EXPLAN: Explaining Black-box Classifiers using Adaptive Neighborhood Generation

Peyman Rasouli, Ingrid Chieh Yu

Presentation by Michał Raszkowski

Introduction

EXPLAN is a local model-agnostic explanation method applicable to tabular data classification problems. It is a module-based algorithm consisted of dense data generation, representative data selection, data balancing, and rule-based interpretable model.

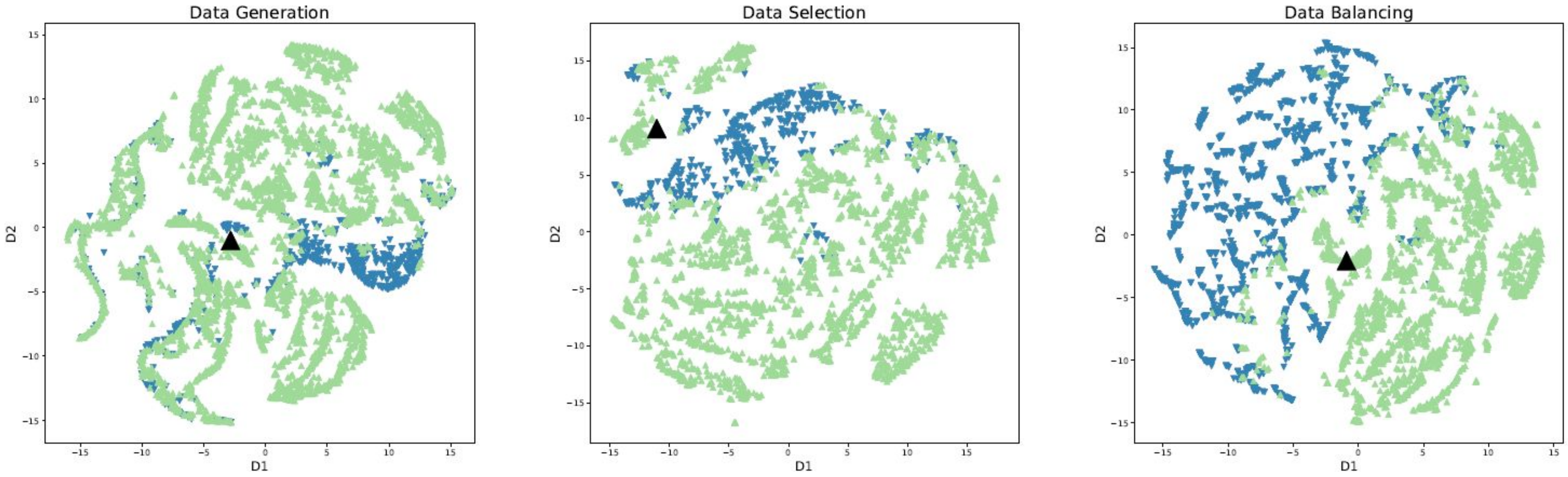


Fig. 2. Visualization of EXPLAN's neighborhood construction process.

Algorithm

Algorithm 1 EXPLAN Explanation Method

Input: $\{x, f, \mathcal{D}, \mathcal{N}, \tau\}$

/* x: *instance to explain,* f: *black-box model,* $\mathcal{D}$: *distribution of training data,* $\mathcal{N}$: *# initial neighborhood samples,* τ: *# minimum samples per class* */

Output: $\{\mathcal{C}, e\}$

/* $\mathcal{C}$: *interpretable model,* e: *explanation of* x */

function EXPLAN$(x, f, \mathcal{D}, \mathcal{N}, \tau)$

  $\mathcal{Z} \leftarrow$ DataGeneration$(x, f, \mathcal{D}, \mathcal{N})$

  $\mathcal{Z}' \leftarrow$ DataSelection$(x, f, \mathcal{Z}, \tau)$

  $\mathcal{X} \leftarrow$ DataBalancing$(f, \mathcal{Z}')$

  $\mathcal{C}, e \leftarrow$ InterpretableModel$(x, f, \mathcal{X})$

  return $\mathcal{C}, e$

Algorithm 2 Dense Data Generation

procedure DATAGENERATION($x, f, \mathcal{D}, \mathcal{N}$)
    procedure RANDOMDATAGENERATION($\mathcal{D}, \mathcal{N}$)
        $\mathcal{S} \leftarrow DataSampling(\mathcal{D}, \mathcal{N})$
        return $\mathcal{S}$ /* *random data points* */
    procedure SURROGATEMODELCONSTRUCTION($f, \mathcal{S}$)
        $\mathcal{T} \leftarrow RandomForestConstructor(\mathcal{S}, f(\mathcal{S}))$
        return $\mathcal{T}$ /* *RF surrogate model* */
    procedure CONTRIBUTIONEXTRACTION($x, \mathcal{S}, \mathcal{T}$)
        $\mathcal{V}(x) \leftarrow TreeInterpreter(\mathcal{T}, x)$
        for all $s \in \mathcal{S}$ **do**
            $\mathcal{V}(s) \leftarrow TreeInterpreter(\mathcal{T}, s)$
        return $\mathcal{V}$ /* *feature importance* */
    procedure SAMPLEMANIPULATION($x, \mathcal{S}, \mathcal{T}, \mathcal{V}$)
        $l_x \leftarrow \mathcal{T}(x)$
        $\mathcal{Z} \leftarrow \{\}$
        for all $s \in \mathcal{S}$ **do**
            $l_s \leftarrow \mathcal{T}(s)$
            for $j \leftarrow 1, \mathcal{F}$ **do**   /* *$\mathcal{F}$: feature dimension* */
                if $(s_j \neq x_j)$ **then**
                    if $(\mathcal{V}^{l_x}_{s_j} = \mathcal{V}^{l_x}_{x_j}) \wedge (\mathcal{V}^{l_s}_{s_j} = \mathcal{V}^{l_s}_{x_j})$ **then**
                        $s_j \leftarrow x_j$
            $\mathcal{Z} \leftarrow \mathcal{Z} \cup s$
        return $\mathcal{Z}$ /* *meaningful dense data w.r.t x* */
    return $\mathcal{Z}$

Draw perturbed samples from training set distribution

↓

Train surrogate random forest

↓

Observation-level feature importance from random forest explanations

↓

Values of features in s with the same contributions as in x set to values from x

Algorithm 3 Representative Data Selection

procedure DataSelection($x, f, \mathcal{Z}, \tau$)
  $n_c \leftarrow 2$ /* n_c: *number of clusters* */
  $\mathcal{Z}' \leftarrow \{\}$
  for all $l \in \mathcal{L}$ **do** /* $\mathcal{L}$: *set of labels* */
    $\mathcal{G}_l \leftarrow \{z \in Z \mid f(z) = l\}$
    $\mathcal{G}_l \leftarrow x \cup \mathcal{G}_l$
    while `True` **do**
      $c_x, c_{\neg x} \leftarrow AgglomerativeClustering(\mathcal{G}_l, n_c)$
      if $|c_x| \geq \tau$ **then**
        $\mathcal{G}_l \leftarrow c_x$
      else
        break
    $\mathcal{Z}' \leftarrow \mathcal{Z}' \cup \mathcal{G}_l$
  return $\mathcal{Z}'$ /* *representative data set* */

Data Balancing

Generate samples in classes with small representations via equation:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i)$$

Rule-based Interpretable model

Build a classifier (decision tree) on obtained dataset and values of black-box model

Experiments

Experimental setup

- Matched against LIME, LORE, and Anchor explanation methods on:
  - fidelity comparison
  - neighborhood analysis
  - explanation comparison
- Black-boxes: Neural Network, Logistic Regression, Gradient Boosting
- Datasets:

TABLE I
DESCRIPTION OF THE DATA SETS.

Data set	# Instances	# Features	Class imbalance
Adult	49K	14	<=50K: 76% - >50K: 24%
German	1K	20	Good: 70% - Bad: 30%
COMPAS	7K	52	Medium-Low: 72% - High: 28%

TABLE II

COMPARISON OF $fidelity_x$ SCORES.

Data set	Black-box	EXPLAN	LIME	LORE
Adult	**GB**	0.994±.1	0.838±.4	0.980±.1
	LR	0.992±.1	0.940±.2	0.989±.1
	NN	0.992±.1	0.859±.3	0.977±.2
German	**GB**	1.000±.0	0.910±.3	0.950±.2
	LR	0.990±.1	0.940±.2	0.910±.3
	NN	1.000±.0	0.930±.3	0.990±.1
COMPAS	**GB**	1.000±.0	0.911±.3	0.999±.0
	LR	1.000±.0	0.925±.3	0.981±.1
	NN	0.999±.0	0.915±.3	0.986±.1

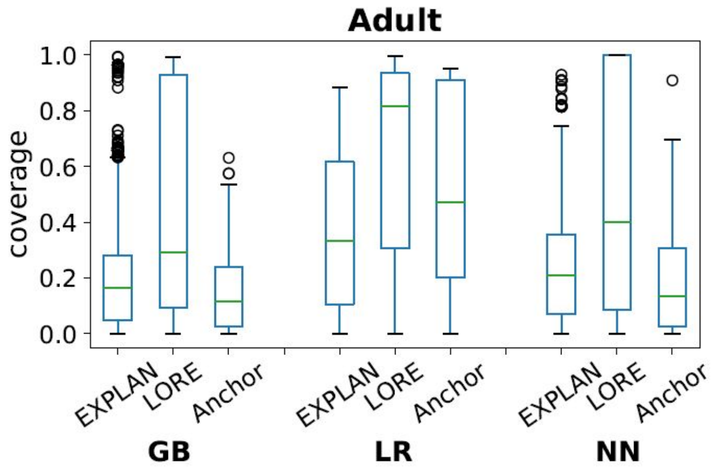
instance-level

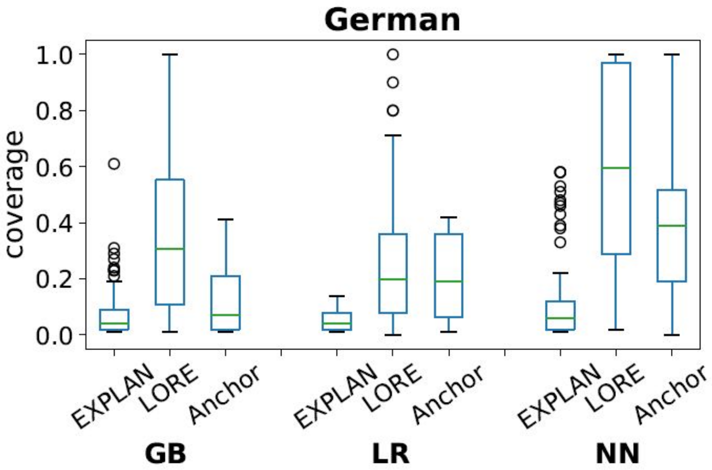
TABLE III

COMPARISON OF $fidelity_{\mathcal{X}}$ SCORES.

Data set	Black-box	EXPLAN	LIME	LORE
Adult	**GB**	0.971±.0	0.738±.0	0.996±.0
	LR	0.990±.0	0.793±.1	0.995±.0
	NN	0.980±.0	0.804±.0	0.993±.0
German	**GB**	0.942±.0	0.223±.1	0.979±.0
	LR	0.972±.0	0.179±.1	0.944±.2
	NN	0.981±.0	0.037±.1	0.987±.0
COMPAS	**GB**	0.984±.0	0.897±.0	0.982±.1
	LR	0.988±.0	0.919±.0	0.975±.1
	NN	0.988±.0	0.896±.0	0.974±.1

neighborhood-level





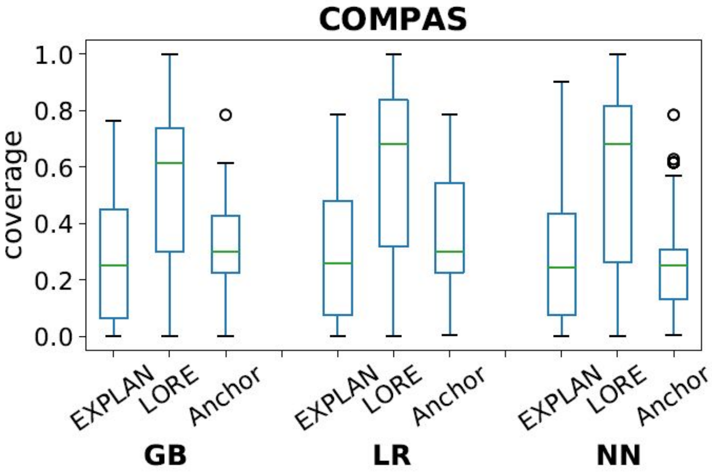


TABLE IV
COMPARISON OF $\boldsymbol{precision_e}$ SCORES.

Data set	Black-box	EXPLAN	LORE	Anchor
Adult	**GB**	0.924±.1	0.852±.2	0.980±.1
	LR	0.966±.1	0.894±.2	0.963±.0
	NN	0.895±.2	0.815±.2	0.971±.1
German	**GB**	0.897±.2	0.816±.2	0.984±.0
	LR	0.937±.1	0.835±.3	0.994±.0
	NN	0.950±.1	0.879±.2	0.976±.1
COMPAS	**GB**	0.914±.2	0.855±.2	0.963±.0
	LR	0.912±.2	0.862±.2	0.963±.0
	NN	0.898±.2	0.851±.2	0.979±.0

TABLE V
COMPARISON OF FEATURE FREQUENCY VARIANCE.

Data set	Black-box	EXPLAN	LORE	Anchor
Adult	**GB**	1.191±.2	2.263±.5	1.497±.1
	LR	1.087±.2	2.259±.5	1.523±.1
	NN	1.174±.5	2.273±.5	1.509±.1
German	**GB**	0.469±.0	2.296±.6	0.472±.0
	LR	0.513±.0	2.293±.7	0.469±.0
	NN	0.467±.0	2.334±.6	0.468±.0
COMPAS	**GB**	0.529±.1	1.342±.4	0.682±.1
	LR	0.535±.1	1.335±.4	0.678±.1
	NN	0.533±.1	1.314±.4	0.681±.1

TABLE VI
COMPARISON OF JACCARD MEASURE OF STABILITY.

Data set	Black-box	EXPLAN	LORE	Anchor
Adult	**GB**	0.827±.1	0.821±.1	0.755±.1
	LR	0.859±.1	0.799±.1	0.671±.1
	NN	0.856±.1	0.728±.1	0.744±.2
German	**GB**	0.702±.1	0.694±.2	0.754±.1
	LR	0.729±.1	0.698±.2	0.819±.2
	NN	0.846±.1	0.779±.1	0.884±.1
COMPAS	**GB**	0.888±.1	0.858±.2	0.859±.1
	LR	0.886±.1	0.859±.2	0.854±.1
	NN	0.848±.1	0.807±.2	0.822±.1

Explanation generated by EXPLAN for an input from *Adult* data set and *GB* black-box

x = { **age:** 30;
workclass: Private;
marital-status: Never-married;
occupation: Prof-speciality;
relationship: Unmarried;
race: White;
sex: Male;
capital-gain: 0;
capital-loss: 0;
hours-per-week: 40;
native-country: United-States
class: $\leq$ 50K

e = { **age:** $\leq$ 30,
capital-gain $\leq$ 0
hours-per-week $\leq$ 44
} -> **class:** $\leq$ 50K

Thank you for your attention

Time for Q&A