World Wide Web (2023) 26:3965–3981
https://doi.org/10.1007/s11280-023-01213-8

CoSP: co-selection pick for a global explainability of black box machine learning models

Dou El Kefel Mansouri[1] · Seif-Eddine Benkabou[2] · Khaoula Meddahi[2] · Allel Hadjali[3] · Amin Mesmoudi[2] · Khalid Benabdeslem[4] · Souleyman Chaib[5]

Received: 25 April 2023 / Revised: 12 September 2023 / Accepted: 15 September 2023 /
Published online: 18 October 2023

Abstract
Recently, few methods for understanding machine learning model's outputs have been developed. SHAP and LIME are two well-known examples of these methods. They provide individual explanations based on feature importance for each instance. While remarkable scores have been achieved for individual explanations, understanding the model's decisions globally remains a complex task. Methods like LIME were extended to face this complexity by using individual explanations. In this approach, the problem was expressed as a submodular optimization problem. This algorithm is a bottom-up method aiming at providing a global explanation. It consists of picking a group of individual explanations which illustrate the global behavior of the model and avoid redundancy. In this paper, we propose CoSP (Co-Selection Pick) framework that allows a global explainability of any black-box model by selecting individual explanations based on a similarity preserving approach. Unlike submodular optimization, in our method the problem is considered as a co-selection task. This approach achieves a co-selection of instances and features over the explanations provided by any explainer. The proposed framework is more generic given that it is possible to make the co-selection either in supervised or unsupervised scenarios and also over explanations provided by any local explainer. Preliminary experimental results are made to validate our proposal.

Keywords Machine learning models · Explicability · Local explanation and global aggregation

1 Introduction

Nowadays, a wide range of real-life applications such as computer vision [1, 2], speech processing, natural language understanding [3], health [4], and military fields [5, 6] make use of Machine Learning (ML) models for decision making or prediction/classification purpose.

This article belongs to the Topical Collection: *Special Issue on Web Information Systems Engineering 2022*
Guest Editors: Richard Chbeir, Helen Huang, Yannis Manolopoulos and Fabrizio Silvestri.

✉ Dou El Kefel Mansouri
douelkefel.mansouri@univ-tiaret.dz

Extended author information available on the last page of the article

However, those models are often implemented as black boxes which make their predictions difficult to understand for humans. This nature of ML-models limits their adoption and practical applicability in many real world domains and affect the human trust in them. Making ML-models more explainable and transparent is currently a trending topic in data science and artificial intelligence fields which attracts the interest of several researchers.

Explainable AI (XAI) refers to the tools, methods, and techniques that can be used to make the behavior and predictions of ML models to be understandable to human [7]. Thus, the higher the interpretability/explainability of a ML model, the easier it is for someone to comprehend why certain decisions or predictions have been made.

Multiple interpretability approaches are based on additive models where the prediction is a sum of individual marginal effects like feature contribution [8], where a value (denoting the influence on the output) is assigned to each feature. One of the latest proposed methods is based on mathematical Shapeley Values and was introduced by Scott et al. [9] as SHAP (for SHapley Additive exPlanations). It relies on combining ideas from cooperative game theory and local explanations [10]. LIME (Local Interpretable Model-agnostic Explanations), introduced by Ribeiro et al. [11], is also one of the most famous local explainable models. It explains individual predictions of any classifier or regressor in a faithful and intelligible way, by approximating them locally with an interpretable model (e.g., linear models, decision trees). However, having a global explanation of the model can be challenging as it is more complicated to maintain a good fidelity - interpretability trade off. To this end, authors in [11] proposed an approach, called submodular Pick which is an algorithm aiming to maximize a coverage function of total feature importance for a set of instances. While maximizing the coverage function is NP-Hard, authors make use of a greedy algorithm which adds iteratively instances with the highest marginal coverage to the solution set, offering a constant-factor approximation to the optimum. The selected set is the most representative, non-redundant individual explanations of the model.

In this paper, our aim is to introduce a new approach to select individual instances (explanations) to be considered for global explanation to ensure that the picked group reflects the global behavior of the black-box model. Unlike submodular optimization proposed in [11], we advocate to consider the problem of picking representative instances as a co-selection task. The idea is to apply a similarity preserving co-selection approach to select a set of instances and features on the explanations provided by any explainer. In fact, feature or instance selection has been widely considered separately in the literature to remove noise, irrelevant and redundant features or instances in datasets [12–16]. Unfortunately, selecting features and instances separately and sequentially is time consuming, especially when dealing with large scale datasets. To overcome this problem, co-selection or the simultaneous selection of features and instances is proposed making use of the duality between feature space and instance space. In this context, several approaches have been proposed. For instance, Kuncheva et al. [17] proposed a genetic algorithm that simultaneously select features and reference cases to improve the performance of nearest neighbor classifiers. Derrac et al. [18] suggested an evolutionary model based on cooperative coevolution to perform co-selection in nearest neighbor classification. GarcíA-Pedrajas et al. [19] proposed a scalable, almost any size, method for concurrent instance and feature selection. In another side, similarity preserving approaches have been considered in the literature with the aim of evaluating features by their ability to preserve locality. For instance, Zhao et al. [20] introduced a similarity preserving feature selection framework that overcomes common weakness in handling feature redundancy. Ma et al. [21] proposed a similarity preserving method that generate unseen visual features from random noises concatenated with semantic descriptions. Shang et al. [22] suggested UFSRL, a framework that used local similarity preserving for feature selection.

Contributions

The technical contributions of this paper are summarized as follows.

- We propose a new approach, called CoSP, for a global explainability of black box machine learning.
- The proposed approach selects individual explanations to provide global explanation for machine learning models.
- CoSP is based on a similarity preserving co-selection approach.
- Experiments are conducted to validate the efficacy of these contributions.
- We release a performant implementation of CoSP at [23].

The original version of this work is published at WISE'22 [24]. The main changes in this paper are presented below:

- Creating a new section (Related Work) that presents some modern explainability methods proposed in the literature to further clarify the importance of interpretability of machine learning models.
- Detailing, in the Proposed approach section, the alternative optimization procedure applied to the objective function.
- Adding computational complexity in the Algorithm Analysis subsection.
- Conducting further experiments to validate the effectiveness of CoSP by comparing it against four approaches including, Random, Greedy [25], Parzen [26] and LIME [11], combined with Submodular Pick (SP) and Random Pick (RP).

The paper is structured as follows. Section 2 introduces the related work. Section 3 provides a necessary background on LIME method. In Section 4, we present our approach allowing for a global explanation of black box ML models. Section 5 shows the preliminary experiments done to validate our proposal. In Section 6, we conclude the paper and draw some research lines for future work.

2 Related work

Interpretability of ML models reflects the ability to provide meaning in understandable terms to human. It is crucial to trust the system and get insights based on its decisions. Quality of an explanation could be improved by making it more Interpretable, Faithful, and model-agnostic [27]. Faithfulness represents how the explanation is describing the reality of the model. Model-agnostic methods are used for any type of model. Several explainability methods are proposed in the literature. LIME introduced by Ribeiro et al. [11], is one of the well-known examples. It is a framework which explains a prediction by approximating it locally using an interpretable model. Other methods were proposed later, for instance, Burkart et al. [28] provided a survey that presents the main explainable of supervised machine learning methods. Lundberg et al. [29] suggested a novel explanation that improves the interpretability of tree-based models by directly measuring local feature interaction effects. Vlahek et al. [30] introduced an iterative approach to learning explainable features, where new features are generated with each iteration and high quality dissimilars are selected. Dinh et al. [31] suggested a consistent feature selection for analytic deep neural networks. Cancela et al. [32] proposed E2E-FS, a feature selection algorithm providing both precision and explainability in a smart way. Wang et al. [33] proposed RC-Explainer, a Reinforced Causal Explainer for Graph Neural Networks. A poweful framework that generate faithful and concise explanations

to unseen graphs. Moritz et al. [34] introduced CoDA Nets, a powerful classifiers with a high degree of inherent interpretability. Table 1 gives a concise overview of other existing explainability algorithms.

3 Background on LIME

The basic idea of LIME is to replace a data instance x by its interpretable representations x' thanks to a mapping function $\Phi(x)$. For example, an image will be represented as a group of super-pixels, a text as binary vectors indicating the presence or the absence of a word. The interpretable representations are more easily understandable and close to human intuition. Then, x' is perturbed to generate a set of new instances. The black box model is used to make predictions of generated instances from x' which are weighted according to their dissimilarity with x'. Now, for the explanation purpose, an interpretable model, such as linear models, is trained on weighted data to explain prediction locally (see, Algorithm 1).

Algorithm 1 Sparse linear explanations LIME.

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Lengths of explanation K
$Z \longleftarrow \{\}$
for ($i \in \{1, 2, 3, ..., N\}$)
  $z'_i \longleftarrow$ sample-around(x')
  $Z \longleftarrow Z \cup \langle z'_i, f(z_i), \pi_x(z_i)\rangle$
end for
$w \longleftarrow K$-Lasso(Z, K), z'_i as features, $f(z)$ as target
return w

3.1 LIME: fidelity-interpretability trade-off

Authors in [11] define an explanation as a model $g \in G$, where G is a class of potentially interpretable models (e.g., linear models, decision trees). Let $\Omega(g)$ be a measure of complexity (as opposed to interpretability) of the explanation g. For example, for linear models $\Omega(g)$ may be the number of non-zero weights. The model being explained is denoted by $f : \mathbb{R}^d \longrightarrow \mathbb{R}$. Let now π_x defines a locality around x and $\mathcal{L}(f; g; x)$ be a measure of how unfaithful g is in approximating f in the locality π_x. The explanation produced by LIME is then obtained by the following minimization problem [11]:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, \mathcal{L}(f; g; \pi_x) + \Omega(g) \tag{1}$$

3.2 Explaining global behavior

LIME explains a single prediction locally. Then, it picks K explanations which must be representative to show to the user. The *Submodular Pick* is used to choose instances to be inspected for global understanding. The quality of selected instances is critical to get insights from the model in a reasonable time (see, Algorithm 2). Let $\mathbf{X}$ (with $|\mathbf{X}| = n$) be the set of instances to explain, Algorithm 2 calculates $\mathbf{W} \in \mathbb{R}^{n \times d'}$ an explanation matrix using

Table 1 Overview of explainability methods

Name	Ref	Year	Data type	Explanator	Framework	Github Link
LIME	Ribeiro et al. [11]	2016	Any	Perturbation-based	Tensorflow	https://github.com/marcotcr/lime
DeepLIFT	Shrikumar et al. [35]	2016	Image	Correlation-Score	Tensorflow	https://github.com/kundajelab/deeplift
SHAP	Lundberg et al. [9]	2017	Any	Perturbation-based	Tensorflow	https://github.com/shap/shap
PALM	Krishnan et al. [36]	2017	Any	Decision Tree	-	-
Anchor	Ribeiro et al. [37]	2018	Tabular	Adversarial-based	Tensorflow	https://github.com/marcotcr/anchor
ND	Bolei et al. [38]	2018	Image	Concept-based	Caffe/Pytorch	https://github.com/CSAILVision/NetDissect
CXPlain	Schwab et al. [39]	2019	Image	Perturbation-based	tensorflow	https://github.com/d909b/cxplain
BAM	Yang et al. [40]	2019	Image	Heatmap	tensorflow	https://github.com/google-research-datasets/bam
CFX	Albini et al. [41]	2020	Tabular	Other	-	-
NAM	Agarwal et al. [42]	2020	Image	Other	tensorflow	https://neural-additive-models.github.io/
XAI-Bench	Liu et al. [43]	2021	Image	Heatmap	tensorflow	https://github.com/abacusai/xai-bench
XAI-Eval	Graziani et al. [44]	2021	Image	Heatmap	tensorflow	https://github.com/maragraziani/XAI_evaluation
OpenXAI	Agarwal et al. [45]	2022	Tabular/Structured	Heatmap	torch	https://github.com/AI4LIFE-GROUP/OpenXAI
IntepretDL	Li et al. [46]	2022	Image/Text	Feature importance	sklearn	https://github.com/PaddlePaddle/InterpretDL
BARBE	Motallebi et al. [47]	2023	Text	Feature importance	scikit-learn	https://github.com/changyaochen/rbo
GraphXAI	Agarwal et al. [48]	2023	Graph	Feature importance	torch	https://github.com/mims-harvard/GraphXAI

each individual explanation given by Algorithm 1. Then, it computes (I_j) global feature importance for each column j in W, such that the highest importance score is given to the feature explaining an important number of different instances. Submodular Pick aims then at finding the set of instances V, $|V| < \mathbf{B}$ that scores the highest coverage, defined as the function which calculates total importance of features in at least one instance. Finally, greedy algorithm is used to build V by adding the instance with highest marginal coverage gain.

Algorithm 2 Submodular Pick (SP).

Require: Instances X, Budget **B**
for (all x_i in X)
  $\mathbf{W}_i \longleftarrow explain(x_i, x'_i)$ {Using LIME}
end for
for $j \in 1...d'$ **do**
  $I_j \longleftarrow \sqrt{\sum_{i=1}^{n} |\mathbf{W}_{ij}|}$ {Compute the feature importance}
end for
V $\longleftarrow$ {}
while$|V| < \mathbf{B}$
  $V \longleftarrow V \bigcup argmax_i \quad c(V \bigcup \{i\}, \mathbf{W}, I)$
end while
return

4 Proposed approach

The approach we propose in this paper consists of two sequential phases (see Figure 1). The first is to use LIME (without loss of generality, any other explainer can be used) to obtain the explanations of the predictions for the test data. While the second phase focuses on global explainability by co-selecting the most important test instances and features. Thus, we provide a global understanding of the black-box model.

4.1 Notation

Table 2 summarizes the significant notations used in this paper. Let $\mathbf{E}$ be an explanation matrix of n instances and m features. The $l_{2,1}$-norm of $\mathbf{E}$ is:

$$\| \mathbf{E} \|_{2,1} = \sum_{i=1}^{m} \| \mathbf{E}_i \|_2 = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} \mathbf{E}_{ij}^2} \tag{2}$$

and its Frobenius norm $(l_{2,2})$ is:

$$\| \mathbf{E} \|_F = \left(\sum_{i=1}^{m} \| \mathbf{E}_i \|_2^2 \right) = \left(\sum_{i=1}^{m} \left(\sum_{j=1}^{n} \mathbf{E}_{ij}^2 \right) \right)^{1/2} \tag{3}$$

4.2 Explanation space

Let f be a black box model, and $\mathbf{X}$ a test dataset of n instances and $\Phi(\mathbf{X}) = \mathbf{X}'$ its interpretable representation in $\mathbb{R}^p$. First, to obtain an individual explanation of the prediction made by f for each instance x_i we use LIME by fitting a linear model on a generated dataset around x'_i,

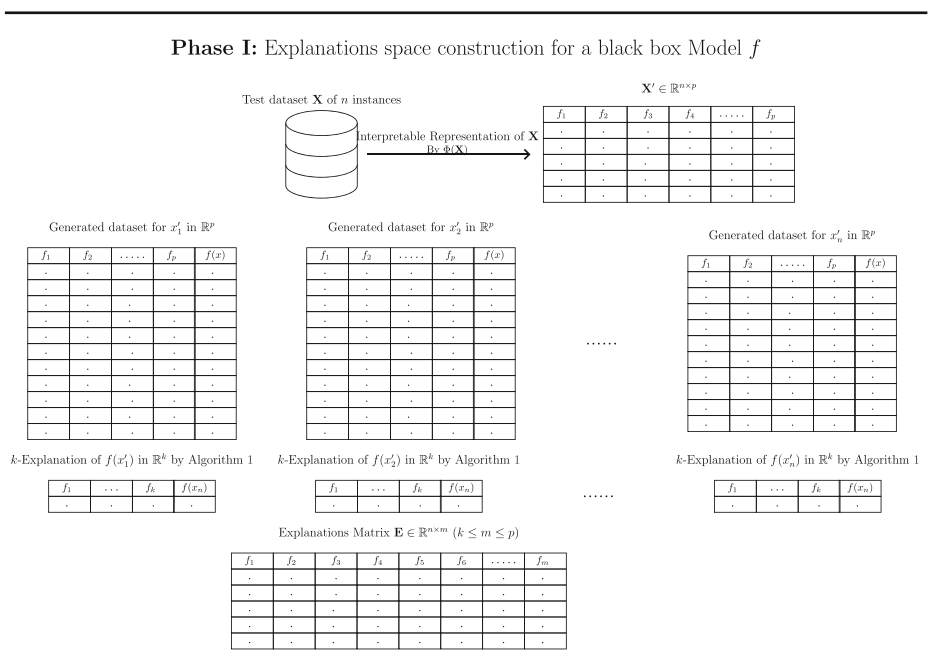

Figure 1 The proposed framework for a global explanation using a co-selection of features and instances

Table 2 Summary of symbols and notations

Symbol	Definition
n	Number of instances
m	Number of features
h	Dimesion of the low dimensional space
$\mathbf{E} \in \mathbb{R}^{n \times m}$	Explanation matrix
$\mathbf{A} \in \mathbb{R}^{n \times n}$	Pairwise similarity matrix over $\mathbf{E}$
$\mathbf{R} \in \mathbb{R}^{n \times h}$	Instance coefficient matrix
$\mathbf{W} \in \mathbb{R}^{m \times h}$	Feature coefficient matrix
$\mathbf{Z} \in \mathbb{R}^{n \times h}$	Eigen-decomposition of $\mathbf{A}$
$\|\cdot\|_F$; $\|\cdot\|_{2,1}$	Matrix norms

the interpretable representation of x_i . Thus, for each instance x_i, we obtain an explanation of length k ($k < p$). It is worthy to note that the length is a parameter set by the user and corresponds to the number of features retained. Once the individual explanations have been obtained, we construct an explanation space represented by $\mathbf{E} \in \mathbb{R}^{n \times m}$, where the dimension m of the explanations space corresponds to the union of the k features of each explanation. We illustrate this step with the following example:

Example Let $\mathbf{X}'$ be the interpretable representation of 3 instances in $\mathbb{R}^{500}$, and $k = 3$ be the length of the explanation desired for these three instances. By performing LIME algorithm on $\mathbf{X}'$, we obtain 3 explanations of length 3:

$$e_i = \begin{cases} e_1 = \{(f_1, 0.5), (f_{25}, 0.9), (f_4, 0.1)\} \\ e_2 = \{(f_{17}, 0.2), (f_6, 0.3), (f_{78}, 0.4)\} \\ e_3 = \{(f_{500}, 0.8), (f_{25}, 0.7), (f_1, 0.25)\} \end{cases} \tag{4}$$

where e_1, e_2, and e_3 are the explanations of x'_1, x'_2 and x'_3 respectively. Thus, the matrix $\mathbf{E} \in \mathbb{R}^{3 \times 7}$ can be seen as the concatenation of all the explanations and the union of the set of features obtained by each explanation. Note that the dimension m here is equal to 7.

4.3 Global explicability by co-selection

Understanding the model's decisions globally remains a complex task. In fact, some approaches like LIME were extended to face this complexity by only picking a group of individual explanations. In this paper, we advocate a method allowing global explainability by co-selecting the most important instances and features over the explanations provided by any explainer. The idea is to find a residual matrix $\mathbf{R}$ and a transformation matrix $\mathbf{W}$, which transforms high-dimensional explanations data $\mathbf{E}$ to low dimensional data $\mathbf{E}\,\mathbf{W}$, to maximize the global similarity between $\mathbf{E}$ and $\mathbf{E}\,\mathbf{W}$. After the optimal $\mathbf{W}$ and $\mathbf{R}$ have been obtained, the original features and instances are ranked, based on the $\ell_{2,1}$-norm values of the rows of $\mathbf{R}$ and $\mathbf{W}$, and the top features and instance are selected accordingly.

4.4 Co-selection pick (CoSP)

To perform a co-selection of instances and features on the explanations matrix, we must minimize the following problem as pointed out in [49]:

$$\min_{\mathbf{W},\mathbf{R}} \| \mathbf{E}\,\mathbf{W} - \mathbf{R}^T - \mathbf{Z} \|_F^2 + \lambda \| \mathbf{W} \|_{2,1} + \beta \| \mathbf{R} \|_{2,1} \tag{5}$$

Where:

- $\mathbf{Z}$ is the eigen-decomposition of the pairwise similarity matrix, $\mathbf{A}$, computed over the explanation matrix $\mathbf{E}$. Note that the similarity matrix $\mathbf{A}$ can be calculated in supervised fashion (e.g. adjacency matrix, fully binary matrix) if the labels of test instances are available, or in unsupervised mode as follows:

$$\mathbf{A}_{ij} = e^{-\frac{\|e_i - e_j\|^2}{2\delta^2}} \tag{6}$$

- $\mathbf{R} = \mathbf{W}^T\,\mathbf{E}^T - \mathbf{Z}^T - \Theta.$, is a residual matrix and Θ is a random matrix, usually assumed to be multi-dimensional normal distribution [50]. Note that the matrix $\mathbf{R}$ is a good indicator of outliers and less important and irrelevant instances in a dataset according to [51, 52].

- λ and β are regularization parameters, used to control the sparsity of $\mathbf{W}$ and $\mathbf{R}$ respectively; and δ is a parameter for the RBF kernel used to compute the matrix $\mathbf{A}$ in the unsupervised mode in (6).

The first term of the objective in (5) exploits the $\mathbf{E}$ structure by preserving the pairwise explanations similarity while the second and third terms are used to perform feature selection and instance selection, respectively.

Optimization

In order to minimize (5), we adopt an alternating optimization over $\mathbf{W}$ and $\mathbf{R}$ as in[49], by solving two reduced minimization problems :

Problem 1 Minimizing (5) by fixing $\mathbf{R}$ to compute $\mathbf{W}$ (for feature selection). To solve this problem, we consider the lagrangian function of (5):

$$\mathcal{L}_{\mathbf{W}} = trace(\mathbf{W}^T\,\mathbf{E}^T\,\mathbf{E}\,\mathbf{W} - 2\mathbf{W}^T\mathbf{E}^T(\mathbf{R}^T + \mathbf{Z})) + \lambda \parallel \mathbf{W} \parallel_{2,1} . \tag{7}$$

Then, we calculate the derivative of $\mathcal{L}_W$ *w.r.t* $\mathbf{W}$:

$$\frac{\partial \mathcal{L}_{\mathbf{W}}}{\partial \mathbf{W}} = 2\,\mathbf{E}^T\,\mathbf{E}\,\mathbf{W} - 2\,\mathbf{E}^T(\mathbf{R}^T + \mathbf{Z}) + 2\lambda\mathcal{D}_{\mathbf{W}}\mathbf{W}. \tag{8}$$

Where $\mathcal{D}_W$ is a $(m \times m)$ diagonal matrix with the i^{th} element equal to $\frac{1}{2\|\mathbf{W}(i,:)\|_2}$. Subsequently, we set the derivative to zero to update $\mathbf{W}$:

$$\mathbf{W} = (\mathbf{E}^T\mathbf{E} + \lambda\mathcal{D}_{\mathbf{W}})^{-1}\,\mathbf{E}^T(\mathbf{R}^T + \mathbf{Z}) \tag{9}$$

Problem 2 Minimizing (5) by fixing $\mathbf{W}$ to compute the solution for $\mathbf{R}$ (for explanation selection). To solve this problem, we consider the Lagrangian function of (5):

$$\mathcal{L}_{\mathbf{R}} = trace(\mathbf{R}^T\mathbf{R} - 2\mathbf{R}^T(\mathbf{E}\,\mathbf{W} - \mathbf{Z})) + \beta \parallel \mathbf{R} \parallel_{2,1} . \tag{10}$$

Then, we calculate the derivative of $\mathcal{L}_{\mathbf{R}}$ *w.r.t* $\mathbf{R}$:

$$\frac{\partial \mathcal{L}_{\mathbf{R}}}{\partial \mathbf{R}} = 2\mathbf{R}^T - 2(\mathbf{E}\,\mathbf{W} - \mathbf{Z}) + 2\beta\mathcal{D}_{\mathbf{R}}\mathbf{R}^T. \tag{11}$$

Where $\mathcal{D}_{\mathbf{R}}$ is a $(n \times n)$ diagonal matrix with the i^{th} element equal to $\frac{1}{2\|\mathbf{R}^T(i,:)\|_2}$.
Subsequently, we set the derivative to zero to update $\mathbf{B}$:

$$\mathbf{R} = (\mathbf{E}\,\mathbf{W} - \mathbf{Z})^T((\mathbf{I} + \beta\mathcal{D}_{\mathbf{R}})^{-1})^T \tag{12}$$

Where $\mathbf{I}$ is a $(n \times n)$ identity matrix. All of the above developments are summarized on Algorithm 3.

Algorithm 3 Co-Selection pick (CoSP).

1: **Require:** Instances $\mathbf{X}$, Budget $\mathbf{B}$ and $\mathbf{L}$, hyper-parameters: $\lambda, \beta, \delta, h$.
2: **for** (all x_i in $\mathbf{X}$)
3: $e_i \longleftarrow explain(x_i, x_i')$ {Using LIME}
4: **end for**
5: Build the explanations matrix $\mathbf{E}$ (see Figure 2).
6: Calculate $\mathbf{A}$ {according to (6) for unsupervised mode or as adjacency matrix for supervised mode}.
7: Eigen-decomposition of $\mathbf{A}$ such as $\mathbf{A} = \mathbf{Z}\mathbf{Z}^T$.
8: Initialize $\mathcal{D}_{\mathbf{W}}$ and $\mathcal{D}_{\mathbf{R}}$ as identity matrices.
9: **repeat**
10: Update $\mathbf{W}$ by $(\mathbf{E}^T\mathbf{E} + \lambda\mathcal{D}_W)^{-1}\mathbf{E}^T(\mathbf{R}^T + \mathbf{Z})$
11: Update $\mathbf{R}$ by $(\mathbf{E}\mathbf{W} - \mathbf{Z})^T((\mathbf{I} + \beta\mathcal{D}_{\mathbf{R}}))^{-1})^T$
12: Update $\mathcal{D}_{\mathbf{R}}$ and $\mathcal{D}_{\mathbf{W}}$.
13: **until** $Convergence$
14: Rank the features according to $\| \mathbf{W}(j,:) \|_2$ in descending order, and the instances according to $\| \mathbf{R}(:,i) \|_2$ in ascending order.
15: Pick the top $\mathbf{B}$ instances and the top $\mathbf{L}$ features.

4.5 Algorithm analysis

In the Algorithm 3, the final user expects a selection of $\mathbf{B}$ instances (e.g., explanations) and $\mathbf{L}$ features which are most relevant to provide global explanation of the model. In order to achieve this, CoSP requires four hyper-parameters λ, β , δ and h that will be used later on to build the set of chosen instances and features. Firstly, we build the explanations matrix $\mathbf{E}$ using any explainer, in our case we use LIME. Secondly, we compute the similarity matrix $\mathbf{A}$ either in supervised mode (as adjacency matrix or a binary matrix) or in an unsupervised way according to the availability of the labels of the test instances $\mathbf{X}$. Then, we eigen-decompse $\mathbf{A}$ to find $\mathbf{Z}$. From line 9 to line 13 $\mathbf{W}$ and $\mathbf{R}$ are updated until convergence according to (9) and (12). Following the alternate optimization, we rank the instances and the features according to $\mathbf{R}$ and $\mathbf{W}$ respectively. So, the higher the norm of $\| \mathbf{R}(:,j) \|_2$, the more the j^{th} explanation is not representative, while the higher the norm $\| \mathbf{W}(i,:) \|_2$, the more the i^{th} feature is important. The computational complexity of Algorithm 3 is presented by the following lemma.

Lemma CoSP is computed in time of $\mathcal{O}(nmh + m^3 + n^3 + nm^2 + n^2h)$.

Proof The time complexity of CoSP essentially depends on the rule of (9) as well as the rule of (12). These two rules are for updating the two matrices $\mathbf{W}$ and $\mathbf{R}$ which consists of some matrix multiplication and inversion operations at each iteration. Specifically, the computation of the derivative w.r.t $\mathbf{W}$ requires $\mathcal{O}(nmh + nm^2 + m^3)$. The derivative w.r.t $\mathbf{R}$ needs $\mathcal{O}(nmh + n^2h + n^3)$.

	f_1	f_4	f_6	f_{17}	f_{25}	f_{78}	f_{500}
x_1	0.5	0.1	0	0	0.9	0	0
x_2	0	0	0.3	0.2	0	0.4	0
x_3	0.25	0	0	0	0.7	0	0.8

Figure 2 Explanation matrix $\mathbf{E}$ (this matrix is given as input for CoSP Algorithm 3)

5 Experiments

In this section, we conduct some experiments to validate our framework on some known sentiment datasets.

5.1 Datasets and compared methods

We use a binary sentimental classification dataset. Sentimental analysis is the task of analyzing people's opinions, reviews, and comments presented as textual data. It gives intuition about different points of view and feedback by detecting relevant words used to express specific sentiments [53]. Today, companies rely on sentimental analysis to improve their strategy. People's opinions are collected from different sources like Facebook, Tweets, product reviews and processed in order to understand customer's needs and improve marketing plans. When the sentiment is divided into positive and negative ones, it is called binary sentimental analysis which is the most common type and the one used in our case. While multi-class sentiment analysis classifies text into groups of possible labels. We use multi-Domain Sentiment Dataset[1], which contains multiple domains reviews (books and dvd) from Amazon.com, where for each type of product there are hundred of thousands of collected reviews. Then, we use an experiment introduced in [11] which aims to evaluate if explanations could help a simulated user to recognize the best model from a group of models having the same accuracy on validation set. In order to do this, a new dataset will be generated by adding 10 artificial features to the train and validation set from original public dataset (reviews). For the train examples, each of those features appears in 10% of instances in one class and in 20% of the other class. In the test examples, an artificial feature appears in 10 % of examples in both classes. This represents the case of having spurious correlations in the data introduced by non informative features.

Furthermore, we train pairs of classifiers until their validation accuracy is within 0.1% of each other. However, their test accuracy should differ by at least 5% which will make one classifier better than the other. Then, we explain global behaviors of both classifiers using our proposed approach CoSP.

We compare CoSP against Random, Greedy [25], Parzen [26] and LIME approches [11] combined with Submodular Pick (SP) and Random Pick (RP). In the following, we briefly describe each approach.

- **Random** randomly chooses the features as an explanation.
- **Greedy** removes features highly contributing to the predicted class until the prediction changes.
- **Parzen** uses parzen windows to globally approximate the classifier
- **Lime** explains the classifier predictions by approximating it locally with an interpretable model.

In the experiment, the explanations were obtained with the above four local explainability techniques. Then, the global explainability approaches CoSP, SP or RP were used to select the relevant instances.

[1] https://www.cs.jhu.edu/~mdredze/datasets/sentiment/

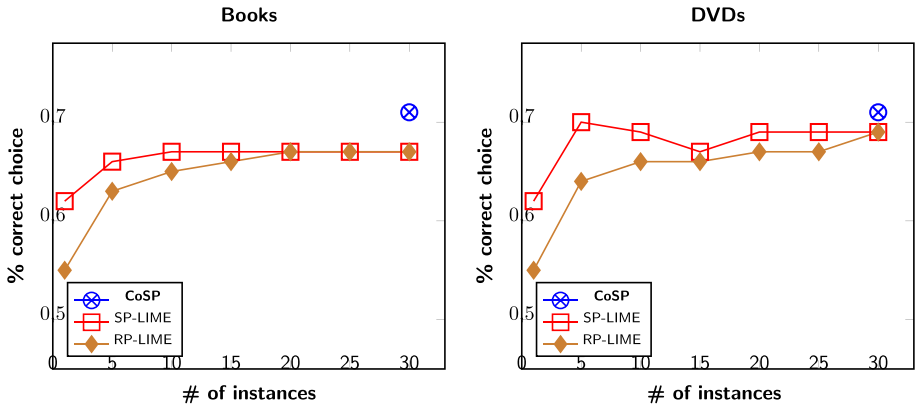

Figure 3 Accuracy of picking the correct classifier over two datasets `Books` and `DVDs`

5.2 Experimental setting

To validate our approach, we use the same experimental setting introduced in [54] by selecting top five important features per class chosen as most relevant ones to be considered for the classification task. Global approach is validated if it selects distinguishing features. Four hyper-parameters necessary for CoSP have been set as follows: $\lambda \approx 2.11$, $\beta \approx 61.79$, $\delta = 1$ and $h = 17000$ (which stands for the number of features selected by CoSP). Parallely, the parameters configuration of compared methods is as follows: K = 10 words in each explanation and B = 10 instances.

5.3 Evaluation and results

In this section, we present the main results of our experiments. Figures 3, 4, 5 and 6 show the experimental results over two datasets, *Books* and *DVDs*. We summarize the main observations of the experimental results in the following points.

Figure 4 Top 5 features per class picked by CoSP global approach for review's binary classification on books dataset

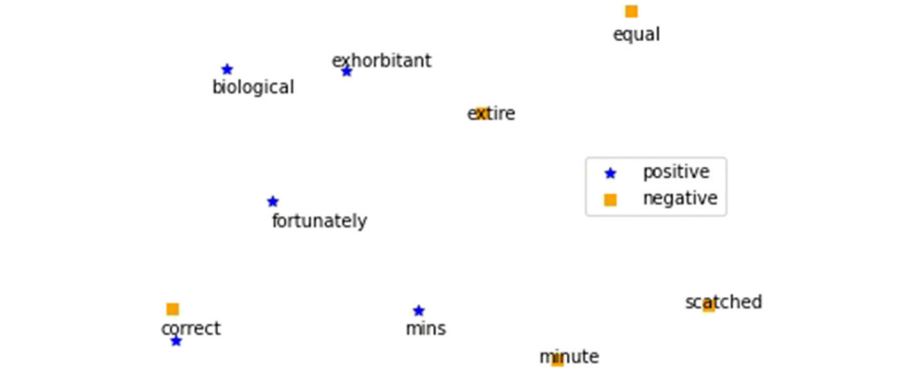

Figure 5 Top 5 features per class picked by CoSP global approach for review's binary classification on DVDs dataset

- In terms of Accuracy, the LIME combined with either Co-Selection Pick (CoSP) or Submodular Pick (SP-LIME) outperforms other comparison algorithms. It means that the explanations provided by LIME are faithful to the models (see Figure 6).
- Regardless of the choice of the explainers, CoSP is significantly better than SP or RP, across the two data sets, followed by SP-LIME (see Figures 3 and 6).
- CoSP further improves the user's ability to select the best classifier comparing with the SP or RP (see Figures 3 and 6).
- From Figures 4 and 5, the displayed perception contains words that are meaningful in order to judge the type of comment. Features are aligned with human intuition and words with no representative meaning like stop words were not selected. Also, noisy features labeled with prefix "FAKE" added to the dataset were not deemed important.

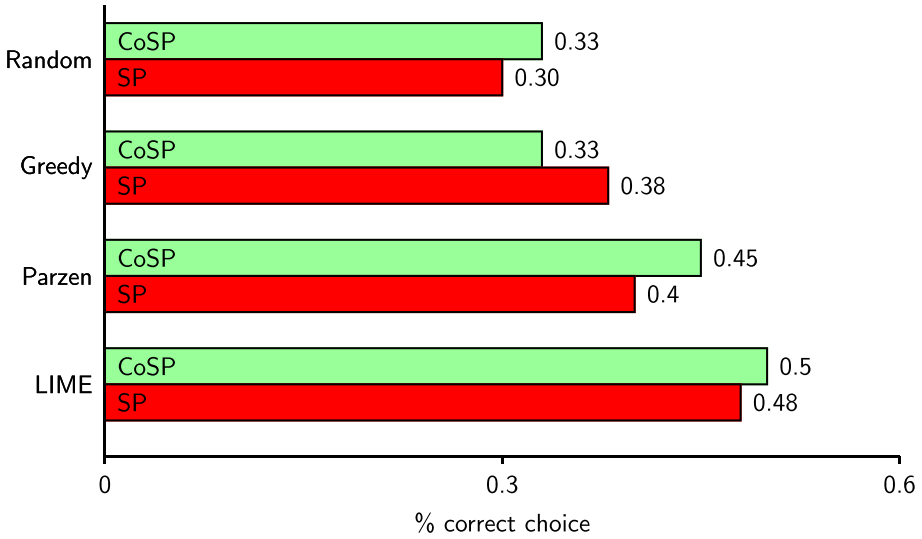

Figure 6 Accuracy of human subject in choosing between two models, using books dataset. The four method are combined with CoSP or Submodular (SP) selections

6 Conclusion

In this paper, we presented CoSP, a generic framework aiming to select individual instances in order to provide global explanation for machine learning models. We used Co-selection based on similarity as foundation to build global understanding of the black box internal logic over any local explainer. Furthermore, we conducted some experiments showing that CoSP offers representative insights. This study is a another step towards understanding machine learning models globally. For future work, we would like to explore this method in the context of time series data, as it is a challenging to find representative illustration for this type of data. The approach we proposed is independent of the type of data, since it is based on the explanations provided by a local explainer. Concerning time series, the local explainer must be capable of processing this type of data. This involves in particular the choice of an efficient representation of the time series. In the case of LIME, it is necessary to find a vector representation of the series to be able to apply LASSO and have the explanations. Among the applications on which we want to apply our approach, there is the detection of contextual anomalies in time series. The idea is then not only to detect abnormal segments in a time series but to explain why such a segment was detected as abnormal.

Author Contributions
1- Dou El Kefel Mansouri: Methodology, Software, Validation, Formal analysis, Investigation, Writing- Original draft preparation.
2- Seif-Eddine Benkabou: Conceptualization of this study, Methodology, Software, Validation, Formal analysis, Investigation, Writing- Original draft preparation.
3- Khaoula Meddahi: Conceptualization of this study, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing- Original draft preparation.
4- Allel Hadjali: Methodology, Investigation, Visualization.
5- Amin Mesmoudi: Software, Methodology, Investigation, Visualization.
6- Khalid Benabdeslem: Methodology, Investigation, Visualization.
7- Souleyman Chaib: Methodology, Investigation, Visualization.
**All authors reviewed the manuscript.

Funding Not applicable.

Availability of data and materials Not applicable.

Declarations

Conflicts of interest The authors declare that they have no confict of interest.

Ethical Approval Not applicable.

References

1. Mohaghegh, F., Murthy, J.: Machine learning and computer vision techniques to predict thermal properties of particulate composites. CoRR. **abs/2010.01968** (2020). arXiv:2010.01968
2. Holm, E.A., Cohn, R., ao, N., Kitahara, A.R., Matson, T.P., Lei, B., Yarasi, S.R.: Overview: Computer vision and machine learning for microstructural characterization and analysis. CoRR. **abs/2005.14260** (2020). arXiv:2005.14260. https://doi.org/10.1007/s11661-020-06008-4
3. Kosowski, P.: Deep learning for natural language processing and language modelling. In: 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pp. 223–228 (2018). https://doi.org/10.23919/SPA.2018.8563389

4. Shailaja, K., Seetharamulu, B., Jabbar, M.A.: Machine learning in healthcare: a review. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 910–914 (2018). https://doi.org/10.1109/ICECA.2018.8474918
5. Bistron, M., Piotrowski, Z.: Artificial intelligence applications in military systems and their influence on sense of security of citizens. Electronics. **10**(7), (2021). https://www.mdpi.com/2079-9292/10/7/871
6. Gunning, D., Aha, D.: Darpas explainable artificial intelligence (xai) program. AI. Mag. **40**(2), 44–58 (2019)
7. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), (2018). https://doi.org/10.48550/arXiv.1802.01933
8. Strumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**, 647–665 (2013)
9. Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)
10. Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.: Explainable ai for trees: From local explanations to global understanding. ArXiv. **abs/1905.04610**, (2019)
11. Ribeiro, M., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: al., B.K. (ed.) Proc. of the 22nd ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, pp. 1135–1144. ACM, ??? (2016). https://doi.org/10.1145/2939672.2939778
12. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: A data perspective. ACM computing surveys (CSUR). **50**(6), 1–45 (2017)
13. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**(Mar), 1157–1182 (2003)
14. Olvera-López, J.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Kittler, J.: A review of instance selection methods. Artif. Intell. Rev. **34**, 133–143 (2010)
15. Liu, H., Motoda, H.: On issues of instance selection. Data Min. Knowl. Disc. **6**(2), 115 (2002)
16. Li, Y.-F., Zhou, Z.-H.: Improving semi-supervised support vector machines through unlabeled instances selection. In: Proceedings of the AAAI Conference on Artificial Intelligence, **25**, pp. 386–391 (2011)
17. Kuncheva, L.I., Jain, L.C.: Nearest neighbor classifier: Simultaneous editing and feature selection. Pattern Recogn. Lett. **20**(11–13), 1149–1156 (1999)
18. Derrac, J., García, S., Herrera, F.: Ifs-coco: Instance and feature selection based on cooperative coevolution with nearest neighbor rule. Pattern Recogn. **43**(6), 2082–2105 (2010)
19. GarcíA-Pedrajas, N., De Haro-GarcíA, A., PéRez-RodríGuez, J.: A scalable approach to simultaneous evolutionary instance and feature selection. Inf. Sci. **228**, 150–174 (2013)
20. Zhao, Z., Wang, L., Liu, H., Ye, J.: On similarity preserving feature selection. IEEE Trans. Knowl. Data Eng. **25**(3), 619–632 (2011)
21. Ma, Y., Xu, X., Shen, F., Shen, H.T.: Similarity preserving feature generating networks for zero-shot learning. Neurocomputing **406**, 333–342 (2020)
22. Shang, R., Chang, J., Jiao, L., Xue, Y.: Unsupervised feature selection based on self-representation sparse regression and local similarity preserving. Int. J. Mach. Learn. & Cybernet. **10**, 757–770 (2019)
23. Code for COsP, howpublished = https://github.com/KhaoulaBF/CoSPIctai/blob/main/dvd_features_scos%20(2).ipynb,
24. Meddahi, K., Benkabou, S.-E., Hadjali, A., Mesmoudi, A., El Kefel Mansouri, D., Benabdeslem, K., Chaib, S.: Towards a co-selection approach for a global explainability of black box machine learning models. In: International Conference on Web Information Systems Engineering, pp. 97–109 (2022). Springer
25. Martens, D., Provost, F.: Explaining data-driven document classifications. MIS quarterly. **38**(1), 73–100 (2014)
26. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.-R.: How to explain individual classification decisions. J. Mach. Learn. Res. **11**, 1803–1831 (2010)
27. Ribeiro, M., Singh, S., Guestrin, C.: Fairness, Accountability, and Transparency in Machine Learning, paper 'Why Should I Trust You?' Explaining the Predictions of Any Classifier. https://www.fatml.org/schedule/2016/presentation/why-should-i-trust-you-explaining-predictions (2016)
28. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. J. Artif. Intell. Res. **70**, 245–317 (2021)
29. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I.: From local explanations to global understanding with explainable ai for trees. Nat. Mach. Intell. **2**(1), 56–67 (2020)

30. Vlahek, D., Mongus, D.: An efficient iterative approach to explainable feature learning. IEEE Trans. Neural Netw. & Learn, Syst (2021)
31. Dinh, V.C., Ho, L.S.: Consistent feature selection for analytic deep neural networks. Adv Neural Inf. Proc. Syst. **33**, 2420–2431 (2020)
32. Cancela, B., Bolón-Canedo, V., Alonso-Betanzos, A.: E2e-fs: An end-to-end feature selection method for neural networks. IEEE Trans. Pattern Anal. & Mach, Intell (2022)
33. Wang, X., Wu, Y., Zhang, A., Feng, F., He, X., Chua, T.-S.: Reinforced causal explainer for graph neural networks. IEEE Trans. Pattern. Anal. & Mach, Intell (2022)
34. Böhle, M., Fritz, M., Schiele, B.: Optimising for interpretability: Convolutional dynamic alignment networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2022)
35. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv:1605.01713 (2016)
36. Krishnan, S., Wu, E.: Palm: Machine learning explanations for iterative debugging. In: Proceedings of the 2Nd Workshop on Human-in-the-loop Data Analytics, pp. 1–6 (2017)
37. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proc. AAAI Conf. Artif. Intell. vol. 32 (2018)
38. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE Trans. Pattern. Anal. & Mach. Intell. **41**(9), 2131–2145 (2018)
39. Schwab, P., Karlen, W.: Cxplain: Causal explanations for model interpretation under uncertainty. Adv. Neural Inf. Proc. Syst. **32** (2019)
40. Yang, M., Kim, B.: Benchmarking attribution methods with relative feature importance. arXiv:1907.09701 (2019)
41. Albini, E., Rago, A., Baroni, P., Toni, F.: Relation-based counterfactual explanations for bayesian network classifiers. In: IJCAI, pp. 451–457 (2020)
42. Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., Hinton, G.E.: Neural additive models: Interpretable machine learning with neural nets. Adv. Neural Inf. Proc. Syst. **34**, 4699–4711 (2021)
43. Liu, Y., Khandagale, S., White, C., Neiswanger, W.: Synthetic benchmarks for scientific research in explainable machine learning. In: Adv. Neural Inf. Proc. Syst. Datasets Track. (2021)
44. Graziani, M., Lompech, T., Müller, H., Andrearczyk, V.: Evaluation and comparison of cnn visual explanations for histopathology. In: Proceedings of the AAAI Conference on Artificial Intelligence Workshops (XAI-AAAI-21), Virtual Event, pp. 8–9 (2021)
45. Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., Lakkaraju, H.: Openxai: Towards a transparent evaluation of model explanations. Adv. Neural Inf. Proc. Syst. **35**, 15784–15799 (2022)
46. Li, X., Xiong, H., Li, X., Wu, X., Chen, Z., Dou, D.: Interpretdl: explaining deep models in paddlepaddle. J. Mach. Learn. Res. **23**(1), 8969–8974 (2022)
47. Motallebi, M., Anik, M.T.A., Zaïane, O.R.: Explaining decisions of black-box models using barbe. In: Int. Conf. Database & Expert Syst. Appl. pp. 82–97 (2023). Springer
48. Agarwal, C., Queen, O., Lakkaraju, H., Zitnik, M.: Evaluating explainability for graph neural networks. Scientific Data. **10**(1), 144 (2023)
49. Benabdeslem, K., Mansouri, D.E.K., Makkhongkaew, R.: scos: Semi-supervised co-selection by a similarity preserving approach. IEEE Trans. Knowl. Data Eng. **34**(6), 2899–2911 (2022). https://doi.org/10.1109/TKDE.2020.3014262
50. She, Y., Owen, A.-B.: Outlier detection using nonconvex penalized regression. CoRR. **abs/1006.2592** (2010). arXiv:1006.2592
51. Tong, H., Lin, C.: Non-negative residual matrix factorization with application to graph anomaly detection. In: Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA, pp. 143–153. SIAM / Omnipress, ??? (2011)
52. Tang, J., Liu, H.: Coselect: Feature selection with instance selection for social media data. In: Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA, pp. 695–703. SIAM, ??? (2013)
53. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning-based text classification. ACM Computing Surveys (CSUR). **54**, 1–40 (2021)
54. Linden, I.-V.-D., Haned, H., Kanoulas, E.: Global aggregations of local explanations for black box models. CoRR. **abs/1907.03039**, (2019). arXiv:1907.03039

Authors and Affiliations

Dou El Kefel Mansouri[1] · Seif-Eddine Benkabou[2] · Khaoula Meddahi[2] · Allel Hadjali[3] · Amin Mesmoudi[2] · Khalid Benabdeslem[4] · Souleyman Chaib[5]

Seif-Eddine Benkabou
seif.eddine.benkabou@univ-poitiers.fr

Khaoula Meddahi
kh.meddahi@gmail.com

Allel Hadjali
allel.hadjali@ensma.fr

Amin Mesmoudi
amin.mesmoudi@univ-poitiers.fr

Khalid Benabdeslem
khalid.benabdeslem@univ-lyon1.fr

Souleyman Chaib
s.chaib@esi-sba.dz

1 Department of Biology, Ibn Khaldoun University, 14000 Tiaret, Algeria

2 Laboratory of Computer Science and Automatic Control for Systems (LIAS)/École Nationale Supérieure de Mécanique et d'Aérotechnique Poitiers Futuroscope (ISAE-ENSMA), University of Poitiers, 86000 Poitiers, France

3 LIAS/ENSMA, 86000 Poitiers, France

4 Laboratoire d'InfoRmatique en Image et Systémes d'information (LIRIS), University of Lyon 1, 69000 Villeurbanne, France

5 LabRi Laboratory, École Supérieure en Informatique, 22000 Sidi Bel Abbés, Algeria