Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25–29 May 2026, Vancouver, BC

Enhancing Stability in Rule-Based Post-Hoc Explanations

Abstract

Many AI models being explained today are pre-trained and static, so the reasoning for a particular decision does not change across predictions. To trust an explanation, it must also stay the same – or at least be similar – across multiple runs. This is a problem when many users need to be able to trust the explanations, and an issue which we attribute to poor data being used to train surrogate-based post-hoc explainer. In this work, we explore powerful explainers which use rules, where explanation instability stemming from training data becomes more apparent. Inconsistent rules are easier to discover and can lead users to trusting these explanations less. We theorize why multivariate normal distributions, producing correlated perturbed training data (+P) improve stability and fidelity. We also balance classes in the training data (+B) to further improve stability, along with exploring the use of clustering (+C) for locality improvements to explanations. By providing both theoretical reasoning for the improvements and experiments on seven diverse datasets, with two different black-box architectures we found that the rule-based method we employed, BARBE, sharply increased in fidelity and stability when trained with our modified process, making BARBE+PBC which exceeded other methods that improve stability like S-LIME and LORE. Explanations from BARBE+PBC provides clear, trustworthy, and most importantly, stable rules. Stable rules have lower risk of betraying the previous insights a user relied on for an explanation. The final code is available as a package on GitHub at https://anonymous.4open.science/r/Stable-BARBE-014F/.

Keywords: Explainable AI, post-hoc explanations, stability, associative classifiers

1. Introduction

In recent years, eXplainable AI (XAI) has been prioritized to address critical problems with black-box machine learning algorithms. An explanation for a black-box generally outlines the reason a specific decision, or many decisions were made. We can use an example of a bank's loan approval process. If a bank rejects a loan then the client may want to know why their application was rejected. Many explanation methods focus on single samples to explain because a local reason may be easier to understand and serve the client better. For example, a local explanation may be that the bank rejected this loan since the user has a low credit score.

Explanations for black-box decisions can be important to users and are even becoming a regulatory requirement in Canada [1]. So, proactive organizations are using post-hoc explanations of their pre-trained black-box models as a critical tool to fulfill these requirements. If such an explanation is untrue or inconsistent in a critical area like medicine, then it risks harm to a patient. While an untrue explanation causes obvious harm – like pursuing and incorrect treatment plan – an inconsistent explanation harms trust, a difficult to measure metric which is still of upmost importance to explanations.

The first measure a user will look use for an explanation is the feature importance, if it is provided. Feature importance usually consists of ranked values that give an estimate of how important a feature is to the prediction, either positively or negatively. It was popularized by Local Interpretable Model-agnostic Explanations (LIME) [2], and it is clear that this type of explanation is welcoming to users. Since these are easy to use, issues with stability are also easier to spot. If two explanations for the same sample switch from one important feature to another, then a simple conclusion which a user may make is that the explainer is incorrect. Returning to the loan example, in one case the explainer ignores the user's credit score entirely and bases the prediction on the amount requested for the loan. While the

This article is © 2026 by author(s) as listed above. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (https://creativecommons.org/licenses/by/4.0/legalcode), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the author(s) identified above.

inverse happens in the second prediction. This change happens because the explanation is not stable, and it impacts user trust.

Inconsistent explanations could be caused by a link between two features that are not always assigned to importance. Since the surrogate model of a post-hoc explanation determines feature importance it could be most at fault for stability issues. However, missing and outlying data may also be a source of instability, and due to a lack of severe evaluation in XAI most methods do not expose this source of instability and test against it [3]. In this work, we aim to improve the stability of some post-hoc explanation methods and heighten the severity of evaluations to address this significant gap.

Rule-based explanations uniquely offer a clear, causal explanation for a sample's prediction. Due to this, the source of the explanation, the rules, can be evaluated to establish conditions that improve stability. We focus on Black-box Association-Rule Based Explanations (BARBE) [4] which uses multiple rules to explain a point-of-interest. Previous work shows that BARBE and other explainers can benefit from more complex data, taking feature correlation into account [5]. Our approach targets the most simple changes which can be made to improve stability further. This is done to avoid complicating the process of explaining a sample and to save time and resources.

In this work we propose BARBE+PBC, a modification to the BARBE explainer to improve stability. We evaluate improvements theoretically and identify multivariate perturbations (+P) as an existing – but unproven – improvements to stability, balancing classes in the perturbations (+B), and clustering training data (+C). Unlike other approaches we also use severe statistical testing and identify multiple sources of instability in the training data which tend to be left constant and lead to unclear results. The contributions of this work are as follows:

- C1: Showed theoretically that multivariate perturbations (+P) and class balancing (+B) improve the stability of rule-based explanation methods.
- C2: Showed that the error added by clustering (+C), when there are 2 groups, does not outweigh prior stability improvements, making it a viable option for further improving the "locality" of an explanation.
- C3: Provided strict and severe significance-based tests on seven datasets to show that the theory holds in application, using the Jaccard index and Relative Input Stability for evaluation.
- C4: Provided these modifications in an easy-to-access GitHub repository (`https:\anonymous.4open.science/r/Stable-BARBE-014F/`) implemented in the Python programming language.

We present the background related to post-hoc explainability in Section 2. In Section 3, we discuss BARBE – the method we advocate for. In Section 4, we formally describe the stability problem for post-hoc explanations along with some theory that is useful for improving BARBE. We apply the theory in Section 5, splitting up between the known multivariate perturbations and our new improvements, class balancing and clustering. Section 6, describes the experiments we use to show these results on real data and we present the results and conclusions in Sections 7 and 8.

2. Background

Post-hoc explanations deal with pre-trained "black-box" model, a model that does not provide an explanation and is difficult or improbable to sufficiently interpret. Many post-hoc explanation approaches exist [2, 6, 7] and a particular area of focus for this work is methods that use an interpretable surrogate, or "white-box," model. A white-box is explainable or interpretable and it is trained to give similar predictions to the black box. Many of these approaches create a "local" explanation for a point-of-interest by training the white-box on

perturbations of the point-of-interest, like LIME [2], which also provides an explanation in the form of feature importance values. LORE [7] improves explanations by instead providing a supporting rule from a trained decision tree. These are single rules, which may be limited as an explanation since the understanding of a complex black-box is reduced to one branch of a decision tree.

The stability of a post-hoc explanation is usually quantified by how consistent the explanation is, this can be done using the Jaccard index [7, 8], which prioritizes the relative order of importance values, or by comparing the importance of a single feature across runs using the Relative Input Stability (RIS) [9]. The stability of popular methods like LIME has been contentious, with many approaches aiming to solve the problem [8, 10].

Approaches like S-LIME [8] do not correct for other important metrics so the amount S-LIME can be trusted is in question. Along with their improvements requiring many perturbed points, which may not be ideal given the white-box being used. VAE-LIME [10] and in-turn VAE-SLIME [11] help by providing an alternative, producing better perturbations using a variational auto-encoder (VAE). Since VAEs are complex some additional, confounding tasks fall onto the user. Selecting layer sizes and hyper-parameters require expert users and many trials that may not be possible for lay-users. Better approaches should attempt to simplify what the user needs to provide or limit reliance on training data.

3. BARBE

Black-box Association-Rule Based Explanations (BARBE) [4] is a post-hoc explanation approach that uses an associative classifier to explain decisions made by a black-box. An associative classifier uses association rules R of the format $R : X \rightarrow Y = y$ for a set of feature-value pairs X predicting a class y. These rules can also be put into a human readable format, for example: "$\$5k \leq income < \$10k$ (X) implies $(\rightarrow)$ rejected loan $(Y = y)$". The associative classifier is trained using perturbations of the point-of-interest x_p, $x \sim \mathcal{N}(x_p, \Sigma)$ on a multivariate normal distribution with covariance Σ, either provided by a user or estimated using training data. The most significant rules for predicting the black-box's labeling $f(x) = y$ are found by using the Fisher's exact score [12]:

$$p(R) = \sum_{i=0}^{min(m_c, m_X)} \frac{\binom{\sigma(X)}{\sigma(X, Y = y) + i} \binom{\sigma(\neg X)}{\sigma(\neg X, \neg(Y = y)) + i}}{\binom{|D|}{\sigma(Y = y)}}$$

Where $\sigma(X)$ is the number of cases in training which have the combination of feature-value pairs in X, $\sigma(X, Y = y)$ are the number of those cases where the class is also y, and $|D|$ is the size of the dataset.

The $-log(p(R))$ is then used to estimate the feature importance. Where for each rule which applies to x_p with a certain number of input conditions $|R|$, the feature importance is $imp(A) = \sum_{A \in R} (-1)^{\mathbf{1}(R \rightarrow Y = y)} \frac{log(p(R))}{|R|}$. So, the feature importance is based on the cumulative significance of the rules that feature A appears in, scaled by the number of conditions which must be satisfied, with a sign based on the class being predicted.

To our knowledge, no stability improving work exists for the BARBE explainer. In addressing this issue we aim to provide better, theory-based reasoning for the improvement, and severe testing, which is a continual issue with XAI [3].

4. Theoretical Improvements to Stability

Earlier, in Section 2 we discussed that stability metrics are usually based on the amount that feature importance changes between runs, so we should focus on methods for improving

the bounds of the feature importance, $imp(A)$, to yield continual improvements to stability. First, we consider the distribution of the importance as a continuous value sampled based on the training data.

$$imp(A) = E[imp(A)] + (L + L_\epsilon)\epsilon \leq E[imp(A)] + \frac{d(d+1)}{2}\epsilon \tag{4.1}$$

Where $(L_\epsilon)_{ij} \sim \mathcal{N}(0,1)$ are errors created when estimating the covariance matrix from a finite dataset, which is decomposed into the parts used for perturbing data $\Sigma = L^T L$. While $\epsilon \sim \mathcal{N}(0,1)$ is similarly distributed and is the usual source of noise from simulations. The value d is the number of noisy parameters in the training data. We simplify to $d(d+1)/2$ since the same error is reflected over the diagonal of Σ. Now, suppose that we found a set of important rules a target feature A which satisfies the following inequality with a set of comparison rules for another feature B.

Assumption 1. *The importance value of a feature is stable if the difference between the low bound ϵ_L of a target feature A and the high bound ϵ_H of an alternative feature B is greater than the noise from the sampled distribution of importance values, $(L + L_\epsilon)\epsilon \leq \frac{d(d+1)}{2}\epsilon$.*

$$p((imp(A) - \epsilon_L) - (imp(B) + \epsilon_H) \geq \frac{d(d+1)}{2}\epsilon) \geq \delta \tag{4.2}$$

If the above criterion holds, then we can consider the result to be stable to the right-hand-side (RHS) error $d(d+1)\epsilon/2$ with a significance δ. Higher distance between importance values means that the feature importances are more likely to remain greater or lesser relative to each other. So if we are able to produce tighter bounds, ϵ_L and ϵ_H, on the estimated importance values, then we have higher stability. At present, should we use a standard normal distribution without correlation then the inequality will have a lower error, because only the diagonal of Σ is estimated, thus $d\epsilon \leq \frac{d(d+1)}{2}\epsilon$. So, we must improve the bounds based on the existing change – the use of correlations – and find other approaches which either use this improved stability or improve it further.

4.1. Improvements using Correlated Rules

The importance of a feature in BARBE is the sum of the importance of each rule for a feature, A, which is analogous to the rules appearing as part of the explanation of A, $A = \bigcup_{A \in R} R$. So the value $imp(A)$ represents this sum and the expected value $E[imp(A)]$ is the expected importance of the significant rules which apply to x_p. The probability of rules appearing in the explanation form a discrete distributions, sampled using Bonferroni boxes, however by assuming that there are some set of correlated features c between the rules we can get improved bounds by using a Bonferroni nest as per Bax and Ouimet [13]. Based on previous work, the bound should improve by a ratio of $\sqrt{c}$. As a result, we can adjust the bound on the correlated side given the following assumptions:

Assumption 2. *The value being estimated, $imp(A)$, is the function of a set of discrete events occurring, in our case these are significant rules being used in the explanation. Where the true average $E[imp(A)]$ is for the actual explanation of the black-box. Which is the case for the importance values of BARBE [4].*

Assumption 3. *Frequencies of these rules can be "combined," reducing the total number of rules that need to be used when calculating the bound. As per Bax and Ouimet [13], this leads to an improvement of $\sqrt{c}$ on the error of each bound involving the c combined frequencies.*

Assumption 4. *We cannot reasonably exceed a training set size of 1000 samples used to estimate the value without high time or processing costs. Since bound improvements above other approaches were found for this number of samples [13].*

$$p((imp(A) - \frac{\epsilon_L}{\sqrt{c}}) - (imp(B) + \epsilon_H) \geq \frac{d(d+1)}{2}\epsilon) \geq \delta \tag{4.3}$$

This improvement of $\sqrt{c}$ is only yielded by the rules which have combined probabilities, however for brevity we will not be including the remaining error. As we find below, the extra error does not have an impact on the expected effect of this improvement; we can just leave the remaining error on the left-hand-side (LHS) of the inequality. Now, by moving the error terms we can tell how well the improvement helps the stability.

$$p(imp(A) - (imp(B) + \epsilon_H) \geq \frac{d(d+1)}{2}\epsilon + \frac{\epsilon_L}{\sqrt{c}}) \geq \delta \tag{4.4}$$

So long as the new error is less than the old error $d\epsilon + \epsilon_L$, then this is a more stable result $\frac{d(d+1)}{2}\epsilon + \frac{\epsilon_L}{\sqrt{c}} \leq d\epsilon + \epsilon_L$. By separating the two distinct error terms we get a cleaner relationship $\epsilon \leq \frac{2\epsilon_L}{d(d-1)}(1 - \frac{1}{\sqrt{c}})$. For an existing set of value's which satisfy the relationship in Equation 4.4, as $1/\sqrt{c} \to 0$ the error of the importance ϵ_L decreases. Specifically, this inequality improves when the errors from the bounds are very high in comparison to the errors on the distribution. This will occur when we have a low number of perturbed samples (large ϵ_L and ϵ_H, fairly likely), or when there is a very low amount of noise from the training data (ϵ is small, which is less likely).

4.2. **Improvements using Correlated Bias Rules**

In the previous section, we were able to show that with correlated rules – a result of correlated features – the bounds on importance values are improved and therefore the stability of the explainer is improved. This property should help us yield further gains, since there are more possible "rules" to use in the explanation. Recall, that for BARBE there are both positive and negative importance values, dependent upon the class, and since we measure – rather than sample from – the responding class labels $Y = y$ we currently cannot benefit from the improved bounds for these labels as an extra "rule".

Assumption 5. *The "class label," $Y = y$, is also in the distribution as the bias rule which contributes to the importance of all the features "$\phi \to Y = y$". By removing the bias, this rule instead contributes to $imp(A)$ based on its correlation to the other rules.*

$$p((imp(A, Y = y) - \epsilon_L) - (imp(B) + \epsilon_H) \geq \frac{d(d+1)}{2}\epsilon) \geq \delta \tag{4.5}$$

If Assumption 5 applies, then we get one more correlated rule $c + 1$, so similar to before, this inequality improves while $\epsilon \leq \frac{2\epsilon_L}{d(d-1)}(1 - \frac{1}{\sqrt{c+1}})$. When we have many correlated rules it may not contribute much to improving stability, but when there are few – or when many other rules are correlated with the bias rule – it can increase stability quite a bit. However, the correlation requirement of the new rule is not necessarily true; the rule is purely bias '$\phi \to Y = y$' and we do not technically sample the class label from a distribution. To fulfill the correlation requirement, a modification must be made which causes us to induce a correlation.

If we choose a required proportion of each class in Y, then we are technically sampling the event and there are now correlated rules which matter to each $\phi \to Y = y$, leading to the further improved bounds above. It is noteworthy that this improvement should always be possible, whether the surrogate training data used a covariance or not. So if this property holds, then we expect to see that balancing alone will improve the stability of the explainer, which we can test for later.

4.3. Effects of Mixed Distributions of Rules

Keep in mind, the goal of a local explanation is to be more true to the local distribution of the data about x_p. Thus, we aim to represent the complex behaviour of the training data's distribution, to do so, we assume:

Assumption 6. *The local $imp(A)$ is a mixture of several distributions rather than one single distribution. Therefore, $imp(A) = E[imp(A)] + \pi_1(L + L_\epsilon)\epsilon + ... + \pi_k(L + L_\epsilon)\epsilon \leq E[imp(A)] + kd(d+1)\epsilon/2$. For some probabilities $\pi_i \in [0, 1]$ such that $\sum \pi_i = 1$, which we simplify for potentially noisy choices of clusters and probabilities to just $kd(d+1)\epsilon/2$.*

We get the RHS term by not only assuming the instability of the mixture, but the instability of whether an individual cluster will be selected. So we expect there are around k different clusters that we should select from for our point to be in, if we select just one. However, if we use the appropriate proportional representation we may be able to get better stability from the same number of clusters, which we investigate further in this work.

$$p((imp(A, Y = y) - \epsilon_L) - (imp(B) + \epsilon_H) \geq k\frac{d(d+1)}{2}\epsilon) \geq \delta \tag{4.6}$$

In this case, if we increase the number of clusters we expect the required stability to be too high to improve the result. However, if both sides compared importances A and B have correlated features we may be able to make use of the improved bounds to still have better stability over the original standard normal distribution situation.

$$p((imp(A, Y = y) - \frac{\epsilon_L}{\sqrt{c+1}}) - (imp(B, Y = y) + \frac{\epsilon_H}{\sqrt{c+1}}) \geq k\frac{d(d+1)}{2}\epsilon) \geq \delta \tag{4.7}$$

Now this improves beyond our starting point so long as $\epsilon \leq \frac{2(\epsilon_L+\epsilon_H)}{kd(d-1)}(1 - \frac{1}{\sqrt{c+1}})$, in other words we can usually improve stability – even beyond multivariate normal distributions – when there are at least 2 sufficiently correlated features and the error of both groups is similar (i.e. $\epsilon_L + \epsilon_H \approx 2\epsilon_L$). For our purposes, we assume that the most minimal non-trivial number of clusters, $k = 2$, still improves stability, going beyond this may be difficult since clusters can disappear, adding extra noise ϵ, then the only result that matters beyond this is how the data are clustered.

5. Applied Improvements to Stability

In this section, we propose the applied algorithms which implement the improved bounds we discussed in Section 4. **Correlated Perturbations (+P)** are already applied to BARBE [5], so in this section we discuss class balancing and clustering techniques which should further improve the locality of an explanation without sacrificing the stability.

5.1. Class Balancing (+B)

In Section 4.2, we established that balancing the perturbed training data's classes can improve the stability of BARBE. Balancing has been used for performance improvements in AI applications before [14] and is usually ensured by explainers that search for samples – like LORE [7] – but balancing perturbed data is still rarely applied in post-hoc explanations. We propose using a simple under-sampling approach where we produce tn perturbations, where n is the original number of perturbations and t is the number of times we need to produce perturbations to have $n/2$ samples of each class. We then under-sample to get only $n/2$ samples of the larger class. The proposed approach is used to keep the sampling as simple as possible, but in testing we also use the SMOTE [15] balancing approach as an alternative approach with a bit more complexity.

5.2. **Clustered Perturbed Data (+C)**

Previously in Section 4.3, we found that using local behavioural groups should not reduce stability significantly if there are two groups, given our loose assumption of some amount of correlated rules in both compared features' rule sets. Since we want a better local representation of data, clustering seems to be an obvious pick for this grouping. If we use a global covariance it may not represent realistic behaviours found in the dataset, where there can be different behaviours locally.

If our goal is to have a simple clustering approach – to avoid over-complication – then we can use K-Means clustering. The K-Means approach is an iterative expectation-maximization (EM) algorithm which selects k cluster mean points, clusters each point based on distance from the means, and recalculates the means, repeating the process until the means do not change. K-Means clustering produces reliable clusters consistently enough that it is usually used as an initialization for other methods. For our purposes, we use the group x_p is in as the "local" behaviour of the perturbations, but this may not be the ideal solution.

A part of Assumption 6 which we did not initially use, was that data can be part of some proportion of each group, a common setup for clustering algorithms [16]. K-Means does not provide group proportions, so it is not ideal for this application. Another issue we may encounter is outliers in the data, so an ideal clustering method should account for these without shifting the distribution. To accommodate these requirements, we propose using clustering based on Mixtures of t-Distribution [16], another EM algorithm, which return a proportional membership for the points and cluster covariances Σ_k while handling outliers in the data since the model-based method assume data can outliers.

We use the proportional membership value when generating perturbations, so as opposed to ordinary BARBE which generates $x \sim \mathcal{N}(x_p, \Sigma)$ for each group of some proportion π_k we generate $\lfloor \pi_k n \rfloor$ samples from $x \sim \mathcal{N}(x_p, \Sigma_k)$. Where n is the total number of perturbations to produce, π_k comes from the probability of x_p being in each cluster k, and $\sum \pi_k = 1$.

6. **Experimental Setup**

In other works [7, 8], experiments for stability keep the training data and input conditions the same and only evaluate instability caused by the random perturbations used to train the white-box model $x \sim \mathcal{N}(x_p, \Sigma)$. However, these approaches are limited because there are far more sources of error than the perturbations. For example, the Relative Input Stability (RIS) – a measure which we use to evaluate stability – measures instability coming from small changes to the sample being explained, x_p. To use this measure, we make small changes, $x'_p \sim \mathcal{N}(x_p, \Sigma/10)$, we divide the covariance Σ by 10 to keep the changes so small that they will not realistically behaviourally change the sample.

We also consider the effect of different training data being passed to the explainer, as this will have an impact on the distribution and the derived Σ used to perturb the sample. We sample a random 1/3 of the training data to use for each of 10 repetitions which are used to calculate the stability of the explanation. This goes beyond what other methods like S-LIME have done for testing, but based on the theory our approach's stability should still improve in this setting.

Experiments are run for 30-50 randomly selected points-of-interest from the testing data (depending on the total number of samples in the dataset) in each of seven diverse datasets, with varying class balance and combinations of categorical and continuous features. The full metadata for these data are shown in Table 1. We use two different black-boxes, both cross validated on a 60-20-20 train-test-validation split. The black-boxes are a neural network and random forest.

For BARBE, we compare each improvement independently and in combination (e.g. +P, +B, +C, +PB, ...) alongside the algorithm used to implement the change (for example

Dataset	# of Samples	# of Features	# of Categorical Features	Balance
Wine	178	13	0	0.33
Loan	614	11	6	0.31
German Credit	1000	20	13	0.30
Employee Attrition	1470	30	7	0.14
Breast Cancer	569	30	0	0.47
COMPAS	7214	28	10	0.45
Spam	4601	57	0	0.39

Table 1. Metadata for each dataset used in the experiments. **Balance** is the minimum proportion of one class to the overall data $min_y(\sigma(Y = y)/|D|)$, so 0.5 would be perfectly balanced for most datasets with two classes.

SMOTE vs. under-sampling). For the final comparison against competitors we use the most stable implementation of BARBE which was found in these tests.

In addition to comparing with competitors (LIME, S-LIME, SHAP, and LORE), results are compared to a baseline, using random feature selection for stability and constant predictions for class prediction accuracy metrics. Comparisons use a one-sided paired sample t-test, with the null hypothesis $H_0 : g(e_{comp}) - g(e_{our}) \geq 0$ and the alternative hypothesis $H_\alpha : g(e_{comp}) - g(e_{our}) < 0$. Where e is a set of explanations and g is a metric used to evaluate the explanations. The paired test ensures that outlying samples across both compared methods are considered equally, rather than them shifting the mean metrics and having a greater impact on the evaluation.

6.1. **Evaluation Metrics**

To measure stability we use the Jaccard index, like other work [7, 8]:

$$Jaccard = \frac{\left(\sum_{e,e' \in \mathcal{E} \forall e \neq e'} \frac{|e \cap e'|}{|e \cup e'|}\right)}{\binom{|\mathcal{E}|}{2}} \tag{6.1}$$

Where $\mathcal{E}$ is the set of all ordered explanations for a single point-of-interest, and e and e' are ordered feature importance based explanations of the form $e = \{e_1, e_2, ..., e_d\}$. When feature rankings are available we consider the top $d = 5$ features, where users are most likely to concern themselves. Sometimes we expect that explanations may require more features, so to generalize better to many datasets we consider the top $d = N/2$ features where N is the number of features, if there are that many features used in the explanation. We set hard cut offs since the length of an explanation may change. Using one feature or all of them would always be the most stable without a flexible size. In our application, we use Jaccard index to understand when feature importance becomes most random.

We also use the inverse Jaccard index, $Inverse\ Jaccard = (1 - Jaccard)$, for sets of 10 different samples – each with feature importance averaged over the 10 repetitions – which have been explained on the same dataset and black-box model. This alternative metric, tells us how similar each explanation is to each other. This is used to evaluate how much stability is caused by the explainer versus how much stability is a result of using the same explanation for every sample, meaning the result would be less likely to be a local explanation if the Inverse Jaccard index is consistently low.

Another metric which we use for stability is the RIS [9]:

$$RIS(x_p, x'_p, e_{xp}, e_{x'p}) = \max_{x'_p \sim \mathcal{N}(x_p, \Sigma/10)} \left(\frac{|| \frac{(e_{xp} - e_{x'p})}{e_{xp}} ||_2}{max(|| \frac{(x_p - x'_p)}{x_p} ||_2, \epsilon_{min})} \right) \tag{6.2}$$

Where x'_p is a minute perturbation of x_p, such that the class does not change $f(x_p) = f(x'_p)$. We produce 1 for each repetition, so we have 9 perturbations x'_p and 1 test with the true x_p. Where e_{xp} is the vector of feature importance of an explanation for x_p. In comparison to the Jaccard Index, this value considers the change in the feature importance value as opposed to the relative scale of the feature importance values. So it may be better when many features are used in the explanation. As with previous work [9], we use the RIS on a log scale, $log(RIS(x_p, x'_p, e_{xp}, e_{x'p}))$.

We additionally use two prior evaluation metrics to ensure that performance is not reduced by our changes: fidelity (weighted agreement between the white-box and the black-box [5]) and hit-rate (how often the white-box could correctly predict original unperturbed input row). We consider these as consequent measures, if fidelity or hit-rate are low then the stability may be a result of a poor classifier. If the stability is low, then we still may not trust a high fidelity and hit-rate approach. As such we are considering comparisons as a goal but ultimately whether the proposal improves stability for BARBE without trading off fidelity or hit-rate.

6.2. **Reproducibility**

Code is available on an anonymized GitHub (https:\anonymous.4open.science/r/Stable-BARBE-014F/) randomization for subset and data selection use pre-defined seeds to ensure data do not change across tests and devices. Pre-trained models are also available. LIME, S-LIME, SHAP, and LORE are run with default settings, some added code is used to make output formats consistent for experiments.

7. **Results**

We compared experiments for SMOTE against Under Sampling, and similar experiments for K-Means, Mixtures of t-Distributions (MixT), and Proportional MixT (prop-MixT). We found that SMOTE had a high error rate due to the secondary class not appearing in any perturbed samples, a problem which Under Sampling did not have, leading us to use Under Sampling for the following tests as +B. Among the clustering approaches, KMeans was relatively stable as we expected, and the other approaches ended up being costly to the fidelity both MixT (0.725) and prop-MixT (0.720) had lower average fidelity than K-Means (0.756), leading us to use K-Means for our final experiments as +C.

As expected in theory, adding perturbations and balancing each significantly ($p<0.001$) improved the stability of BARBE (Figure 1 and Table 2 Jaccard +P, +B, +PB, ...) – while retaining high fidelity – without making the explanations significantly more consistent across samples (Inverse Jaccard ($d = N/2$)). So the explanations from BARBE+PB are still unique even with the added stability.

The improvement to stability from class balancing was beyond our expectations (Table 2, BARBE vs. +B $p<0.001$, +P vs. +PB $p<0.001$). This occurs without increasing the stability between different samples, so it follows that the number of correlated rules is low, $1/\sqrt{c} \geq 1/\sqrt{5}$ thus $c + 1$ represents a significant improvement. Alternatively, the class is correlated with many rules, leading to $1/\sqrt{c} \leq 1/\sqrt{2}$ improvement to many bounds.

While a similar, rule-based method LORE may have slightly higher stability than BARBE, LORE has lower fidelity (Figure 1 Fidelity), so we expect that the explanations from BARBE+PB align better to true, local explanations than LORE does. Clustering does appear to slightly reduce fidelity, this could be since the clusters are not as representative of the local behaviour due to instability. However, the stability can still be improved further with more tightening of the bounds (ϵ_L and ϵ_H) or through reducing the impact the number of dimensions has on the importance noise (reduce d or ϵ in $\frac{d(d+1)}{2}\epsilon$).

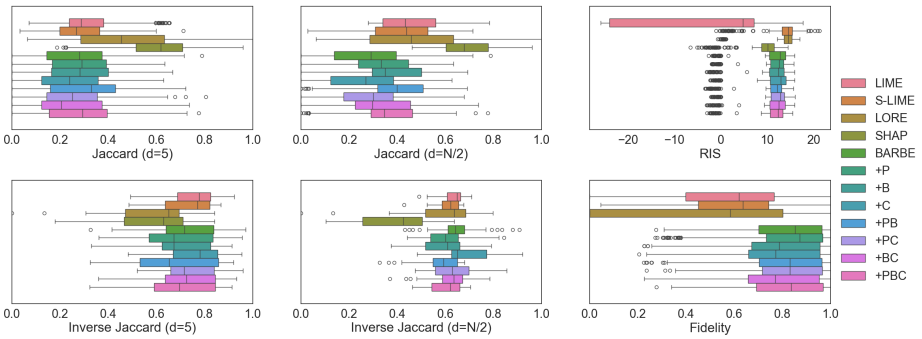


Figure 1. Box plots of each evaluation metric for competitors and our approaches (+P, +B, ...) for **all datasets and both black-box models**. Results indicate initially that the stability of our approach is similar to LIME, S-LIME, and LORE, while it outperforms each in terms of Fidelity (SHAP is not a surrogate, so it does not have Fidelity). We also find that SHAP, which is relatively stable, has high stability across different samples to explain (in plots of the Inverse Jaccard), indicating that it may produce a global explanation rather than a local one.

	LIME	**S-LIME**	**LORE**	**SHAP**	**BARBE**	**BARBE+P**
+P	F**	RIS**, F**	RIS**, F**	-	5**, N/2**, F**	-
+B	F**	RIS**, F**	RIS**, F**	-	5*, N/2**	-
+C	F**	RIS**, F**	RIS**, F**	-	-	-
+PB	F**	5*, N/2*, RIS**, F**	F**	-	5**, N/2**, F**	5**, N/2**
+PC	F**	RIS**, F**	RIS**, F**	-	F**	-
+BC	F**	RIS**, F**	RIS**, F**	-	N/2**	-
+PBC	F**	RIS**, F**	RIS**, F**	-	5**, N/2**, F**	N/2**

Table 2. Paired sample t-Test for **all datasets and black-boxes**, results for our methods compared to existing methods and competitors (BARBE+P [5], previously untested for stability). Results of 5 is for t-tests of Jaccard (d=5), N/2 is for Jaccard ($d = N/2$), RIS is for the RIS statistic, and F is for Fidelity. We use "-" when there are no rejected null hypotheses. We record two levels of significance for rejected null hypotheses, *(pvalue<0.005) and **(pvalue<0.001).

In our experiments, we found that the stability of S-LIME is worse or negligibly better than LIME, unlike previous results which used static training data [8]. SHAP is quite stable, but this also holds across different samples since the Inverse Jaccard value is low. This result suggests that SHAP produces a consistent overall explanation, which is likely the global explanation of the model, rather than a local explanation for the sample, leading every explanation to be global and therefore similar. Even as this reduces the significance of SHAP for local explanations, we may be interested in further improvements to the clustering by exploiting a similar idea, using the global model to anchor unstable training data clusters for additional stability.

When discussing the underlying theoretical work, we emphasized that either ϵ needs to be low or ϵ_L and ϵ_H need to be high. This result is best visualized in Figure 2, where the COMPAS dataset is shown. With a large number of training samples (7214) relative to the number of features (28), we expect that ϵ will be lower. We also see that the best stability improvements for BARBE+PB appear in this group. So, as we expected, not every dataset is going to be clearly significantly improved by our approach, but a key subset of slightly-less noisy datasets will yield significant improvements which are important for establishing user trust in the stability of BARBE+PB.

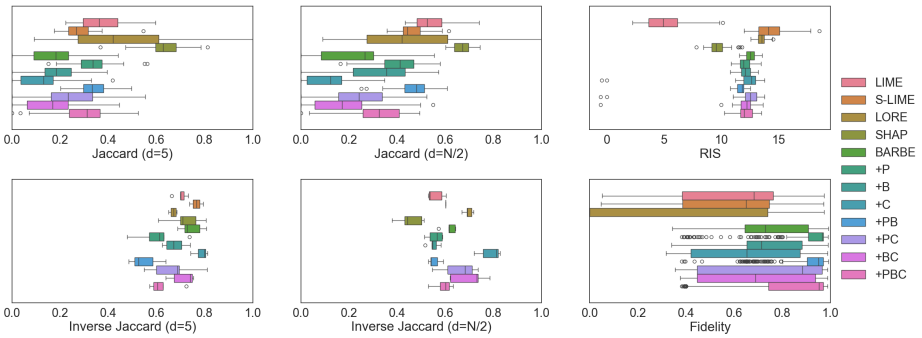


Figure 2. Box plots of each evaluation metric for competitors and our approaches (+P, +B, ...) for the **COMPAS dataset using the neural network model**. These results represent specific cases where BARBE+PB is a significant improvement to stability over BARBE in both Jaccard and RIS, this is a dataset where the theoretical requirements suit the improvements (correlations, causal rules, and lower ϵ).

8. Future Work and Conclusion

8.1. Future Work

Our discussion of SHAP, concluded that the high stability it exhibits is likely a cause of it being anchored to the static model of the black-box. This is done by producing a global explanation rather than using an unstable training sample for local explanations. BARBE+PBC could benefit from this too if we implemented changes to the clustering algorithm to anchor it to the black-box, then we may achieve a similar effect to SHAP. To our knowledge, there are very few clustering algorithms designed which require a class-label. Fewer algorithms exist which balance the classes and even base clusters on their fit to a point-of-interest. This could be an interesting future research direction for post-hoc explanations.

8.2. Conclusion

In this work, we explored two new techniques to surrogate-based post-hoc explanations, balancing (+B) and clustering (+C), as approaches for improving the stability and locality of a post-hoc explanation. We also used a theoretical standpoint to establish the consistent stability improvements yielded by both balancing and the pre-existing improvements to the perturbations using multivariate perturbations (+P). By conducting thorough experiments and statistical testing, we showed that the theoretical implications hold, but also that there is further room to improve for BARBE. As the approach with the highest fidelity – and the most likely to produce meaningfully local explanations – BARBE+PB greatly improved stability over BARBE and BARBE+P. When using the Relative Importance Stability, our approach also significantly outperforms both S-LIME and LORE, where the prior method even underperforms LIME, a result which we found due to testing for unstable training data. Our results additionally serve as evidence for why the stability of BARBE+PBC was slightly lower than BARBE+PB, opening a new avenue of exploration for further stability improvements from clustering. These results not only reinforces why we should conduct severe testing in XAI – as some stability gains were unclear from just medians – but also for why we should use theoretical reasoning. Both have helped us show that BARBE+PB is a significantly more reliably stable explanation approach than the original BARBE and BARBE+P alone.

References

[1] Parliament of Canada. *Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*. 44th Parliament, 1st Session. 2022. URL: https://www.parl.ca/legisinfo/en/bill/44-1/c-27.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin. "\"Why Should I Trust You?\": Explaining the Predictions of Any Classifier". In: *CoRR* abs/1602.04938 (2016). arXiv: 1602.04938. URL: http://arxiv.org/abs/1602.04938.

[3] D. S. Watson. "Conceptual challenges for interpretable machine learning". In: *Synthese* 200.2 (2022), p. 65.

[4] M. Motallebi, M. T. A. Anik, and O. R. Zaïane. "Explaining decisions of black-box models using barbe". In: *International Conference on Database and Expert Systems Applications*. Springer. 2023, pp. 82–97.

[5] I. Smith and O. Zaiane. "Faithful Perturbations and Evaluations for Post-Hoc Local Explanation Methods". In: *Proceedings of the Canadian Conference on Artificial Intelligence* (2025). https://caiac.pubpub.org/pub/gs2ywmlt.

[6] S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, 4768–4777. ISBN: 9781510860964.

[7] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. "Local Rule-Based Explanations of Black Box Decision Systems". In: *ArXiv* abs/1805.10820 (2018). URL: https://api.semanticscholar.org/CorpusID:44063479.

[8] Z. Zhou, G. Hooker, and F. Wang. "S-LIME: Stabilized-LIME for Model Explanation". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. Virtual Event, Singapore: Association for Computing Machinery, 2021, 2429–2438. ISBN: 9781450383325. DOI: 10.1145/3447548.3467274. URL: https://doi.org/10.1145/3447548.3467274.

[9] C. Agarwal, N. Johnson, M. Pawelczyk, S. Krishna, E. Saxena, M. Zitnik, and H. Lakkaraju. "Rethinking Stability for Attribution-based Explanations". In: *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*. 2022. URL: https://openreview.net/forum?id=BfxZAuWOg9.

[10] X. Xiang, H. Yu, Y. Wang, and G. Wang. "Stable local interpretable model-agnostic explanations based on a variational autoencoder". In: *Applied Intelligence* 53.23 (Sept. 2023), 28226–28240. ISSN: 0924-669X. DOI: 10.1007/s10489-023-04942-5. URL: https://doi.org/10.1007/s10489-023-04942-5.

[11] X. Xiang, H. Yu, Y. Wang, and G. Wang. "Stable local interpretable model-agnostic explanations based on a variational autoencoder." In: *Applied Intelligence: The International Journal of Research on Intelligent Systems for Real Life Complex Problems* 53.23 (2023), pp. 28226–28240. ISSN: 0924-669X. URL: https://login.ezproxy.library.ualberta.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edssjs&AN=edssjs.32CD9F24&site=eds-live&scope=site.

[12] W. Hamalainen. "Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures". In: *Knowledge and Information Systems* 32 (2012), pp. 383–414.

[13] E. Bax and F. Ouimet. "Bounding Means of Discrete Distributions". In: *2021 IEEE International Conference on Big Data (Big Data)*. 2021, pp. 5024–5032. DOI: 10.1109/BigData52589.2021.9671544.

[14] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. P. Chen. "A survey on imbalanced learning: latest research, applications and future directions". In: *Artificial Intelligence Review* 57.6 (2024), p. 137.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *J. Artif. Int. Res.* 16.1 (June 2002), 321–357. ISSN: 1076-9757.

[16] D. Peel and G. J. McLachlan. "Robust mixture modelling using the t distribution". In: *Statistics and computing* 10.4 (2000), pp. 339–348.