concept explanation clusters (CEC)* (see Figure 1), which aim to combine aspects of three research areas: (1) XAI (via local explainability), (2) data similarity (inspired by case-based reasoning), and (3) human-centered evaluation (inspired by visual analytics). Following, we present related work of the three inspirational research areas as well as the theoretical methodology of our approach. Further, we provide a first experimental proof of concept on a tabular dataset. Lastly, we discuss results and give an outlook to future work.

Related Work

Our method is inspired by the following three research areas, whose thematic overlap is primarily reflected in the goal of making AI models more understandable for humans.

Local, global, and concept-based XAI methods While the majority of XAI methods has been developed with the goal of creating visually comprehensible and simple explanations, less attention has been paid to the group of developers and model experts, for whom explanations should be designed with the aim of efficient research, exploration and debugging (Biecek and Samek 2024). A subset of XAI methods are concept-based explanations, with TCAV (Kim et. al 2018) as a prominent example using the model's internal concept activation vectors (CAV) and quantifying their importance w.r.t. to user-defined concepts. (Fel et. al 2024) distill three prior concept extraction methods into two fundamental steps: concept extraction and concept importance scoring. Recent approaches, e.g., PCX (Dreyer et. al 2024), leverage layer-wise relevance propagation (LRP) to identify concepts relevant for the model's decision. These approaches mainly use latent space features to identify the NN's concepts. However, this can have the effect that the reference to the actual data can be lost to a certain extent (e.g., because similar concepts are mapped to different representations, or because multiple input data are viewed similarly by

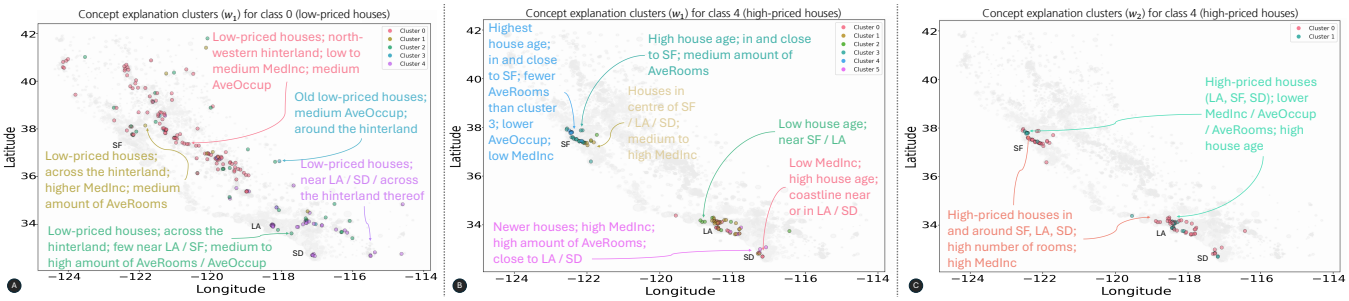

Figure 1: Visualization of the data instances of class 0 (houses with the lowest prices) and class 4 (houses with the highest prices) in a geographical map with the respective concept explanation clusters found (encoded via the color of the data points). As a reference, the entire data set is visualized in the background. For better visibility of the cluster points, the background is displayed in grayscale and with high transparency. The locations of the three big cities San Francisco (SF), Los Angeles (LA), and San Diego (SD) are displayed by their initials. (A): CEC for class 0 that have been identified using weighting function w_1. (C) and (D): A comparison of the CEC for class 4 that have been identified using weighting function w_1 (middle) and w_2 (left). It can be observed that w_1 seems to find more sub-concepts within the data of one class.

the NN and therefore map to the same latent concept). We take inspiration from these approaches, but use input features instead of latent features to overcome this drawback.

Case-based reasoning and explanation-by-example CBR methods employ a distance metric on the training set (the "case-base") to retrieve similar cases to a query-case, e.g., using k-nearest neighbor approaches. In this respect, they do not learn a model and are therefore considered as being naturally transparent, as their reasoning process strongly resembles human behavior in forming conclusions (Leake 2022). To provide explanations-by-example for predictions of NN models (and not only for the training set), the distance metric has to be developed with respect to the model, e.g., by using the learned feature weights of the model (Kenny and Keane 2019). Our method also involves identifying similar cases based on a combination of data and feature importances, but while CBR methods aim to explain only one prediction at a time by showing the user only a few similar cases, we aim to cluster larger groups sharing similar concepts to enable an efficient model understanding.

Human-centered visual analytics A recent field of VA is concerned with VA for human-centered machine learning (ML) (Andrienko et. al 2022). Since DNN models are based on a large amount of data, the research focus here is on efficient integrating of humans into the analysis process of DNNs, while ensuring both the scalability of the analysis task and the necessary level of detail (cp. (Haedecke et. al 2023)), which is also one of our goals.

Methodology

Addressing the missing context of local explanations Our method aims to build a bridge from local explanations to a more efficient and global model understanding, since local explanations in themselves offer no possibility of abstracting from locally occurring feature importances to decisive input feature concepts used in a more global context.

Thus, our method provides context through identifying regional clusters of similar cases, where the similarities are based both on the most important features as well as on the patterns present in the input data, such that the patterns or sub-concepts the model uses for its decisions and which are present within these clusters can efficiently be recognized and understood. Our method therefore involves the following three steps, as also depicted in Figure 2: (1) Generate model predictions and local explanations w.r.t. their predicted class for each instance of the dataset. (2) Compute a weighted distance matrix per class. (3) Cluster the data based on the distance matrix per class. For each of the classes, the found clusters are visualized such that the underlying concepts per cluster can be analysed by the assessor.

In the **first step**, we build a set consisting of the given input data elements, the model predictions for each input, and a local explanation (feature attribution) for each prediction. Using sets per predicted class allows us to identify discrete clusters (and thus subconcepts) per class and thus consider model behaviour for each class separately.

In the **second step**, we combine the information given by input feature values (instead of latent space features) and their corresponding feature importances as computed by a local explanation method. An important part of calculating the distance matrix is weighting the pairwise distances between input feature values of two instances, as here we can influence which instances are more likely to end up in a cluster due to their weighted similarity, and which should be very dissimilar. The main idea for the weighting is that the input feature values should only have a major influence on the distance calculation between instances when they are of high local importance to the model decision. This means that instances should be close to each other if the important features have similar input values (i.e., a small distance), regardless of their possible differences for feature values that have a low importance. In addition, the distances between such input feature values should be increased if their feature importance values are exactly opposite (i.e. high positive and high negative feature importance), even if the input values are similar. We define the weighted distance as $D(i_1, i_2) = \sum_{k=1}^{n} d(feat_{1,k}, feat_{2,k}) \times w(attr_{1,k}, attr_{2,k})$ with the

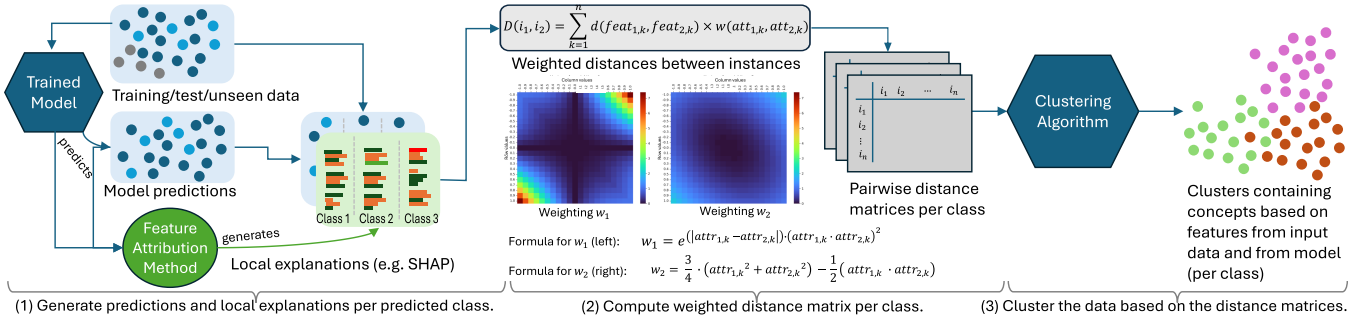

Figure 2: Steps of our method. After generating predictions and local explanations for each class, pairwise distances are computed, weighted w.r.t. the values and differences of their feature importances. Based on the distance matrices, a clustering algorithm identifies the semantic (sub-)concept explanation clusters. The method's process is defined in an agnostic fashion, such that model, data domain, explanation method and clustering algorithm could be adjusted for the use-case at hand.

two instances i_1, i_2, distance function d (e.g., manhattan distance), $feat_{1,k}, attr_{1,k}$ as the k-th input feature and feature attribution values (of n features in total) of i_1, i_2, respectively, and a weighting function w. For this purpose, we have created two weighting functions w_1 and w_2, see the heatmaps in Figure 2, using different approaches to handle opposing and neutral feature attribution. w_1 (left) strongly weights opposing attributions while not taking into account features that are neutral, i.e., not contributing, in either i_1 or i_2. w_2 (right) only neglects features unimportant to both predictions, but less strongly weights opposing features.

Finally, in the **third step**, a clustering algorithm is applied using the respective distance matrices for each class in order to identify the sub-concepts that predominate in them. The clusters found for each class are then visualized in such a way that a user can quickly identify patterns within the input values, the feature importances and the predictions.

Experimental setup We experimentally evaluate our approach on the tabular data set "California Housing" (Kelley Pace and Barry 1997), which originally involves the regression task of predicting house prices based on eight features. For the purpose of identifying dedicated (sub-)concepts per class, we perform equal-width binning into five classes, with prices ascending from bin 0 to 4. We train a xgboost classifier in a k=5-fold on 95% of the data and let the best performing model generate predictions on the 5% hold-out testset. As feature attribution method, we employ SHAP (Lundberg and Lee 2017) to generate local explanations for all testset instances. SHAP values are useful for our approach as they quantify both the positive and the negative importance of features on the local model prediction. As our weighting function is defined for values between -1 and 1, we scale the SHAP values accordingly. Based on the model predictions, we divide the input data and explanations into sets for each of the five classes and compute the weighted distance matrices accordingly for each class, and for each variant of the weighting function.

We employ agglomerative clustering using the distance matrices as input. We evaluate the best hyperparameters according to the Calinski-Harabasz score, which is defined as ratio of the sum of between-cluster dispersion and of within-cluster dispersion and does not require any ground truth.

Experimental Results

The amount of found clusters differs per class c_n and per weighting function: w_1 found 20 clusters (c_0:5, c_1:3, c_2:3, c_3:3, c_4:6) in total with a summarized Calinski Harabasz score of 208.75, while w_2 found 15 clusters (c_0:2, c_1:2, c_2:6, c_3:3, c_4:2) in total with a summarized score of 283.32. To ensure that the concepts underlying the clusters can be recognized efficiently during visual inspection, but still provide a detailed insight into the set, we enhance the data tables displaying the values of the input data and the SHAP feature importances with a kind of bar chart for each row (see Figure 3). For the values of the individual input data, we calculate the length of the bars as a percentage of the prevailing minima and maxima per feature column of the class. For the SHAP values, we calculate the percentage length based on the global minima and maxima of the class. Here we also preserve the sign, so that we can use different colors to represent the negative (red) and positive (green) feature importance values. To highlight feature-wise per-class maximal and minimal SHAP values, we color these bars in brighter green color for positive maxima and brighter red color for negative maxima. Additionally, we include the cluster number, as well as the predicted class and the ground truth class as columns to identify mispredictions of the model. This VA-inspired visualization makes it possible for a user to quickly recognize patterns present in the clusters. Besides the data tables, we additionally visualize the clusters' location in a 2D map with *latitude* and *longitude* as axes (see Figure 1).

Interpretation of clusters Though both geographical features have the highest importance for the model's predictions according to global SHAP values, we observe several sub-concepts that seem to influence the prediction. For instance, the highest priced houses can be found in and close to the three big cities, but the strategies derived from the analysis of the identified clusters do not indicate that the model solely uses this strategy. Instead, most identified concepts also involved other combinations of features. The clusters 4 and 5 of class 4 are good examples for this observation (see Figure 3). A noticeable common feature of the instances in

Figure 3: Visualization of two concept explanation clusters in tabular format for class 4 (highest-priced houses).

cluster 4 is the age of the houses, which reaches the maximum everywhere except for one (still old) house, as well as the very low median income (*MedInc*) and the low average number of rooms per household (*AveRooms*). Cluster 5, on the other hand, consists of much younger houses, which also have a high average number of rooms. The values of *MedInc* are also medium to high. The incorrect classification of instance 15757 also shows that the model focused on the features *HouseAge*, *AveOccup*, *latitude* and *longitude*, but, as a possible reason, it is noticeable that the much higher *AveRooms* value compared to the rest of the cluster was not taken into account. It is also noticeable that the values of the important features show a higher homogeneity than those of the non-important features.

As the two weighting functions identify different amounts of clusters, we additionally compare the differences among both approaches. During concept analysis, we find that though the second approach reached a higher Calinski Harabasz score, the clusters found by weighting function w_1 have a more detailed explanatory level, as the sub-concepts appear to be better separated. For example, while analysing the CEC for the lowest priced houses (class 0), the prevailing strategy of the two identified clusters using weighting function w_2 was the geographical location of the houses (north vs. south), while we visually recognized several subconcepts that occured in both clusters. Using w_1, these sub-concepts were identified, e.g., a cluster of old houses with a small amount of rooms (see also part (A) of Figure 1).

Conclusion and Future Work

We have presented a proof of concept that indicates the usefulness of our method *concept explanation clusters*, with which we were able to efficiently recognize different concepts representing sub-strategies the model uses for predicting the classes, as well as identifying reasons for mispredictions. It also shows that the use of input features instead of latent space features is a promising alternative that maintains the relation to the input data when identifying clusters with an underlying human-understandable concept. Keeping the original data allows an easy integration into existing VA concepts, as demonstrated with the geographical visualization, which supported the visual and semantic interpretation.

During the analysis, we noticed that the efficiency of our approach might further be increased by including a way of identifying prototypes per cluster in the future, such that the underlying concept is summarized in this prototype and can be recognized more quickly, while still offering the detailed view on the full cluster where needed. Additionally, we plan to conduct experiments using other model and data types, e.g., especially unstructured data such as images, to evaluate the applicability of CEC introduced to other domains. Further, we want to test other clustering algorithms. Our method has the advantage of being open to different methods (e.g., regarding feature attribution or clustering). We, however, relegate investigating their impact, especially w.r.t. human efficiency and understandability, to a future user study.

Acknowledgements

The development of this publication was supported by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia as part of the flagship project ZERTIFIZIERTE KI.

References

Andrienko et. al. 2022. Visual Analytics for Human-Centered Machine Learning. *IEEE Computer Graphics and Applications*, 42(1).

Biecek, P.; and Samek, W. 2024. Position: Explain to Question not to Justify. In *41st ICML*.

Dreyer et. al. 2024. Understanding the (Extra-)Ordinary: Validating Deep Model Decisions with Prototypical Concept-based Explanations. In *CVPR Workshop*.

Fel et. al. 2024. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *NeurIPS*, volume 37.

Haedecke et. al. 2023. ScrutinAI: A visual analytics tool supporting semantic assessments of object detection models. *Computers & Graphics*, 114.

Kelley Pace, R.; and Barry, R. 1997. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3).

Kenny, E. M.; and Keane, M. T. 2019. Twin-Systems to Explain Artificial Neural Networks using Case-Based Reasoning: Comparative Tests of Feature-Weighting Methods in ANN-CBR Twins for XAI. In *IJCAI-19*.

Kim et. al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*. PMLR.

Leake, D. 2022. Case-Based Explanation: Making the Implicit Explicit. *CCBR XCBR'22: 4th Workshop on XCBR*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *NeurIPS*, volume 30.

Nguyen et. al. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *NeurIPS*, volume 34.

Poretschkin et. al. 2023. Guideline for Trustworthy Artificial Intelligence – AI Assessment Catalog. arXiv:2307.03681.

Shoeibi et. al. 2024. Automated detection and forecasting of COVID-19 using deep learning techniques: A review. *Neurocomputing*, 577.