How to make the most of local explanations: effective clustering based on influences

Elodie Escriva, Julien Aligon, Jean-Baptiste Excoffier, Paul Monsarrat, Chantal Soulé-Dupuy

How to make the most of local explanations: effective clustering based on influences

Elodie Escriva[1,2][0000−0003−3618−967X], Julien Aligon[2][0000−0002−1954−8733], Jean-Baptiste Excoffier[1][0000−0002−9313−2429], Paul Monsarrat[3,4,5][0000−0002−5473−6035], and Chantal Soulé-Dupuy[2][0000−0002−2637−724X]

[1] Kaduceo, Toulouse (FR)
elodie.escriva@kaduceo.com
[2] Université de Toulouse-Capitole, IRIT, (CNRS/UMR 5505), Toulouse (FR)
[3] RESTORE Research Center, Toulouse (FR)
[4] Artificial and Natural Intelligence Toulouse Institute ANITI, Toulouse (FR)
[5] Oral Medicine Department, Toulouse (FR)

Abstract. Machine Learning is now commonly used to model complex phenomena, providing robust predictions and data exploration analysis. However, the lack of explanations for predictions leads to a black box effect which the domain called Explainability (XAI) attempts to overcome. In particular, XAI local attribution methods quantify the contribution of each attribute on each instance prediction, named influences. This type of explanation is the most precise as it focuses on each instance of the dataset and allows the detection of individual differences. Moreover, all local explanations can be aggregated to get further analysis of the underlying data. In this context, influences can be seen as new data space to understand and reveal complex data patterns. We then hypothesise that influences obtained through ML modelling are more informative than the original raw data, particularly in identifying homogeneous groups. The most efficient way to identify such groups is to consider a clustering approach. We thus compare clusters based on raw data against those based on influences (computed through several XAI local attribution methods). Our results indicate that clusters based on influences perform better than those based on raw data, even with low-accuracy models.

Keywords: Explainable Artificial Intelligence (XAI) · Instance clustering · Prediction explanation · Machine learning explanation.

1 Introduction

Analytic and predictive tools are now commonly based on Machine Learning (ML) methods and used in sensitive domains such as healthcare, finance, insurance, banking and chemistry. These methods give a prediction for a single instance based on its data, which often creates a black-box effect as methods do not inherently explain their decision process [11]. Explanation methods have therefore been perfected, providing global insights about the model's general

behaviour or a local one about a single situation [9] (XAI for eXplainable Artificial Intelligence). Local explanations are increasingly used in AI-assisted tools to offer more information than a single prediction [1]. Among the most popular methods are XAI local attribution methods that produce influences, especially LIME [13] and approximation of Shapley Value such as SHAP [12], the K-depth [7] and Coalitional approaches [8]. Their popularity is due to the instance-level accuracy of these explanations, which links the impact of each attribute to the prediction made for each instance and allows finer differences to be detected between all instances. Yet, providing only local influences seems insufficient to improve decision-making efficiency. Indeed [16,17] show that displaying influences along with an individual prediction did not significantly enhance the utility and understanding for the user as opposed to prediction alone. Moreover, knowing all the local explanations of a dataset does not guarantee a complete data understanding since there are as many explanations as instances in the original raw dataset, with the difficulty of finding explainability patterns in this new dataset.

In this context, we hypothesise that influences can be seen as a new data space that can be explored and used as a basis for further analysis. Indeed, influences provide new information thanks to ML modelling, which considers complex phenomena and interactions. Influence analysis can thus help identify the main trends of explanations, i.e. the characteristic relationships between the attributes. Also, it can be interesting to provide a global view of the explanations to determine whether instances are typical or atypical cases of the data. In this direction, an influence-based clustering approach is a good candidate since it can be the most straightforward approach to detect more homogeneous subgroups of influences and understand the behaviour of the modelling and the underlying dataset. Thus, in this paper, we want to propose a framework for analysing influences through a clustering approach. To the best of our knowledge, this is the first work that studies in a general framework the benefits of using local influences as a new input for clustering to identify more informative and homogeneous groups. We also explored the robustness of this framework regarding low-accuracy models or misclassified instances.

The paper is organised as follows. Section 2 gives an overview of the current local explanation methods used in experiments and how explanations are used to detect subgroups of instances. Then, section 3 details our clustering framework for detecting subgroups based on local influences. Section 4 describes the experiments performed on 104 datasets to compare the use of raw data and influences from multiple local XAI methods. We study the K-medoid clusters quality to show the efficiency of using influences. We detail the metrics used to evaluate the clusters and the different approaches based on the model prediction. Globally, our results demonstrate that local influences produce better-quality clusters than raw data, even with low-accuracy models. Separating instances well classified and misclassified by the model also allows a more precise clustering. Section 5 discusses the advantages of our approach in a broader context, linking results from clustering with knowledge from modelling and explanation methods. Section 6 concludes this paper and gives short and long-term perspectives of works.

2 Related works

In the field of local explanations, one of the first methods was based on the Shapley values, a local attribution XAI method [18], to explain machine learning predictions. With these methods, the influence of each attribute over a prediction is computed as the difference in prediction from the model with and without the attribute. Influences then represent the impact of each attribute over a prediction for each instance of the dataset. Local influences facilitate the prediction understanding without expert data science knowledge as they are easy to interpret and represent graphically. Other methods have emerged with LIME [13] that uses linear surrogate models trained with sampled data to approximate the black box model locally. The Coalitional approaches [8] approximate the Shapley value by precomputing relevant groups of instances and reducing complexity. Finally, SHAP [12] mixes Shapley values with LIME and other methods to simulate the absence of attributes by sampling, find a linear model that explains the black-box model locally and approximate the Shapley values. Nowadays, SHAP is one of the most well-known methods in the literature, easy to use and provides both agnostic and specific methods with KernelSHAP or TreeSHAP.

With the rise of explainability, ML research looks beyond simply explaining the machine learning model. Several papers in the last year have covered use cases combining machine learning explainability and clustering to find relationships between instances [3,10]. Based on a COVID-19 dataset, [3] tries to identify better clusters based on KernelSHAP values. Rather than clustering the original dataset, called raw data, they trained a classification model, computed the KernelSHAP values for each instance and performed DB-SCAN clustering on these influences. They show better identification of clusters with influences than raw data and graphically display the cluster differences using UMAP, a well-known reduction dimension technique. Other papers also used clustering to determine groups and to recommend instances based on the influences on a single dataset [5,6]. [5] was a use case on a urinary disease that explores healthcare risk stratification based on influences from TreeSHAP. Clustering patients by SHAP values allows the selection of representative patients and investigation of the risk factors for each cluster, where only raw data are insufficient to perform the same analysis. The same kind of analysis was performed on a COVID-19 dataset concerning the identification of subgroups of patients during the first lockdown in France [6]. These four papers explored the idea of using influences and clustering to find more knowledge about the data on specific medical examples. However, no paper formally evaluates the contribution of explanation clustering in general. Although their positive conclusions, these papers only use one single dataset with one XAI method, without generalizing the approach or comparing findings with other XAI local attribution methods, in opposition to what we propose in this paper. Finally, none uses prediction to differentiate subgroups of data for clustering.

3 Influence-based clustering framework

In this section, we detail our influence-based clustering framework. Figure 1 shows the step-by-step process to cluster instances based on their influences:

1. A machine learning model is trained with raw data and predicts classes of all the instances from the raw dataset.
2. A local attribution XAI method explains the trained model. Users can choose the data used as input for the method. Influences are computed to explain why the ML model made such predictions.
3. A clustering algorithm is used on influences to create homogeneous groups of instances to detect their important attributes based on the modelling. Users can define the number of clusters they want to compute.

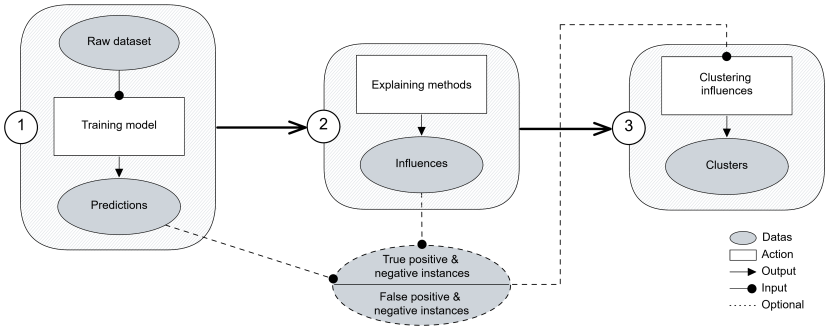


Fig. 1: Our proposal Framework.

In this framework, various elements can be modified according to user preferences. Any classification model can be used in Step 1, as they are all designed to compute predictions, and Step 3 allows any clustering method.

In step 2, the framework is designed to accept XAI local attribution methods. These influences are represented as tabular data, where each instance has a value associated with each attribute. We directly use these influences data as input for the clustering step. Influences are valuable because they provide additional information that the raw data does not: the link between the modelling predictions and the dataset attributes. Compared to raw data, explanations produced by XAI local attribution methods have the same unit across all attributes, thus avoiding any problem of value ranges. Another advantage is that influence values are less noisy since the ML model mainly focuses on attributes relevant to the underlying predictive task and excludes information not explained by the complex attributes interaction, hence the relevance of carrying out clustering. For supervised tasks, XAI local attribution methods usually generate a dataset for each class with identical dimensions as the raw data. For example, if the raw data consists of n instances and m features and the supervised task is a

multi-class problem with c classes, the generated dataset (also called the influence dataset) has a $n \times m \times c$ dimension. To have an influence dataset with the same dimension as raw data ($n \times m$) one can only select a single class and its associated influences. For example, regarding binary classification, the positive class is often chosen as the class of interest for influences. Finally, XAI local attribution methods allow the selection of different data as input than the ones used for training the machine learning model. These inputs are used by XAI methods to explain the model, by creating perturbations in SHAP or LIME, for example.

An additional and optional step is to select a particular subset of the data for clustering. Indeed, it is possible to study the instances correctly and incorrectly classified by the model separately via instance clustering. Considering the model predictions against the data labels, the influences are separated into two distinct groups before being clustered. Two different sets of clusters are then proposed to the users. This step can have several advantages. Since the influences represent the model decisions, separating the instances can provide new knowledge. Studying the well-classified instances can help to identify their characteristic patterns by removing noise and outliers from the misclassified instances. This can give a more accurate idea of general patterns, for example, to check that there is no bias in the dataset. Regarding misclassified instances, they may have several representations. They can be outliers in the data and not correspond to the general behaviours without bias or error. However, misclassified instances can also be a particular subgroup of the data relevant to study. For example, this would be the case of children with cancers usually associated with older people. Due to age, the model may misunderstand this subgroup, as there are few children with non-pediatric cancers, or the input variables may be insufficient to identify this subgroup. However, it is necessary to study this subgroup to understand whether there is any specific behaviour in this subgroup and ultimately understand the overall dataset. Separating the instances can therefore allow the exploration of new patterns that can be invisible if all the data were kept. This may be even more important for influences because of their direct link to the model. Indeed, when the model prediction is incorrect, the influences reflect this error and are directly impacted by the wrong prediction of the model.

The full implementation of our proposal is available here: https://github.com/kaduceo/XAI-based-instance-selection. The source code will evolve with future works. Additional materials are also available.

4 Experiments

4.1 Experimental protocol

For our evaluations, we use 104 datasets from an Open ML collection[6] [14] that meet the following criteria: binary classification, more than 100 instances, more than four attributes and at most nine attributes due to the computational cost of producing influences. Table 1 details statistics about the datasets used.

[6] Available in `https://www.openml.org/s/107/tasks`

Table 1: Statistics of the experimental datasets based on the number of attributes.

# of attributes	4	5	6	7	8	9
# of datasets	14	25	17	16	15	17
Mean # of insts	465	1197	654	554	650	503
Min-Max # of insts	125-1372	100-7129	100-3107	108-4052	130-4177	100-1473

Binary classification is chosen to facilitate the interpretation of influences. We consider that all influences are based on class 1. In this case, influences represent the impact of each attribute on the probability of the instance being in class 1. We train a Random Forest model (RF) with a Grid Search Cross-Validation to optimise hyperparameters. This model was chosen to test tree-specific explanation methods while keeping a limited number of hyperparameters to avoid overfitting (compared to boosted trees). Only to evaluate the performances of the modelling, each dataset is divided into train and test sets according to the 75%/25% ratio. Table 2 shows the performances of all the models trained in our experiments. Models are trained adequately to capture most information of the dataset. The mean and median balanced accuracy are respectively 0.79 and 0.85, meaning most models can accurately classify test instances. Some models also have very low accuracy, the minimum being 0.42. We choose to separate models based on a threshold set to 0.8 to evaluate the behaviour of our framework on models with high and low accuracy. Thus, high-accuracy models have a median balanced accuracy of 0.92, whereas low-accuracy models have a median of 0.6.

We also study the number of instances well classified and misclassified by the ML modelling in Table 2. In all experiments, we call *true instances* well-classified instances, referring to True positive and True negative terms. *False instances* is then related to False positive and False negatives instances, so misclassified instances. We use three different separations of data: all instances together, only true instances and only false instances. As we separate true and false instances, we choose not to evaluate high-accuracy models on false instances as there are not enough instances in most datasets to create clusters and properly evaluate them and compare the results. Then, when studying false instances, we only work with models with low accuracy as the number of false instances is higher and sufficient. Also, the number of true instances is adequate to perform clustering for all models.

For exhaustive purposes, we choose three different XAI local attribution methods to compute influences: KernelSHAP, TreeSHAP, LIME and Spearman coalitional. As explained in [4], each XAI method provides influences with different strengths and disadvantages. Thus, we want to study the relevance of using local influence clustering compared to raw clustering in a global way.

Once influences are computed, instances are clustered by the influence-based approach with K-medoids as the clustering method. This method has the advantage of always selecting actual instances as centroid from the dataset, unlike

Table 2: **Statistics of models trained**. Balanced accuracy and percentages of true and false instances are presented for the 104 datasets, and separataly based on the 0.8 accuracy treshold. For true and falses instances, the median number of instances is presented along the percentage.

Models (#)	Balanced Accuracy			% of True instances			% of False instances		
	Median	Min	Max	Median	Min	Max	Median	Min	Max
All (104)	0.85	0.42	1.0	94% (307)	61%	100%	6% (21)	0%	39%
Acc >0.8 (60)	0.92	0.81	1.0	97% (404)	85%	100%	3% (11)	0%	15%
Acc <0.8 (44)	0.60	0.42	0.79	82% (252)	61%	98%	18% (62)	2%	39%

other clustering methods like *k-means* where centres are not necessarily existing instances. Metrics to compute distances, so clusters, can be selected arbitrarily, with the Euclidean distance being the usually chosen distance. As both raw data and influences data are tabular data of the same dimensions, the distance metrics can be easily applied to both datasets to compute clusters without adapting the clustering method to a specific input. We use ten different percentages to choose the number of clusters: 1%, 2%, 3%, 4%, 5%, 10%, 20%, 30%, 40% and 50%. We define the number of clusters based on the percentage as $n_{cluster} = p * n_{instances}$ with p the selection percentage between 0 and 1. We set a minimum number of two to avoid too few clusters. As the size of the datasets varies greatly as shown in Table 1, we prefer to select a percentage rather than fixed numbers of instances to take into account the diversity of the datasets. As we aim to show how clustering on influences exhaustively performs against the raw data, multiple percentages per dataset can show how cluster quality evolves without looking for the optimal number of clusters (which may also be different for each method).

Finally, we evaluate if clusters are well defined and manage to group similar instances and separate dissimilar instances based on their *a-priori* labels. We select two external clustering metrics, *Entropy* and *Purity*. With external metrics, class labels are needed as metrics assess the distribution of labels within clusters to evaluate how clusters and labels are related and how clusters manage to group similar instances. Entropy measures the distribution of labels in a cluster, i.e. the ability of the algorithm to differentiate between data that do not have the same "real" class. A perfect entropy means all instances from the same class are in the same clusters. In addition, Purity measures the relative size of the majority class in a cluster to evaluate its dominance over other classes. Perfect purity describes that each cluster has only one class. These two metrics give values between 0 and 1. A perfect clustering will usually have an entropy equal to 0 and a purity equal to 1. These metrics are defined as follows [2]:

$$Entropy = \sum_{k=1}^{K} \frac{n_k}{n} \Big(- \frac{1}{\log q} \sum_{i=1}^{q} \frac{n_k^i}{n_k} \log \frac{n_k^i}{n_k} \Big) \qquad Purity = \sum_{k=1}^{K} \frac{1}{n} \max_i (n_k^i)$$

where C_k is a particular cluster of size n_k, q is the number of class in the dataset, K the number of clusters and n_k^i is the number of instances of the ith class assigned to the kth cluster.

4.2 Results

In this section, we describe the results of the experiments by first comparing the clusters based on the influences from XAI methods with the ones made with raw instances. We then study the impact of each data subgroup on the cluster quality for KernelSHAP and Spearman coalitional. For all experiments, results are presented separately based on the machine learning model's accuracy to differentiate the impact of the model performance on the influences and clustering.

Comparing Raw data and XAI influences clustering When comparing raw data clusters to the influence ones, for all instances, Figure 2 shows raw data clusters have lower purity and greater entropy than other clusters, regardless of the percentages, the XAI methods or the model performance. Differences in entropy between clusters from raw data and influences are even greater when the model has an accuracy greater than 80%. Clusters from raw data have poorer quality than clusters from influences, indicating that clustering instances based on their influences from XAI methods gives better results than clustering the raw data. Also, as expected, when models have a lower accuracy, clusters have a lower purity and entropy, whatever the data or cluster percentages. Indeed, when the model's performance is poor while the model is adequately trained, this may indicate that the data is less generalisable or of lower quality. This hypothesis seems to be reflected in the quality of the clusters created.

When taking into account only the true instances (the instances well predicted by the model), Figure 3a shows similar results as Figure 2: clusters based on influences have better quality than the ones based on raw data (for all XAI methods, percentages of selection and model accuracy). The purity and entropy are almost perfect, even with low selection percentages. Clusters have also better quality with only the true instances than with all the instances of the dataset. Figure 3b only considers the false instances (instances misclassified by the model). Globally, the cluster quality is degraded, especially the purity. Since purity checks the proportion of the majority class in each cluster, grouping instances misclassified by the model logically lower the cluster purity. No XAI method seems to have a good result on small percentages, even if they all have better results than raw clustering. With false instances, we analyse cases where the model fails to generalise or describe the data correctly. As influences represent the model decision, influences of misclassified instances may have lower quality than true instances. They may, however, be representative of why the model does not generalise and understand these data. Thus, these clusters can indicate where the problems lie in the data or the model.

Moreover, even if this is not the aim of this paper, we can briefly compare XAI methods between them based on cluster quality. Although all are attribution

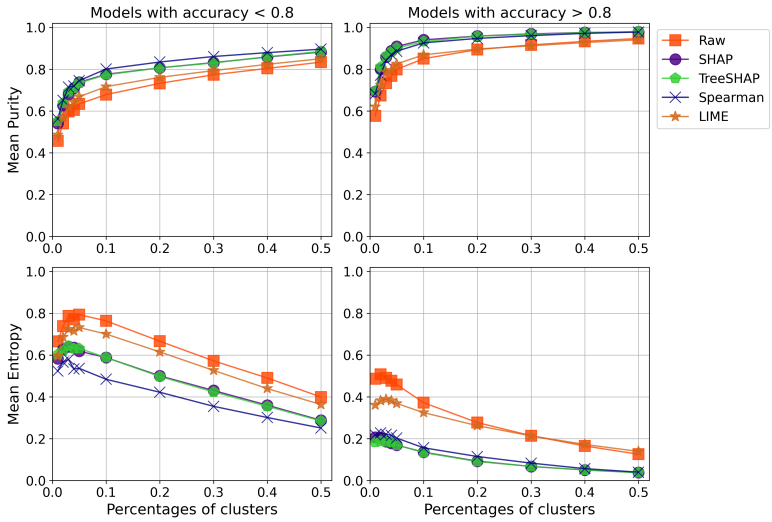


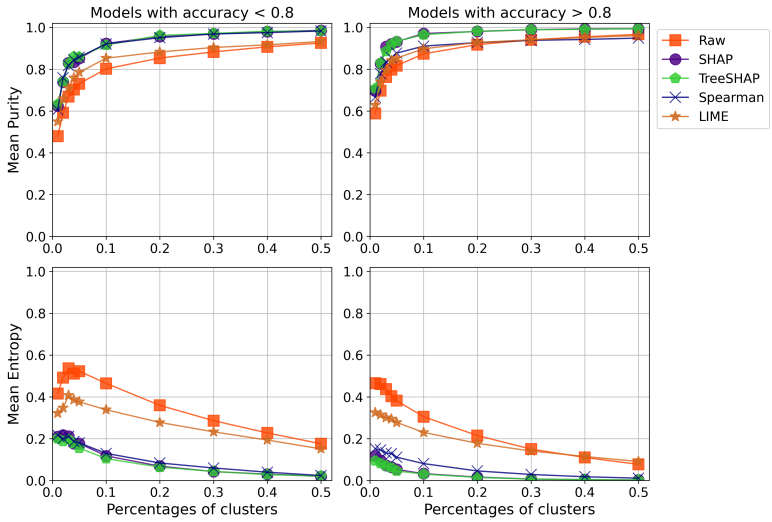
Fig. 2: Comparison of clustering for XAI methods trained on all instances.

methods with similar global behaviour, the calculated influences appear to be sufficiently different to produce dissimilar cluster results, especially in entropy. Spearman coalitional seems slightly better at clustering false instances on models with low accuracy. Clusters based on LIME have purity and entropy close to the clusters based on raw data, making this XAI method the one with the worst results. Based on the subgroup of data studied, one method may be preferable to another depending on the context. This seems consistent with the findings of [4], where depending on the dataset, the interdependence of attributes, the dimensionality or the model, one XAI method can be more efficient than others. The same reasoning seems to apply here, where according to the subgroup studied, one XAI method can be better than the others.

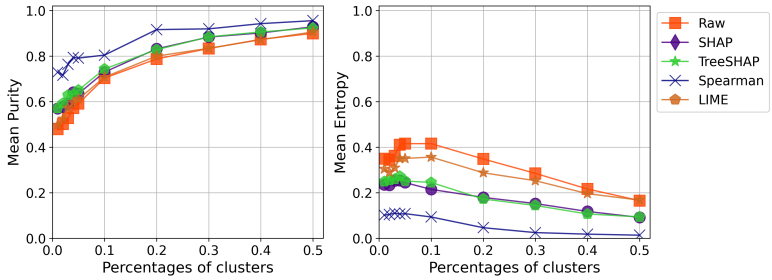
In the next sub-section, we select only KernelSHAP and Spearman coalitional to study the impact of using only specific subgroups of data, as TreeSHAP is almost identical as KernelSHAP and LIME have the worst results.

Comparing the impact of using different data subgroups In this subsection, we aim to show in which circumstances well classified or misclassified instances can be used to produce clusters of good quality (or not), notably in the worst case (degraded accuracy on a set of misclassified instances). Figure 4 and 5 show the cluster quality for the three data modalities, with influences respectively from KernelSHAP and Spearman coalitional.

Figure 4 shows little difference in cluster quality between *all instances* and *true instances* subgroups for models with high accuracy. Purity metric is high and almost equal for both modalities, and the *all instances* subgroups have slightly higher entropy. Influences from true instances produce almost perfect clusters



(a) "True" instances.



(b) "False" instances.

Fig. 3: Comparison of clustering for XAI methods trained on (a) only "true" instances and (b) only "false" instances for models with an accuracy below 0.8.

even with low cluster percentages and are little affected by the model accuracy. As models with high accuracy have fewer false instances, their influences may only produce noises for the clustering. Removing them give slightly better global results, as clusters have better entropy. For models with low accuracy, there are more differences between the subgroups, presumably because the proportion of false instances is greater. The *all instances* and *true instances* subgroups have a 0.4 difference in entropy and a 0.1 difference in purity for almost all percentages. The *false instances* subgroups also have similar purity and better entropy as the *all instances* subgroups. Separating true and false instances to study them separately produces more homogeneous and coherent clusters than keeping all instances together, especially on low-accuracy models. With these models, the number of false instances is higher, and they often represent behaviours not caught by the model.

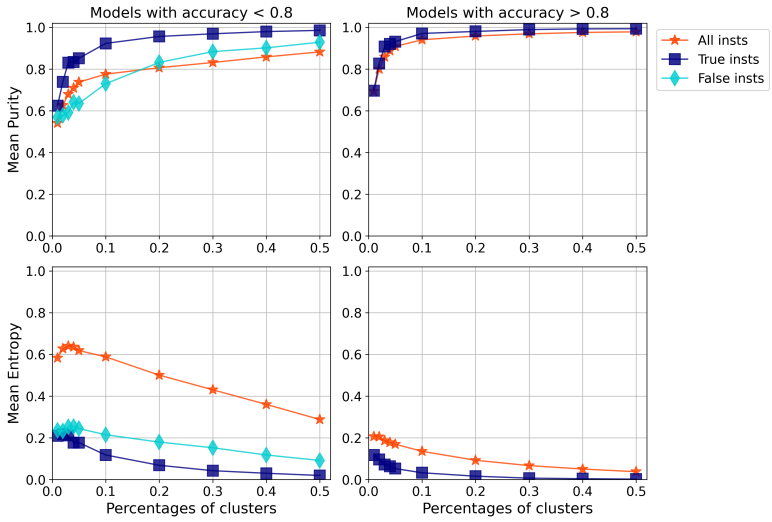


Fig. 4: Comparison of clustering of KernelSHAP influences.

For Spearman coalitional method, Figure 5 reveals a similar overall behaviour to KernelSHAP regarding the cluster quality depending on the subgroups, especially on high-accuracy models and on the *true instances* subgroups. However, for low accuracy models and unlike KernelSHAP, there are some differences when using only false instances. The *false instances* subgroups have slightly higher purity and lower entropy, especially on low percentages. The different use of input data by both methods can explain this behaviour. KernelSHAP use the input to produce perturbations for the model, creating new instances and studying a larger area of the data space than just the input data (here, the false instances). In contrast, Spearman coalitional does not produce any perturbations and uses the input data as is to explain the model. The data space is then smaller, therefore, less exhaustive. Using only false instances may lead to influences more precise for this subgroup, compared to using all instances or instances with perturbations, hence the difference between the two subgroups for Spearman coalitional and the difference with KernelSHAP. Moreover, for low-accuracy models, clusters from *true instances* and *false instances* subgroups are better than the clusters from all instances.

These two figures show that different XAI methods can lead to clusters with distinct qualities or behaviour based on the data subgroups selected. These methods can produce diverse and meaningful clusters to understand the modelling and dataset.

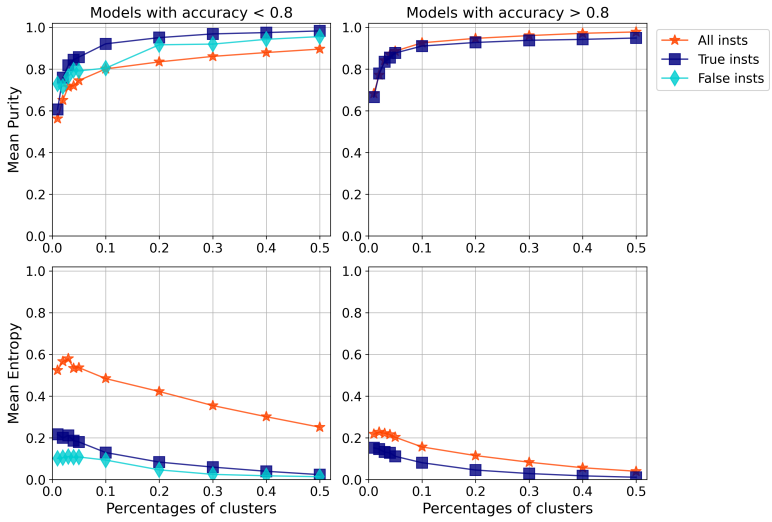


Fig. 5: Comparison of clustering of Spearman coalitional influences.

5 Discussion

Clustering on XAI influences showed better results than clustering on raw data, regardless of the percentage of clusters, the XAI method or the performance of the modelling. The influences seem to contain information allowing a better clustering, probably by highlighting the most significant attributes for each instance or removing noises from raw data. This finding seems consistent with the results of [3] while showing a more global approach, working with other XAI methods than KernelSHAP and a hundred of datasets.

Separating the instances correctly and incorrectly classified by the model also seems to give better results than keeping all the instances together. Since the information in the two subgroups is different, they each seem to create noise in the information of the other subgroup. Indeed, the misclassified instances are often outliers or critical instances in the dataset. Their behaviour is different from the general behaviour of the data, whereas correctly classified instances follow the behaviour that the model detects. However, as some misclassification may result from bias in a subgroup of the data or from the atypical behaviour of that subgroup compared to the whole dataset, it is of great interest to study them as a priority. When separating correctly and incorrectly classified instances, the differences in cluster quality seem to be more pronounced with the Spearman coalitional method than with KernelSHAP. The contribution seems to depend on the XAI method used, probably depending on the XAI method for influences since KernelSHAP creates perturbations on the instances and Spearman coalitional keeps the input data as it is. A limit to these subgroups' separation is also the decrease of its relevance when the accuracy of the model increases. Indeed, the number of false instances logically decreases with increasing accuracy.

Creating an XAI model and clusters with a low instance count does not make sense and can only lead to data misunderstanding. However, as the accuracy increases, the false instances become mostly outliers of the dataset or biased instances rather than subgroups with their behaviours to analyse. Their small number can be analysed manually without any particular clustering method.

Finally, the proposed approach also adds another use of influences. Clusters based on influences can be used to focus on sub-groups of data to be studied. Clustering can be combined with other approaches to understand the clusters created, like rule-based algorithms or instance selection. It reinforces the idea that influences can be considered as new inputs for finer analysis, either directly in the ML pipeline (feature selection [15]) or, afterwards, to gain a more in-depth and concise understanding of the ML model and the underlying data. Examples of how explanation clusters can be used are available on the GitHub mentioned above.

6 Conclusion and perspectives

We propose in this paper a general framework for clustering instances based on influences and predictions. We combine XAI local attribution methods with clustering to explore the space of influence data. We provide clusters of similar instances to assist in analysing modelling and dataset. Experiences validate the valuable contribution of influence-based clustering. The clusters from the influence-based framework are more homogeneous and of better quality, providing insight into the modelling. We also prove that the clusters formed are of good quality and pertinent, even for low-performance models. We also show the advantages of splitting the well- and misclassified instances by the model when studying a dataset as a whole, as it highlights the most important subgroups of data and the behaviour of outliers simultaneously.

Perspectives will first be focused on extending our approach for other supervised tasks. Clusters can also help select informative instances and provide a small number of instances to users. These instances can help to understand datasets and modelling using examples rather than statistical information. With users in the loop, the framework with instance selection added could be tested against example-based XAI methods. New information on the dataset and its subgroups may also provide feedback on the quality of the training data or the trained model to improve it. This idea of possible user feedback may be one way to improve data quality and modelling. Clustering based on influences may help to understand *why* the model is wrong and not just *where* the model is wrong, and allow for detecting bias in the data. Perspectives will then be focused on evaluating this feedback and how it can be implemented in the framework. A long-term perspective is to use the framework in a complete system where users can interact with the modelling and define typical instances to profile new data patterns for user testing.

Acknowledgement We thank the French ANRT and Kaduceo company for providing us with PhD grants (no. 2020/0964).

References

1. Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A., Mooney, C.: Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. Applied Sciences 11 (2021)
2. Conrad, J.G., Al-Kofahi, K., Zhao, Y., Karypis, G.: Effective document clustering for large heterogeneous law firm collections. In: AIL Proceedings (2005)
3. Cooper, A., Doyle, O., Bourke, A.: Supervised clustering for subgroup discovery: An application to covid-19 symptomatology. In: ECML-PKDD Proceedings (2021)
4. Doumard, E., Aligon, J., Escriva, E., Excoffier, J., Monsarrat, P., Soulé-Dupuy, C.: A comparative study of additive local explanation methods based on feature influences. In: DOLAP Proceedings (2022)
5. Excoffier, J.B., Escriva, E., Aligon, J., Ortala, M.: Local Explanation-Based Method for Healthcare Risk Stratification. In: Medical Informatics Europe 2022. Studies in Health Technology and Informatics (2022)
6. Excoffier, J.B., Salaün-Penquer, N., Ortala, M., Raphaël-Rousseau, M., Chouaid, C., Jung, C.: Analysis of covid-19 in patients in france during first lockdown of 2020 using explainability methods. Medical & Biological Engineering & Computing 60 (2022)
7. Ferrettini, G., Aligon, J., Soulé-Dupuy, C.: Improving on coalitional prediction explanation. In: ADBIS Proceedings (2020)
8. Ferrettini, G., Escriva, E., Aligon, J., Excoffier, J.B., Soulé-Dupuy, C.: Coalitional strategies for efficient individual prediction explanation. ISF, Springer (2021)
9. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J.: Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In: CHI Proceedings (2020)
10. Lee, K., Ayyasamy, M.V., Ji, Y., Balachandran, P.V.: A comparison of explainable artificial intelligence methods in the phase classification of multi-principal element alloys. Scientific Reports 12 (2022)
11. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16 (2018)
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NeurIPS Proceedings (2017)
13. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: KDD Proceedings (2016)
14. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: Openml: Networked science in machine learning. SIGKDD Explorations 15 (2013)
15. Wang, H., Doumard, E., Soulé-Dupuy, C., Kémoun, P., Aligon, J., Monsarrat, P.: Explanations as a new metric for feature selection: a systematic approach. IEEE Journal of Biomedical and Health Informatics (2023)
16. Weerts, H.J., van Ipenburg, W., Pechenizkiy, M.: A human-grounded evaluation of shap for alert processing. arXiv preprint arXiv:1907.03324 (2019)
17. Zhang, Y., Liao, Q.V., Bellamy, R.K.: Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: FAccT Proceedings (2020)
18. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41 (2014)