Interpreting Categorical Data Classifiers using Explanation-based Locality

1st Peyman Rasouli
Department of Informatics
University of Oslo
Oslo, Norway
peymanra@ifi.uio.no

2nd Ingrid Chieh Yu
Department of Informatics
University of Oslo
Oslo, Norway
ingridcy@ifi.uio.no

3rd Ernesto Jiménez-Ruiz
Department of Computer Science
City, University of London
London, UK
ernesto.jimenez-ruiz@city.ac.uk

***Abstract*—Local surrogate explanation methods are a popular class of post-hoc interpretability approaches that explain the rationale of machine learning models in the locality of every particular instance. *Fidelity*, which refers to the accuracy of explanation methods in imitating the actual behavior of a model, is highly affected by their strategy for identifying the *locality* of instances. To find the locality of an instance, we need to calculate the distance between the instance and perturbed data points concerning categorical and numerical features. While the distance of numerical features can be measured precisely, the existing works usually adopt a coarse-grained or imprecise approach for comparing categorical features. This is especially problematic in the categorical data setting, where defining a representative locality demands fine-grained semantic similarity information between categories. In this paper, we propose a locality generation approach for categorical data classifiers that makes no assumption about domain knowledge and infers categorical similarities by relying on the model's explanations. Further, we devise a multi-centered sampling approach based on the derived similarity information that, compared to the conventional instance-centered technique, captures the local behavior of the model more effectively. Moreover, we develop a knowledge-based locality generation approach based on knowledge graphs to benchmark our explanation-based method against a scenario where the similarity information is provided by a domain expert. The experiments conducted on various data sets demonstrate the efficacy of our approach in generating faithful explanations.**

***Index Terms*—Interpretable Machine Learning, Local Surrogate Explainability, Categorical Data, Domain Knowledge**

I. Introduction

The research around explaining Machine Learning (ML) classification models in the tabular data setting has received significant attention over the past years [1]. Local explanation methods [2] discover the models' rationale for every single prediction. The generated explanations for the tabular data are commonly represented as feature importance vectors and decision rules. The open problem in the local explanation domain is the *fidelity* of explanations [3]. Fidelity implies that a created explanation for an instance should express the actual behavior of the model in the locality of the instance. Unfaithful explanations provide incorrect information about the model's decision-making logic that can mislead the user of the model.

Local surrogate explanation techniques (e.g., LIME [4], LORE [5], and EXPLAN [6]) that explain an instance by creating an interpretable surrogate model on its locality are considered a popular approach for tabular applications. The core task of these methods is to find a precise locality of the inputs to capture the local behavior of the original model. Mainly, for a given instance, they generate random data points by perturbing numerical and categorical features and select the samples that best represent the locality of the instance. The selected samples are then used for creating an interpretable surrogate model (e.g., a linear model [7]) that interprets the model's rationale regarding the instance's decision.

Similarity measurement between data points plays a vital role in creating an accurate and representative locality [2]. State-of-the-art explanation methods demonstrate high performance for numerical data sets because quantitative values can be subjected to various statistical and arithmetical operations for measuring similarity. For example, *Age* is a numerical feature that can take continuous values within the range `[0, 100]`; we can adopt arithmetic operations to precisely calculate the distance between two individuals regarding *Age*. On the contrary, categorical features denote qualitative data, which in the absence of domain knowledge, the similarity between different categories is undefined. For instance, the categorical feature *Education* that can take various categories from the set {`High-school`, `Bachelors`, `Masters`, `Doctorate`} does not allows to determine a quantitative similarity between two people with different degrees.

To the best of our knowledge, none of the existing explanation methods considers the semantic or fine-grained similarity between categorical values. They either use the Simple Matching Coefficient (SMC) metric [8], which only evaluates whether two values are identical, [4], [9] or convert categorical features to ordinal representation and measure their distance using numerical metrics (e.g., $\ell 1$- or $\ell 2$-norm) [5], [6], [10]. The former approach does not provide detailed and semantic similarity between categories. The latter approach, on the other hand, can be incorrect as categorical values usually do not follow an order. In the case of mixed feature data sets (including both categorical and numerical features), these shortcomings can be somewhat alleviated by calculating the precise distance between numerical features. However, it can be challenging in merely categorical data scenarios, especially if no domain knowledge is available to define the similarity between categories. In this circumstance, calculating

the distance between the input and perturbed samples via SMC is less likely to result in a neighborhood data set that represent the actual behavior of the classifier in the input's locality. One can convert categorical features to dummy variables and apply numerical distance metrics [11]. However, similar to SMC, this approach does not reveal the semantic similarity between categories, and the provided distance information is identical to the SMC method.

Another deterrent factor for capturing the accurate locality in the local surrogate methods, according to [12], is the instance-centered sampling that can neglect the effect of the locally important features in favor of the globally important ones. Precisely, the instance-centered sampling follows a coarse-grained approach that does not take into account the predicted label of data points, a valuable source of information for understanding the decision boundaries of the original model. Disregarding this information leads to creating and/or selecting random samples in any *direction* within a hypersphere centered on the instance being explained. As a result, the surrogate model will deviate from the actual locality of the instance and instead highlights globally important features for the black-box model [12]. The instance-centered sampling can also create class-imbalanced neighborhood data for samples far from the decision boundaries, especially in imbalanced and multi-class data set settings, where the perturbed data follows the distribution of the training data.

In this work, we propose a locality construction approach that resolves the mentioned issues, leading to faithful and accurate explanations. Specifically, we use global explanations to identify semantic similarities between the categories of categorical variables and incorporate this information in a novel multi-centered sampling approach for capturing the actual locality of inputs. We evaluate the performance of our algorithm w.r.t several binary and multi-class classification data sets, different interpretable models, and various baselines. Further, to compare the model's insights versus the expert's insights about data similarities within a domain, we develop a knowledge-based locality generation based on knowledge graphs in the experiments section. Generally, the main contributions of the work can be summarized as follows:

- We propose a tailored locality generation approach for categorical data classifiers that derives the semantic similarity of categories from the models' explanations and utilizes the obtained information in a novel multi-centered sampling technique to capture the actual behavior of the models in the locality of every particular instance.
- We develop a knowledge-based locality generation approach based on knowledge graphs as a baseline to benchmark the efficacy of classifiers' insights versus experts' insights regarding categorical similarities.

The rest of the paper is organized as follows: Section II further investigates the existing challenges and motivates our proposed solution; Section III introduces our explanation-based locality generation technique for categorical data classifiers; Section IV presents and discusses the conducted experiments for evaluating the efficacy of our proposed method; finally, Section V concludes the paper and states the future works.

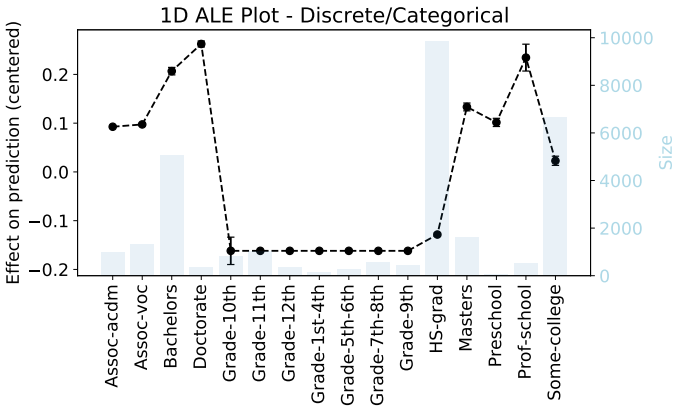
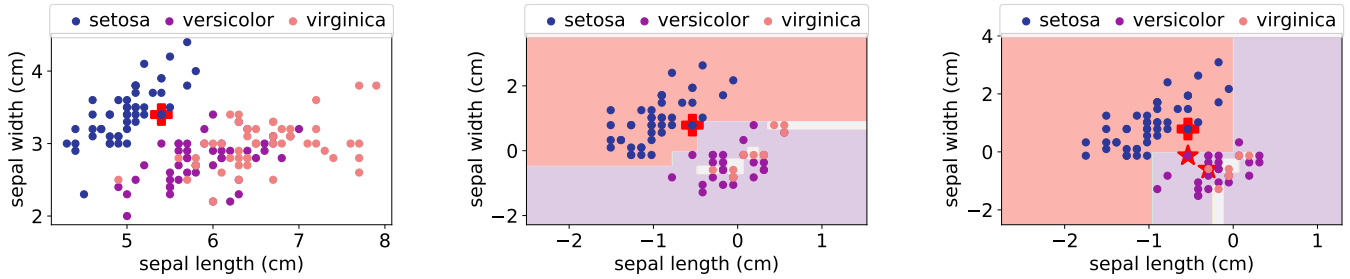


Fig. 1: The global explanation of the *Education* feature from the *Adult Income* data set for the *Gradient Boosting Classifier* model. The *x-axes* demonstrate the feature's categories and the *y-axes* represent their global effect on the model's prediction probability for the target class $Salary > 50K$.

II. Motivation

Categorical Similarity. A potential solution for fine-grained similarity measurement in the categorical data scenario can be incorporating domain knowledge, formulated as taxonomies or knowledge graphs, in the sampling procedure [13]. However, such information is not available for most ML data sets. Moreover, employing domain knowledge for explaining a classifier that is merely created based on observational data may cause an interpretation gap. Thus, it would be useful to identify categorical similarities by relying on the model and its corresponding observational data. Global explanation methods [14] explain the overall behavior of a model w.r.t a group of samples (generally training data). Precisely, they determine the quantitative effect of features' categories (in the domain $\mathbb{R}$) in the model's predictions. The global explanation of a feature reveals categories with similar influence on the model's outcome, providing quantitative semantic similarity between categories from the model's viewpoint. The quantitative similarity information allows applying existing numerical distance metrics (e.g., $\ell 1$- or $\ell 2$-norm) for measuring the distance between data points in the original categorical representation.

In Fig. 1, an example of global feature effect for the *Education* feature in the *Adult Income* data set [15] perceived by the *Gradient Boosting Classifier* [16] is illustrated. The figure demonstrates the global effect of the feature's categories on the overall prediction probability of the model for the target class $Salary > 50K$. The effects represented as black dots on the dashed lines indicate that the categories with positive values tend to increase the probability of data points to be classified as $Salary > 50K$, while the categories with negative values decrease their probability. It can be seen that the global feature effect plot reveals similar categories



(a) Original data. (b) Instance-centered. (c) Multi-centered.

Fig. 2: An illustrative data sampling comparison: instance-centered vs. multi-centered.

from the model's perspective that also conform our subjective knowledge about the feature to a great extent. For example, school-related categories (e.g., `Grade-1st-4th`, `Grade-10th`, and `Grade-12th`) influence the model to classify data points as $Salary \leq 50K$, while university-related items (e.g., `Bachelors`, `Masters`, and `Doctorate`) contribute to classifying instances as $Salary > 50K$.

This approach has several advantages. First, it does not demand external knowledge about the domain, making it applicable for every categorical data set. Second, the global effects are extracted from the classification model, hence, they represent the actual model's perspective about the categories. Third, ML models learn data differently and may have varying decision boundaries that result in different perceptions about features' categories. As explanation methods should always interpret the true behavior of a model regardless of its validity (wrong or correct), this approach is insusceptible to the models' variations and consistently reveals the actual attitude of the models about the data. Last but not least, this is a computationally efficient approach for extracting categorical similarity that can be used for any ML classification model regardless of the model's internal and architecture.

Data Sampling. State-of-the-art local surrogate explanation methods adopt an instance-centered approach that creates and/or selects perturbed data points in a hypersphere centered on the instance of interest. However, this technique usually highlights the effect of globally important features, failing to capture the local behavior of the model for a given input [12]. Moreover, for instances located far from the decision boundaries, it tends to generate neighborhood samples belonging to the same class as the instances, creating imbalanced training data for the surrogate model. Apart from this, since features' categories have different global effects on various classes, it is essential to consider the label of data points while calculating the distance in the global feature effect representation. Therefore, to resolve the stated issues regarding instance-centered sampling and to make explanation-based locality suitable for multi-class classification scenarios, we propose an efficient multi-centered data sampling technique. Instead of using the input instance as a reference for finding the closest neighborhood samples in all classes, we adopt a fine-grained approach by first identifying its counterpart data points in every other class and using them as references for selecting locality samples.

Fig. 2 demonstrates the difference between instance-centered and our proposed multi-centered locality through an example from the *Iris* data set [15]. It can be seen that the instance-centered sampling (Fig. 2(b)), which only relies on the feature space similarity, does not capture the local behavior of the model in the vicinity of the input (denoted by the red cross). In contrast, the multi-centered approach (Fig. 2(c)) uses the labels of the data to understand the decision boundaries of the original model and *narrow* the sampling direction to select instances around local inter-class boundaries. Specifically, it finds the counterparts of the input in other classes (denoted by the red asterisks) and uses them for guiding the data sampling, leading to a balanced data set that can precisely describe the local behavior of the model in the adjacency of the input. The overall F1-score of the surrogate models created on the instance-centered and multi-centered data are 0.93 and 0.96, respectively, indicating the better performance of the multi-centered sampling in representing the instance's locality.

III. Explanation-based Locality for Categorical Data Classifiers

This section describes our proposed approach for creating representative neighborhoods to faithfully explain the decisions of classification models in the categorical data setting. It consists of five procedures outlined in Algorithm 1. Let f be a black-box classifier, $\mathcal{X}_{train}$ be the training data, $\mathcal{X}_{explain}$ be a set of unseen data for explanation, C be the set of classes, M be the set of features, N_{random} be the number of initial random samples, and $N_{samples}$ be the number of final neighborhood samples. For an instance $x \in \mathcal{X}_{explain}$, Algorithm 1 generates a locality (i.e., $\mathcal{X}_{locality}$) that is used as training data for creating an interpretable surrogate model f' that provides explanations for the decision of the original model f regarding x.

A. Global Feature Explanation

We use global explanation methods to obtain fine-grained similarity information of features' categories perceived by the black-box classifier. Accumulated Local Effects (ALE) [17]

is an effective global explanation technique that provides the marginal effect of one or two features on the prediction of a machine learning model. ALE is a faster and unbiased alternative to Partial Dependence Plots (PDPs) [16], a widely-used approach for global explanation of ML models. Compared to PDP, ALE considers feature correlations, leading to more accurate and realistic approximation of global feature effects. This is a valuable property, especially for our use case, because it provides *contextual* similarity information about features' categories. Moreover, it helps to address a common problem in local surrogate explanation methods, which is generating inconsistent and unrealistic neighborhood samples in highly-correlated data sets [18].

Because of the stated advantage of ALE, we employ this technique to measure the effect of every categorical feature on the prediction of the model f. Using this approach we map the feature space from a categorical representation to a numerical representation that allows calculating fine-grained similarity between different categories of a feature. Let us define **ALE** as a function that given a set of features M, a set of categories for every feature K, a set of classes C, a set of training samples $\mathcal{X}_{train}$, and a model f returns a global feature effect matrix $E \in \mathbb{R}^{M \times K \times C}$, where $E_{m,k,c}, m \in M, k \in K, c \in C$, represents the effect of category k existing in feature m on the prediction of the classifier for class c. Formally, the global effect matrix E for a data set is defined as $E = \{\textbf{ALE}(f, \mathcal{X}_{train}, m, k, c) : \forall m \in M, \forall k \in K, \forall c \in C\}$. We need to compute the global effect of each feature with respect to every class because every feature has a dissimilar impact on different classes, leading to distinct effect vectors. The GLOBALFEATUREEXPLANATION function in Algorithm 1 outlines the procedure for extracting feature effects. For a given model f and its corresponding training data $\mathcal{X}_{train}$, this function is only executed once to generate the global effect matrix E that is later used for creating the locality of every input $x \in \mathcal{X}_{explain}$.

B. Class-wise Nearest Neighborhood Model

This widely-used instance-centered sampling can fail to capture the effect of locally important features and may create class-imbalanced neighborhood data for samples far from the decision boundaries. We resolve these issues by proposing a multi-centered sampling approach that selects neighborhood data from every class that captures inter-class decision boundaries. For every class in the data set, we create a 1-nearest neighborhood model using training data, denoted by $\mathcal{N}_c, \forall c \in C$. The NEARESTNEIGHBORHOODMODEL procedure in Algorithm 1 outlines the required steps for constructing the neighborhood models. The models are created in the initialization stage of the algorithm. During runtime, they are queried to find the closest samples to an instance in other classes. Formally, for an input $x \in \mathcal{X}_{explain}$, we use the models to find the following set: $\mathcal{R} = \{\mathcal{N}_c(x) : \forall c \in C\}$. The set $\mathcal{R}$ contains x and its immediate neighborhood samples in other classes, acting as representatives for x. Using the obtained set, we can perform a multi-centered sampling in which perturbed samples from other classes are compared with the counterparts of x, rather than being compared with x itself. This results in a fine-grained sampling strategy that considers not only the similarity of feature values but also the class of samples for finding an accurate locality.

Algorithm 1 Explanation-based Locality

Input: x: instance to explain, f: black-box model, $\mathcal{X}_{train}$: training data, $N_{samples}$: number of neighborhood samples, M: set of features, K: set of features' categories, C: set of classes
Output: $X_{locality}$: a set of samples representing the locality of x

procedure GLOBALFEATUREEXPLANATION($f, \mathcal{X}_{train}, M, K, C$)
  $E = \{\}$
  for all $m \in M$ **do**
    for all $k \in K$ **do**
      for all $c \in C$ **do**
        $E_{m,k,c} = ALE(f, \mathcal{X}_{train}, m, k, c)$
  return E

procedure NEARESTNEIGHBORHOODMODEL($f, \mathcal{X}_{train}, C$)
  $\mathcal{N} = \{\}$
  for all $c \in C$ **do**
    $\mathcal{X}^c_{train} = \{x : f(x) = c, \forall x \in \mathcal{X}_{train}\}$
    $\mathcal{N}_c = CreateModel(\mathcal{X}^c_{train}, n_{neighbor} = 1)$
  return $\mathcal{N}$

procedure CAT2NUMTRANSFORMER($f, M, E, \mathcal{X}^{cat}$)
  $\mathcal{X}^{num} = \{\}$
  for all $x^{cat} \in \mathcal{X}^{cat}$ **do**
    $c = f(x)$
    $x^{num} = \emptyset$
    for all $m \in M$ **do**
      $k = x^{cat}_m$
      $x^{num}_m = E_{m,k,c}$
    $\mathcal{X}^{num} = \mathcal{X}^{num} \cup x^{num}$
  return $\mathcal{X}^{num}$

procedure RANDOMDATASAMPLING($\mathcal{X}_{train}, M, N_{random}$)
  $\mathcal{S} = \{\}$
  for all $m \in M$ **do**
    $freq_m = CalculateFrequencyHistogram(\mathcal{X}_{train_m})$
    $\mathcal{S}_m = GenerateRandomData(freq_m, N_{random})$
  return $\mathcal{S}$

procedure NEIGHBORHOODSAMPLESELECTION($x, f, \mathcal{X}_{train}, C, M, E, \mathcal{N}, N_{samples}$)
  $N_{random} = 10 * N_{samples}$
  $\mathcal{S} =$ RANDOMDATASAMPLING($\mathcal{X}_{train}, M, N_{random}$)
  $\mathcal{R} = \{\}$
  for all $c \in C$ **do**
    $\mathcal{R}_c = \mathcal{N}_c(x)$
  $\mathcal{S}^{num} =$ CAT2NUMTRANSFORMER($f, M, E, \mathcal{S}$)
  $\mathcal{R}^{num} =$ CAT2NUMTRANSFORMER($f, M, E, \mathcal{R}$)
  $\mathcal{D} = \{\}$
  for $i \leftarrow 1$ to N_{random} **do**
    $cs = f(\mathcal{S}_i)$
    $cx = f(x)$
    $\mathcal{D}_i = ||\mathcal{R}_{cs} - \mathcal{S}_i||_0 + ||\mathcal{R}^{num}_{cs} - \mathcal{S}^{num}_i||_1 + ||\mathcal{R}^{num}_{cx} - \mathcal{R}^{num}_{cs}||_1$
  $\mathcal{X}_{locality} = SelectNClosestSamples(\mathcal{S}, \mathcal{D}, N_{samples})$
  return $\mathcal{X}_{locality}$

C. Categorical to Numerical Transformation

This section defines a function to map features in categorical representation to continuous numerical representation (i.e., global feature effect). Let $\mathcal{X}^{cat} \in \mathbb{Z}^M$ be a set of data points with M-dimensional categorical features. Given the black-box

classifier f and the obtained global feature effect matrix E, we define a function $\Phi : \mathcal{X}^{cat} \rightarrow \mathcal{X}^{num}, \mathcal{X}^{num} \in \mathbb{R}^M$, to map categorical data to real-valued representation:

$$\Phi(\mathcal{X}^{cat}) = \{x_m \leftarrow E_{m,k,c}, \forall x \in \mathcal{X}^{cat} \mid x_m = k \wedge f(x) = c\} \quad (1)$$

The CAT2NUMTRANSFORMER function in Algorithm 1 implements this mapping. By doing this transformation, we will achieve a real-valued representation of features' categories that allows applying various distance metrics available for numerical features (e.g., $\ell 1$- or $\ell 2$-norm) to calculate the precise distance between categories.

D. Random Data Sampling

Our proposed approach relies on the distribution of the observed data for creating localities. For every feature, we sample random data based on the frequency histogram of its values (categories) in the training data. Compared to feature manipulation techniques [5], [10], this is a more reliable approach as it is less likely to generate out-of-distribution and outlier data points. The RANDOMDATASAMPLING procedure in Algorithm 1 outlines the mentioned data generation approach that given the training data $\mathcal{X}_{train}$, a set of features M, and a user-defined number of samples $N_{samples}$, it returns a set of random samples $\mathcal{S}$ following the original data distribution.

E. Neighborhood Sample Selection

In this section, we propose a function (NEIGHBORHOODSAMPLESELECTION in Algorithm 1) that utilizes the defined procedures and outputs of previous stages for selecting a representative neighborhood data for an instance of interest. The function starts by creating a set of random samples denoted by $\mathcal{S}$. Given the number of desired samples for the locality, i.e., $N_{samples}$, initially, we generate a diverse set of random data points (e.g., $N_{random} = 10 * N_{samples}$) to properly cover the feature space. Next, we find the set counterparts of x in every class using neighborhood models $\mathcal{N}$, denoted by $\mathcal{R}$ (it also contains x). We use these instances to perform a fine-grained comparison between x and random samples $\mathcal{S}$. In other words, we consider the label of data points for measuring their similarity alongside their feature values. Further, we transform the set of random samples $\mathcal{S}$ and representative instances $\mathcal{R}$ in categorical representation to numerical (i.e., global feature effects) representation denoted by $\mathcal{S}^{num}$ and $\mathcal{R}^{num}$, respectively. The distance of the random samples to x is represented by a vector $\mathcal{D}$ that contains the aggregation of the $\ell 0$-norm and $\ell 1$-norm between every random sample and its same-class representative instance in categorical and numerical representations, respectively, as well as the $\ell 1$-norm between x and its representative instance in the numerical representation to put more emphasis on the vicinity of x for sample selection than its counterparts. We chose $\ell 0$-norm for calculating distance in categorical representation to count the number of features having distinct values, whereas we used $\ell 1$-norm in numerical representation, a suitable metric for real-valued data that takes into account the sparsity of feature differences in distance calculation. Finally, we select $N_{samples}$ data points from $\mathcal{S}$ that are closest to x based on the distance vector $\mathcal{D}$, denoted by $\mathcal{X}_{locality}$. The $\mathcal{X}_{locality}$, hence, is the neighborhood data set of x in the categorical (original) representation that is used for creating an interpretable surrogate model that explains the x's decision.

TABLE I: Performance of the classifiers in terms of *F1-score*.

Data set	**NN**	**GB**
Adult Income	0.816	0.829
COMPAS	0.784	0.785
German Credit	0.718	0.744
Breast Cancer	0.695	0.637
Car Evaluation	0.931	0.968

IV. EXPERIMENTS AND DISCUSSION

We designed two types of experiments to evaluate the efficacy of our proposed approach in generating faithful explanations: i) evaluation of the proposed explanation-based locality method (called XBL) versus local surrogate explainability baselines concerning standard ML data sets and ii) evaluation of XBL against a knowledge-based locality approach (called KBL) with respect to the *Adult Income* data set.

We used several binary and multi-class classification data sets including *Adult Income*, *COMPAS*, *German Credit*, *Breast Cancer*, and *Car Evaluation*. Except from *COMPAS* data set that is available at [19], all data sets can be found at [15]. It should be mentioned that the numerical features existing in some of the data sets are either removed or discretized to form purely categorical data sets. The data sets were split into $\mathcal{X}_{train}$ and $\mathcal{X}_{test}$ sets that were used for training and testing the classifiers, respectively. TABLE I reports the performance of the models in terms of *F1-score*.

We employed *Multi-layer Perceptron Neural Networks* [20] (NN) consisted of one hidden layer with 100 neurons and *Gradient Boosting Machines* [16] (GB) comprised of 100 estimators as black-box classifiers. We explain inputs using two widely-used interpretable models: CART Decisions Trees (DT) [21] and Ridge Linear Regression (LR) [7]. This enables evaluating the efficacy of our approach for different surrogate models and its generalizability for different explanation formats, i.e., decision rules and feature importance.

The only hyper-parameter required by our approach, which is also common in baselines, is the number of neighborhood samples that is set to $N_{samples} = 1000$. We executed the baseline with their default hyper-parameter settings. For every data set, we randomly selected $Min(500, |\mathcal{X}_{test}|)$ samples from their test set $\mathcal{X}_{test}$ as explaining set $\mathcal{X}_{explain}$.

Our framework has been developed using Python programming language, and experiments were run on a system with Apple M1 Pro processor and 16GB of memory. We provide an open-source implementation of our approach, including data sets and experiments, to facilitate reproducible research: https://github.com/peymanrasouli/categorical_locality.

TABLE II: Fidelity evaluation of XBL versus local surrogate explainability baselines in explaining NN classifiers constructed for standard ML data sets.

	Interpretable Model	DT				LR			
Data Set	Method	*F1-score*	*Precision*	*Accuracy*	*M-score*	*R2-score*	*MAE*	*MSE*	*M-score*
Adult Income	GEN	**0.922**	0.896	0.942	0.844	**0.607**	0.081	0.010	0.671
	MDS	0.921	0.944	**0.948**	0.791	0.397	0.095	0.015	0.683
	RIS	0.753	0.747	0.784	0.787	0.327	0.106	0.017	0.445
	XBL	0.910	**0.962**	0.944	**0.930**	0.600	**0.075**	**0.010**	**0.820**
COMPAS	GEN	0.977	0.991	0.986	0.933	0.761	0.056	0.005	0.897
	MDS	0.941	0.952	0.964	0.945	0.732	0.054	0.005	0.882
	RIS	0.884	0.854	0.920	0.768	0.536	0.075	0.009	0.722
	XBL	**0.990**	**0.992**	**0.994**	**0.994**	**0.933**	**0.021**	**0.001**	**0.964**
German Credit	GEN	0.903	0.87	0.930	0.828	0.444	0.091	0.012	**0.696**
	MDS	0.724	0.743	0.835	0.697	-0.200	0.129	0.027	0.513
	RIS	0.499	0.644	0.505	0.656	-0.952	0.180	0.043	0.429
	XBL	**0.992**	**0.997**	**0.995**	**0.840**	**0.508**	**0.078**	**0.011**	0.581
Breast Cancer	GEN	0.934	0.900	0.964	0.910	0.693	0.060	0.006	0.860
	MDS	0.808	0.827	0.911	0.829	0.534	0.066	0.008	0.728
	RIS	0.868	0.839	0.929	0.824	0.198	0.090	0.014	0.733
	XBL	**0.962**	**0.990**	**0.982**	**0.913**	**0.790**	**0.039**	**0.004**	**0.886**
Car Evaluation	GEN	**0.992**	0.989	0.988	0.655	0.403	0.112	0.018	0.773
	MDS	0.515	0.556	0.896	0.459	-0.590	0.184	0.049	0.604
	RIS	0.902	0.941	0.977	0.818	0.485	0.102	0.016	0.821
	XBL	0.969	**0.991**	**0.991**	**0.947**	**0.693**	**0.076**	**0.010**	**0.834**

A. Evaluation of XBL vs Local Surrogate Explainability Baselines on Standard ML Data sets

We benchmark our proposed explanation-based locality approach (XBL) against various locality generation strategies including:

- Genetic-based Neighborhood (GEN) [5]: for an input x, it uses a genetic algorithm to generate a set of neighborhood data containing a similar number of samples for every class that are closest to x.
- Meaningful Data Sampling (MDS) [10]: for an input x, it generates random samples based on the distribution of the training data, and then makes data points closer to x by setting x's values for the features having a similar contribution to the model's output.
- Random Instance Selection (RIS) [5]: for an input x, it generates stratified random samples based on the distribution of the training data that are closest to x, and then under-samples the data points based on the Condensed Nearest Neighbour Rule [22].

TABLE II presents the fidelity results of the DT and LR interpretable models created using various locality generation techniques for explaining NN classification models. The explanations are evaluated regarding various metrics. For the DT interpretable model, where the model's output is the predicted class of the instance, we used *F1-score*, *Precision*, and *Accuracy* metrics to measure the similarity between the predictions of the interpretable model and the original classification model; the higher these metrics are, the more faithful explanations are generated. Similarly, for the LR interpretable model, where the model's output is the prediction probability of the instance, we applied *R2-score*, *Mean Absolute Error* (*MAE*), and Mean Squared Error (*MSE*) metrics; in this case, the higher values for *R2-score* and lower values for *MAE* and *MSE* indicates the crated localities have represented the actual behavior of the original classification model, leading to more accurate explanations. Moreover, we measure the overall score of the interpretable models (using *F1-score* for DT and *R2-score* for LR), denoted by *M-score*, w.r.t the entire neighborhood instances. We believe this is an essential metric for evaluating the performance of local surrogate explanation methods. Because if an interpretable model only demonstrates a good performance concerning the input instances and not their locality data points, it implies that the interpretable model either has over-fitted on the inputs (caused by training on very similar neighborhood samples to the inputs, i.e., narrow localities) or has not covered the actual locality of the inputs, leading to a poor overall score.

According to the reported results in TABLE II, our proposed approach, i.e., XBL, has outperformed the baselines regarding the majority of scenarios. For the DT interpretable model, the generated localities by XBL resulted in highest performance regarding the most instance-level metrics (i.e., *F1-score*, *Precision*, and *Accuracy*), except for the *Adult Income* and *Car Evaluation* data sets for which GEN demonstrates a better performance regarding *F1-score*. The notable point is that XBL has led to the highest overall score (i.e., *M-score*) for the DT model regarding all data sets, indicating the representativeness of the created localities for the explained instances. The reported results for the LR intepretable model are similar to DT, confirming the efficacy of XBL localities regardless of the employed surrogate model. Except for the *Adult Income* data set, XBL significantly outperforms the

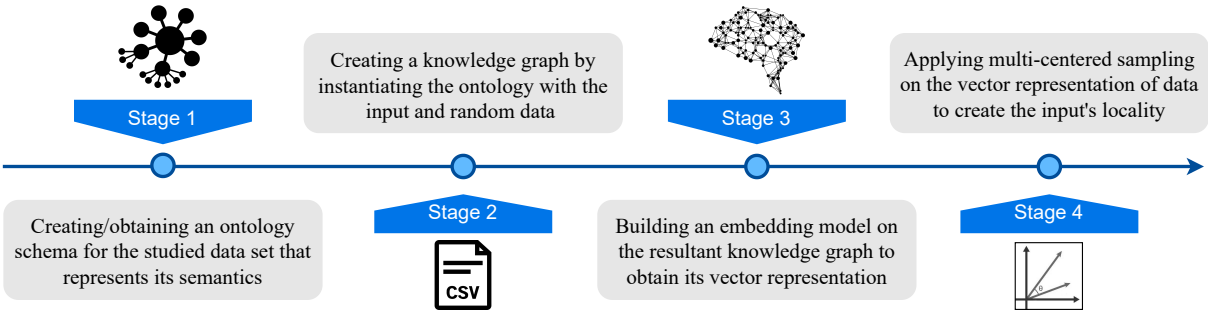


Fig. 3: The procedure of the knowledge-based locality (KBL) approach.

baseline methods in generating faithful explanations in the LR setting. Furthermore, we notice that MDS and RIS baselines show high fidelity in the DT scenario whereas their performance is considerably diminished in the LR setting. This is because, compared to the LR model, DT is more prone to overfitting and class-imbalance data. However, the similar (high) performance of XBL in both DT and LR scenarios confirms that the DT model is not over-fitted and the created localities captured the actual behavior of the classifier.

Apart from high fidelity performance, our proposed approach generates localities in a computationally efficient manner. It should be noted that global feature effects (Section III-A) and class-wise nearest neighborhood models (Section III-B) are created once in the initialization phase of XBL and used later for explaining every instance. This eases the computational burden of our approach and leaves the locality generation of every instance with lightweight operations and procedures. We evaluated the execution time of XBL and the baselines concerning *Adult Income* which is considered a standard data set for tabular classification. Our proposed XBL approach, on average, takes 1.71 seconds to generate neighborhood data points. In contrast, GEN, MDS, and RIS take 2.76, 0.28, and 7.94 seconds, respectively. These results validate the efficiency of our locality generation approach.

B. Evaluation of XBL vs KBL on Adult Income Data set

We benchmark our proposed XBL approach against a knowledge-based locality method (KBL). This experiment aims to compare the model's insight (obtained by global feature effects) versus the expert's insight (achieved via a knowledge graph) about categorical similarities within the domain for generating faithful explanations. The expert's knowledge is represented by the concept of *ontology* which is defined as *"a representation of a shared, agreed and detailed model or set of concepts of a certain problem, domain"* [23]. The created ontology schema is used to express and interpret the data of a *knowledge graph*, an effective framework for data integration, unification, linking and reuse [24]. Such a knowledge graph formally represents semantics by describing entities and their relationships that allows logical inference for retrieving implicit knowledge rather than only explicit knowledge. A common way to use knowledge graphs in ML tasks is to derive latent feature representations of entities and relations using *Knowledge Graph Embedding* (KGE) models [25]. We implement the above-mentioned approach to convert categorical data points into feature vectors that enable measuring their conceptual similarity from the expert's perspective.

The procedure of KBL is illustrated in Fig. 3. The prerequisite of the KBL approach is the existence of an ontology schema for a studied data set (Fig. 3: Stage 1). However, since such domain formalisms are rarely available for the existing ML data sets, we decided to create an ontology schema for a widely-used data set in the explainability research community, i.e., the *Adult Income* data set. We created an ontology $\mathcal{O}$ for the data set based on our knowledge about the features and gathered information from valid sources like regional groupings by United Nations [26], standard classifications of occupations by ILOSTAT [27], annual wages in the USA based on major occupational groups by STATISTA [28], and structure of the USA education system by WENR [29]. In the second stage of the KBL approach (Fig. 3: Stage 2), the created ontology $\mathcal{O}$ is instantiated with a CSV data containing an input of interest x and randomly generated tabular data $\mathcal{S}$, resulting in an RDF-based knowledge graph KG. The knowledge graph KG is composed of RDF triples in the form of $\langle sb, p, ob \rangle$, where sb represents a subject (a class or an instance), p represents a predicate (a property), and ob represents an object (a class, an instance, or a literal); an example of RDF triple for an individual x in the *Adult Income* data set can be $\langle x, hasEducation, Masters \rangle$. To measure the similarity between entities in KG we employed OWL2Vec* [30], a random walk and word embedding based KGE model, in the third stage of the KBL algorithm (Fig. 3: Stage 3) that transforms entities in triple representation $\langle sb, p, ob \rangle$ into a vector representation $\mathbf{e}_{sb}, \mathbf{e}_p, \mathbf{e}_{ob} \in \mathbb{R}^k$, where k is the dimension of the vector. Having the vector representation of the data points provided by the KGE model, we apply the multi-centered data sampling to create the locality of x in the last stage of the KBL algorithm (Fig. 3: Stage 4).

We constructed a GB classifier for the *Adult Income* data set and explained individual instances using interpretable models created on XBL and KBL localities. The evaluation results of this experiment are reported in TABLE III. It can be seen that XBL demonstrates a superior performance than KBL in both surrogate models. It should be mentioned

TABLE III: Fidelity evaluation of XBL versus KBL in explaining a GB classifier created for the *Adult Income* data set.

Interpretable Model	Method	Metric			
		F1-score	*Precision*	*Accuracy*	*M-score*
DT	XBL	**0.932**	**0.976**	**0.960**	**0.922**
	KBL	0.844	0.829	0.896	0.893
		R2-score	*MAE*	*MSE*	*M-score*
LR	XBL	**0.803**	**0.051**	**0.004**	**0.869**
	KBL	0.412	0.086	0.013	0.850

that designing an ontology schema is a subjective task. The designer can include/exclude different types of information, adopt a desired level of abstraction, and define customized categorizations and relationships between entities. Although the performance of the knowledge-based locality is highly dependent on the ontology and KGE model, this experiment confirms the explanation-based locality, which merely relies on the model's insight, can effectively represent the actual behavior of the classification model in the vicinity of the inputs. This is because the model's insight provides precise and objective information on categorical similarities that are necessary for creating representative localities.

V. Conclusion and Future Work

In this work, we studied the locality generation problem of local surrogate explanation methods in the categorical data setting. To create representative localities in this setting, semantic similarity information about features' categories is required. We proposed an intuitive and efficient explanation-based approach that obtains such information from the model's insight and utilizes them in a multi-centered sampling technique to capture the actual behavior of the model in the locality of every particular instance. To evaluate the efficacy of our algorithm, we also designed a knowledge-based locality generation technique that incorporates the domain expert's insight about categorical similarities into the data sampling. Several evaluation scenarios demonstrated the efficacy of our proposed approach in generating faithful explanations compared to the knowledge-based and state-of-the-art locality generation baselines. In future work, we aim to combine the model's insight with the expert's insight to obtain more precise information about categorical similarities. Moreover, we will study the generalization of the proposed approach for mixed-feature data sets.

References

[1] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[2] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[3] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable artificial intelligence: objectives, stakeholders, and future research opportunities," *Information Systems Management*, vol. 39, no. 1, pp. 53–63, 2022.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[5] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," *arXiv preprint arXiv:1805.10820*, 2018.

[6] P. Rasouli and I. C. Yu, "Explan: Explaining black-box classifiers using adaptive neighborhood generation," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–9.

[7] G. C. McDonald, "Ridge regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 93–100, 2009.

[8] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[10] P. Rasouli and I. C. Yu, "Meaningful data sampling for a faithful local explanation method," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2019, pp. 28–38.

[11] M. Hardy and J. Reynolds, "Incorporating categorical information into regression models: The utility of dummy variables," *Handbook of data analysis*, pp. 229–55, 2004.

[12] T. Laugel, X. Renard, M.-J. Lesot, C. Marsala, and M. Detyniecki, "Defining locality for surrogates in post-hoc interpretablity," *arXiv preprint arXiv:1806.07498*, 2018.

[13] G. Zhu and C. A. Iglesias, "Computing semantic similarity of concepts in knowledge graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 72–85, 2016.

[14] G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," *arXiv preprint arXiv:2006.00093*, 2020.

[15] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[16] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[17] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 82, no. 4, pp. 1059–1086, 2020.

[18] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to shapley values," *Artificial Intelligence*, vol. 298, p. 103502, 2021.

[19] ProPublica, "Compas data set," 2017. [Online]. Available: https://www.kaggle.com/danofer/compass

[20] G. P. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000.

[21] W.-Y. Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[22] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[23] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[24] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. d. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier *et al.*, "Knowledge graphs," *Synthesis Lectures on Data, Semantics, and Knowledge*, vol. 12, no. 2, pp. 1–257, 2021.

[25] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[26] UN, "Sdg indicators," (accessed: 24.05.2022). [Online]. Available: https://unstats.un.org/sdgs/indicators/regional-groups/

[27] I. L. O. ILOSTAT, "International standard classification of occupations (isco)," (accessed: 24.05.2022). [Online]. Available: https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/

[28] Statista, "Mean annual wage in the united states in 2020, by major occupational group," (accessed: 24.05.2022). [Online]. Available: https://www.statista.com/statistics/218235/median-annual-wage-in-the-us-by-major-occupational-groups/#professional

[29] WENR, "Education in the united states of america," (accessed: 24.05.2022). [Online]. Available: https://wenr.wes.org/2018/06/education-in-the-united-states-of-america

[30] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, and I. Horrocks, "Owl2vec*: Embedding of owl ontologies," *Machine Learning*, vol. 110, no. 7, pp. 1813–1845, 2021.