Towards a Co-selection Approach for a Global Explainability of Black Box Machine Learning Models

Khoula Meddahi[1], Seif-Eddine Benkabou[2(✉)], Allel Hadjali[1], Amin Mesmoudi[2], Dou El Kefel Mansouri[3], Khalid Benabdeslem[4], and Souleyman Chaib[5]

[1] ISAE-ENSMA (LIAS), Chasseneuil-du-Poitou, France
allel.hadjali@ensma.fr
[2] Université de Poitiers (LIAS), Poitiers, France
{seif.eddine.benkabou,amin.mesmoudi}@univ-poitiers.fr
[3] University Ibn Khaldoun Tiaret, Tiaret, Algeria
douelkefel.mansouri@univ-tiaret.dz
[4] Université de Lyon (LIRIS), Lyon, France
khalid.benabdeslem@univ-lyon1.fr
[5] École Supérieure d'Informatique (ESI-SBA), Sidi Bel Abbès, Algeria
s.chaib@esi-sba.dz

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 · Global aggregation

R. Chbeir et al. (Eds.): WISE 2022, LNCS 13724, pp. 97–109, 2022.
https://doi.org/10.1007/978-3-031-20891-1_8

1 Introduction

Nowadays, a wide range of real-life applications such as computer vision [5,11], speech processing, natural language understanding [6], health [14], and military fields [2,4] make use of Machine Learning (ML) models for decision making or prediction/classification purpose. 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 [3]. 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 [16], 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 [8].

Let us also mention the LIME (Local Interpretable Model-agnostic Explanations) method which is one of the most famous local explainable models [13].

LIME 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 [13] 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 [13], 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.

The paper is structured as follows. Section 2 provides a necessary background on LIME method. In Sect. 3, we present our approach allowing for a global explanation of black box ML models. Section 4 shows the preliminary experiments

done to validate our proposal. In Sect. 5, we conclude the paper and draw some research lines for future work.

2 Background on LIME

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 [12]. Faithfulness represents how the explanation is describing the reality of the model. Model-agnostic methods are used for any type of model. LIME introduced by Ribeiro et al [13], is one of the well-known examples of such methods. It is a framework which explains a prediction by approximating it locally using an interpretable model (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
1: $Z \longleftarrow \{\}$
2: **for** ($i \in \{1, 2, 3, ..., N\}$) **do**
3: $\quad z'_i \longleftarrow$ sample-around(x')
4: $\quad Z \longleftarrow Z \cup \langle z'_i, f(z_i), \pi_x(z_i)\rangle$
5: **end for**
6: $w \longleftarrow K$-Lasso$(Z, K), z'_i$ as features, $f(z)$ as target
7: return w

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.

2.1 LIME: Fidelity-Interpretability Trade-off

Authors in [13] 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 [13]:

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

2.2 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* explained in Algorithm 2 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.

Algorithm 2. Submodular Pick (SP)

Require: Instances X, Budget **B**
for (all x_i in X) **do**
  $\mathbf{W}_i \longleftarrow explain(x_i, x_i')$ {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 $|V| < \mathbf{B}$ **do**
  $V \longleftarrow V \bigcup argmax_i \quad c(V \bigcup \{i\}, \mathbf{W}, I)$
end while
return

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 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.

3 Proposed Approach

The approach we propose in this paper consists of two sequential phases (see Fig. 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.

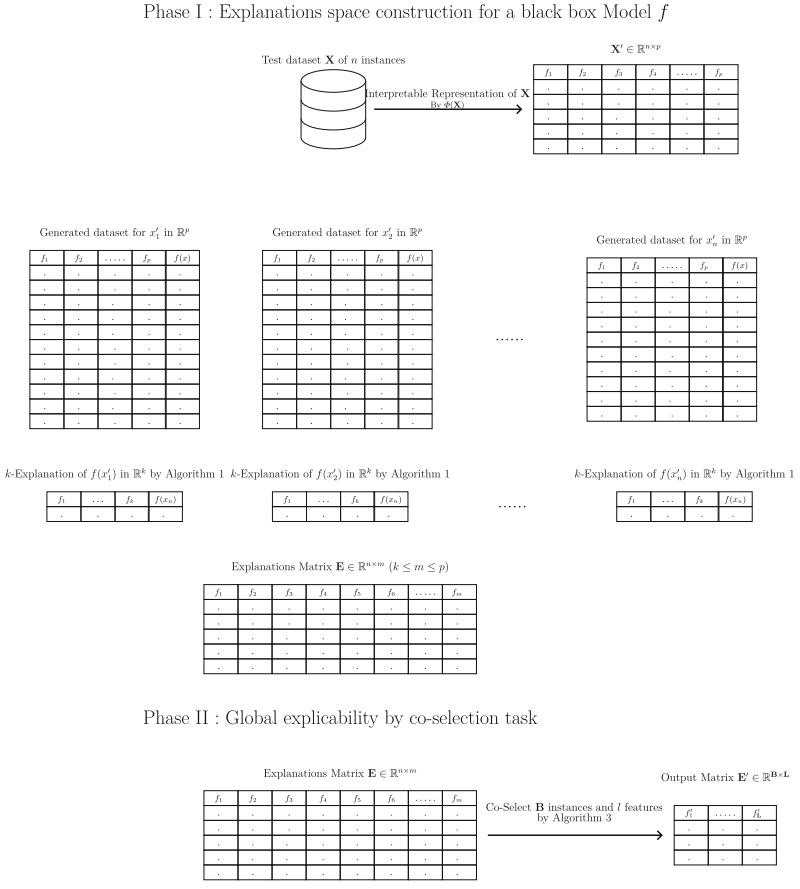

Fig. 1. The proposed framework for a global explanation using a co-selection of features and instances.

3.1 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, 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 1
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:

$$e_i = \begin{cases} e_1 = \{(f_1, 0.5), (f_{25}, 0.9), (f_4, 0.1)\} \\ e_2 = \{(f_{17}, 0.2), (f_6, 0.3), (f_{78}, 0.4)\} \\ e_3 = \{(f_{500}, 0.8), (f_{25}, 0.7), (f_1, 0.25)\} \end{cases} \tag{2}$$

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.

	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

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

3.2 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 $\mathbf{EW}$, to maximize the

global similarity between $\mathbf{E}$ and $\mathbf{EW}$. 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.

3.3 Notation

First, we present the notation we use 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:

$$\| \mathbf{E} \|_{2,1} = \sum_{i=1}^{m} \| \mathbf{E}_i \|_2 = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} \mathbf{E}_{ij}^2} \tag{3}$$

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

$$\| \mathbf{E} \|_F = \left(\sum_{i=1}^{m} \| \mathbf{E}_i \|_2^2 \right) = \left(\sum_{i=1}^{m} \left(\sum_{j=1}^{n} \mathbf{E}_{ij}^2 \right) \right)^{1/2} \tag{4}$$

Table 1. Summary of symbols and notations

Symbol	Definition
n	Number of instances
m	Number of features
h	Dimension 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

3.4 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 [1]:

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

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 [15]. Note that the matrix $\mathbf{R}$ is a good indicator of outliers and less important and irrelevant instances in a dataset according to [17,18].
- 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 Eq. (6).

The first term of the objective in Eq. (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. In order to minimize Eq. (5), we adopt an alternating optimization over $\mathbf{W}$ and $\mathbf{R}$ as in [1], by solving two reduced minimization problems:

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

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

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 $\mathbf{X}$) **do**
3: $e_i \longleftarrow explain(x_i, x_i')$ {Using LIME}
4: **end for**
5: Build the explanations matrix $\mathbf{E}$ (see Fig. 2).
6: Calculate $\mathbf{A}$ {according to Eq. (6) for unsupervised mode or as adjacency matrix for supervised mode}.
7: Eigen-decomposition of $\mathbf{A}$ such as $\mathbf{A} = \mathbf{ZZ}^T$.
8: Initialize $\mathcal{D}_W$ 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 $(\mathbf{EW} - \mathbf{Z})^T((\mathbf{I} + \beta\mathcal{D}_{\mathbf{R}}))^{-1})^T$
12: Update $\mathcal{D}_{\mathbf{R}}$ and $\mathcal{D}_W$.
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.

3.5 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-decompose $\mathbf{A}$ to find $\mathbf{Z}$. From line 9 to line 13 $\mathbf{W}$ and $\mathbf{R}$ are updated until convergence according to Eqs. (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.

4 Experiments

In this section, we conduct some experiments to validate our framework[1] on some known sentiment datasets.

4.1 Datasets

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 [10]. 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[2], 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 [13] 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 the 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 none informative features.

4.2 Evaluation and Results

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**.

To validate our approach, we use the same experimental setting introduced in [7] 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. Results shown in Figs. 3 and 4 were produced by applying **CoSP** with its hyper-parameters: $\lambda \approx 2.11$, $\beta \approx 61.79$, $\delta = 1$ and $h = 17000$ (which stands for the number of features selected by CoSP). First, the displayed perception contains words that are meaningful in order to judge the

[1] https://github.com/KhaoulaBF/CoSPIctai.
[2] https://www.cs.jhu.edu/~mdredze/datasets/sentiment/.

Fig. 3. Top 5 features per class picked by CoSP global approach for review's binary classification on books dataset

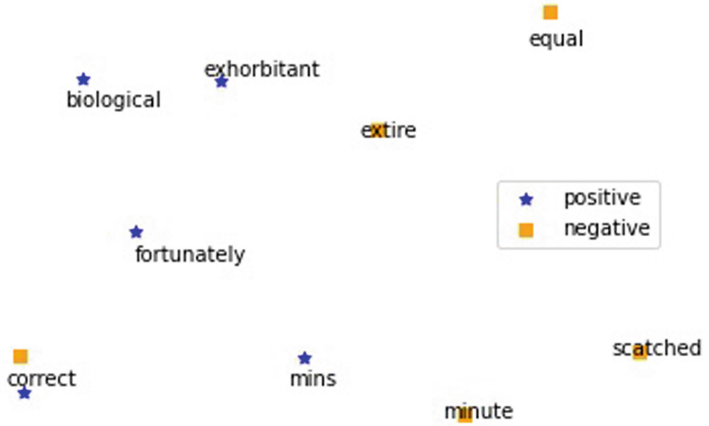

Fig. 4. Top 5 features per class picked by CoSP global approach for review's binary classification on kitchen dataset

type of comment. Features are aligned with human intuition and words with no representative meaning like stop words were not selected. Second, noisy features labeled with prefix "FAKE" added to the dataset were not deemed important.

5 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 methods in the context of time series data, as it is a challenging to find representative illustration for this type of data.

References

1. 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
2. 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
3. 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)
4. Gunning, D., Aha, D.: DARPA's explainable artificial intelligence (XAI) program. AI Mag. **40**(2), 44–58 (2019)
5. Holm, E.A., et al.: Overview: computer vision and machine learning for microstructural characterization and analysis. CoRR abs/2005.14260 (2020). https://doi.org/10.1007/s11661-020-06008-4
6. Kłosowski, 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
7. Linden, I.V.D., Haned, H., Kanoulas, E.: Global aggregations of local explanations for black box models. CoRR abs/1907.03039 (2019). https://arxiv.org/abs/1907.03039
8. Lundberg, S., et al.: Explainable AI for trees: from local explanations to global understanding. ArXiv abs/1905.04610 (2019)
9. Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)
10. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning-based text classification. ACM Comput. Surv. (CSUR) **54**, 1–40 (2021)
11. Mohaghegh, F., Murthy, J.: Machine learning and computer vision techniques to predict thermal properties of particulate composites. CoRR abs/2010.01968 (2020). https://arxiv.org/abs/2010.01968
12. 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 (2016). https://www.fatml.org/schedule/2016/presentation/why-should-i-trust-you-explaining-predictions
13. Ribeiro, M., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram, B., et al. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016, pp. 1135–1144. ACM (2016). https://doi.org/10.1145/2939672.2939778

14. 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
15. She, Y., Owen, A.B.: Outlier detection using nonconvex penalized regression. CoRR abs/1006.2592 (2010). https://arxiv.org/abs/1006.2592
16. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**(3), 647–665 (2013). https://doi.org/10.1007/s10115-013-0679-x
17. 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, 2–4 May 2013. Austin, Texas, USA, pp. 695–703. SIAM (2013)
18. 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, 28–30 April 2011, Mesa, Arizona, USA, pp. 143–153. SIAM/Omnipress (2011)