Machine Learning Interpretability through Contribution-Value Plots

Citation for published version (APA):

DOI:
10.1145/3430036.3430067

Document status and date:
Published: 08/12/2020

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Machine Learning Interpretability through Contribution-Value Plots

Dennis Collaris
Eindhoven University of Technology

Jarke J. van Wijk
Eindhoven University of Technology

ABSTRACT
The field of explainable artificial intelligence aims to help experts understand complex machine learning models. One key approach is to show the impact of a feature on the model prediction. This helps experts to verify and validate the predictions the model provides. However, many challenges remain open. For example, due to the subjective nature of interpretability, a strict definition of concepts such as the contribution of a feature remains elusive. Different techniques have varying underlying assumptions, which can cause inconsistent and conflicting views. In this work, we introduce Local and Global Contribution-Value plots as a novel approach to visualize feature impact on predictions and the relationship with feature value. We discuss design decisions, and show an exemplary visual analytics implementation that provides new insights into the model.

CCS CONCEPTS
• Human-centered computing → Visualization. • Computing methodologies → Machine learning.

KEYWORDS
visualization, machine learning, interpretability

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

VINCI 2020, December 8–10, 2020, Eindhoven, Netherlands
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8750-7/20/12... $15.00
https://doi.org/10.1145/3430036.3430067

1 INTRODUCTION
The past decade has witnessed a sharp increase in the popularity of artificial intelligence and machine learning. This prevalence has resulted in a wide variety of new approaches and techniques (e.g., deep learning) that have achieved astounding results previously not deemed possible [10, 15]. Clearly, these models have advanced over their predecessors in terms of predictive performance (e.g., accuracy, precision, recall, F1-score). However, there are more properties of these models that have not received as much attention, such as complexity, interpretability, and fairness [5]. As a consequence, state-of-the-art techniques are ever increasing in complexity, yielding black-box models that cannot easily be inspected or verified.

The field of explainable artificial intelligence (XAI) has recently gained a lot of traction as it aims to alleviate these issues. It exposes more details about the behavior of complex machine learning models, which helps experts to verify and validate model predictions. XAI has proposed a variety of new techniques to show the impact of a feature on the model prediction [6, 8, 14, 18]. However, due to the novelty of the field many challenges remain open.

In particular, the complex and ill-defined nature of interpretability hinders a strict definition of concepts such as contribution of a feature. Different techniques have varying underlying assumptions, which can cause different and conflicting results. In this work, we present Local and Global Contribution-Value plots as a novel technique to explain machine learning models. The plots visualize the feature contribution to a prediction, as well as the relationship with feature value. Such information about the model is typically conveyed with multiple techniques, which could lead to contradictory results. We discuss relevant design decisions, and show an exemplary visual analytics instrumentation and show it enables insights into the model that were previously not possible.

2 BACKGROUND AND RELATED WORK

Figure 1: Design space of interpretability methods. Blue boxes indicate our contribution. The x-axis denotes feature values; the y-axis denotes a prediction (A, B, C) or a contribution (D, E, F).

Visualization can help data scientists to get a better understanding of black box models. For trivial prediction problems this can be done by inspecting the predictions of a model directly (Figure 1A). Scatter plots can be used to show the relationship between prediction probability \( \hat{y} \) and feature value \( x \). However, for any non-trivial prediction problems, there are likely many interactions between features which make it impossible to identify patterns and trends.

2.1 Local Partial Dependence Plot
To help to gain insight into models, Friedman [6] introduced the Partial Dependence Plot (PDP). This is a sensitivity analysis technique that shows how the prediction \( \hat{y} \) changes as target features \( z_t \) are varied over their marginal distributions (Figure 1B).

To define partial dependence for a data point \( x \), let \( z_t \subset \{x_1, \ldots, x_n\} \) be a set of target features, and \( z_c \) the complement of \( z_t \) such that

\[
z_c \cup z_t = x, \quad z_c \cap z_t = \emptyset \tag{1}\]

The prediction \( \hat{f}(x) \) in principle depends on both subsets:

\[
\hat{f}(x) = \hat{f}(z_c) + \hat{f}(z_t) - \hat{f}(x) \]
with a local sampling region around an instance. The coefficients work to explain machine learning predictions [12, 13].

An alternative approach to gain insight into machine learning models is the feature contribution technique (Figure 1D). Such methods yield feature contribution vectors that indicate how much every feature contributed to a prediction.

2.2 Global PDP and ICE plot

Local PDPs provide a great insight into a single prediction. However, for many applications such a local explanation is not sufficient. In an explorative setting, experts would like to inspect much more than just a single prediction. For example, the explanation of a single prediction is not helpful for diagnosing problems with a model, or for model refinement. Even if there is a single prediction of interest, instance-level explanations do not show whether they are specific to that instance, or generalize to a larger set of instances. For these cases we need global explanations. To get a global insight into the entire model, Friedman [6] proposes averaging the local partial dependence lines of all training data as follows:

\[
\hat{f}(z_t) = \mathbb{E}_{x \sim P} \left[ f(x) \right] = \int f(z_t, x) M_x \, dx \approx \frac{1}{N} \sum_{i=0}^{N} \hat{f}(z_{t,i}, z^N_x) \tag{4}
\]

where \( M_x \) is the marginal probability density of \( z_x \). This global PDP is used in visualization work to explain and compare machine learning models [20, 21]. However, Friedman notes that Equation 4 does not hold when there is a strong interdependence amongst features, which often the case for complex black box models.

To deal with interdependence, Goldstein et al. proposed an alternative called Individual Conditional Expectation (ICE) plot [8] by superimposing all individual local partial dependence lines. This reveals patterns that would be otherwise be hidden by averaging. For example, the plot in Figure 1C shows two clusters of partial dependence lines that would not be apparent in a global PDP.

2.3 Feature Contribution

An alternative approach to gain insight into machine learning models is the feature contribution technique (Figure 1D). Such methods yield feature contribution vectors that indicate how much every feature contributed to a prediction.

Initially, Baehrens et al. [1] showed that machine learning models can be explained using the derivative of the class probability function. The reasoning is that if a small change in feature value leads to a large change in the prediction probability (or regression output), that feature is relevant for the prediction. They note, however, that an exact derivative for the majority of models does not exist.

To this end, LIME [18] was proposed. It solves this issue by fitting a linear regression surrogate model to the class probability gradient with a local sampling region around an instance. The coefficients of the linear model effectively approximate the derivative of the probability function, regardless of whether a formal derivative exists. Next, the approximation can be used to show which features have the most impact on a prediction [3].

Another prominent approach for feature contribution are Shapley values [11, 14, 19]. This method estimates the contribution of a feature by comparing the class probability of a prediction including and not including this feature [16]. The absence of a feature is estimated by averaging the predictions for different values for that feature sampled from the training data distribution.

Any of these techniques yield feature contribution vectors that give a quick overview of which feature had an impact on a single prediction. However, it remains unclear for which values in general that feature is relevant. For example, in a medical trial where feature attribution shows that ‘dosage’ is important predictor for recovery, we would also like to know what values of ‘dosage’ were most relevant. In addition, these methods only target single predictions whereas some use cases require a global perspective on the model.

3 LOCAL CONTRIBUTION-VALUE PLOT

To alleviate the limitations of previous techniques, we propose the Local Contribution-Value (LCV) plot. The curves are generated in the same way as PDPs (Section 2.1), but instead of class probability values we use feature contribution values (Figure 1E). This yields a plot that reveals how the feature contribution varies for changes in feature value. It has some key advantages over local PDPs.

First, contrary to a PDP [6], the LCV plot is also effective when features are heavily correlated. For example, if feature \( k \) and \( f \) are correlated, changing the value of either does not change the prediction, while changing both would. As the sensitivity analysis used in PDPs only alters the value of a single feature at a time, the PDP would not show variation in the prediction. In contrast, LCV plots use an explanation technique that considers a wider region of feature space (compared to a single point), which enables them to show variation in contribution even when features are correlated.

Next, for certain use cases the LCV plot may be easier to read and compare. To infer relevance of a feature in a PDP plot, experts have to consider the slope of the line. Previous work has shown that human slope estimation is not trivial and prone to be biased (i.e., angle contamination) as our visual system is geared towards judging angle rather than slope [2]. Hence, our graphical perception of slopes prohibits any exact judgement of contribution or importance in prediction-value plots. LCV plots encode feature contribution with position, making it easier to read exact values.

This does mean that the prediction probability is not directly encoded in LCV plots. We argue PDP and LCV plots serve a different (and complementary) purpose. When experts are interested in the predictions for specific data points, PDPs are more suitable. However, when trying to understand how the model makes predictions, LCV plots are more suitable.

Finally, the LCV remains a local approach focusing on a single instance. This makes it difficult to get a global overview of the model, and whether a feature that is locally relevant is always relevant, or only for a small number of instances.

4 GLOBAL CONTRIBUTION-VALUE PLOT

For a global overview, we propose using the same procedure as for an ICE plot: to superimpose LCV plots to show the contribution for an entire data set (Figure 1F). This helps experts to get a global overview of the model behavior for typical data. We refer to this approach as the Global Contribution-Value (GCV) plot.
The GCV shows more clearly which values of a feature have a significant impact on the model prediction, which helps to understand the model. As an example, we examine the Wine Quality dataset [4]. Figure 2b shows two different thresholds (3.05 and 3.35) for pH that the Random Forest model uses to determine wine quality.

Next, GCV plots enable the comparison of feature importance at different feature values. For instance, for the selected instances in Figure 2b, the first threshold contributes more than the second.

In addition, in a GCV plot it is much easier to find patterns and clusters compared to ICE plots. Such expert-guided subgroup discovery can for instance be used to assess model fairness, and to discover different 'strategies' a model has for predicting the same class. There are two reasons for this.

First, in ICE plot lines the differences in prediction probability lead to vertical dispersion that obscures global patterns. Second, feature contribution techniques have to simplify in order to approximate the reference model. For instance, LIME averages contribution over a sampling region around an instance. We found that this simplification yields smooth curves in LCV and GCV plots, making it easier to spot more subtle patterns [9]. This also gives an intuitive visual interpretation of the kernel size parameter in LIME: changing this parameter affects the smoothness of the curves.

For example, in the ICE plot shown in Figure 2a it is difficult to spot patterns. The corresponding GCV plot in Figure 2b clearly highlights two different clusters. These clusters can be thought of as different 'strategies' of the model, as these correspond to subsets of the data set with different explanations (i.e., contribution values).

The lower vertical dispersion also enables intuitive interactive selection by means of lasso brushing [17], as shown in Figure 2b. In an ICE plot, lasso selection does not yield any interesting clusters; this would require selecting lines based on angle.

![Image](a) ICE plot: Due to the huge variance in class probabilities it is difficult to find patterns.

![Image](b) GCV plot: Selected polylines in blue, revealing two clusters with diverging contribution values.

Figure 2: Two visualizations of a Random Forest (trees = 100) trained on the Wine Quality dataset [4], showing feature "pH".

5 DESIGN

We built a visual analytics instrumentation of all discussed techniques (Figure 1), as they are valuable in different situations. More detail and a usage scenario are shown in the supplementary video. It can be used by data scientists to understand how a feature impacts model predictions on a global level. In addition, it also shows what values of a feature are relevant. Through interaction, different patterns in feature contribution can be analyzed.

5.1 Feature Contribution technique

Even though feature contribution techniques can provide great insight into model predictions, the output of different techniques may vary significantly, making it challenging to compare them.

The examples in Figure 3 show that LCV plots with different explanation techniques can vary significantly. The difference is that LIME contribution values are approximate (partial) derivatives, whereas Shapley contributions are additive: the sum of all feature contributions (plus the constant base rate, i.e., the average predicted value) equals the class probability \( \hat{y} \). For LIME, the contribution values need to be composed with the feature values first. Next, the linear regression intercept (\( \alpha \)) is not constant but varies per instance.

For this paper we focus on LIME contribution as it has a more straightforward interpretation (i.e., which small change in feature value results in a big change in prediction) than Shapley values, and a lower computational cost [7]. A kernel size of 0.5 was used, but we encourage tweaking this parameter on a per-dataset basis.

5.2 Visual encoding

In PDP and LCV plots, a single instance is traced over the entire marginal distribution of a feature. This may yield data points that are out-of-distribution (e.g., a person with age 5 and height 200cm). Such data points force the model to extrapolate to an unseen part of the feature space, which could be misleading.

To account for this, we gradually fade out polylines as they get further away from the original data point. Any kernel can be applied, but in our implementation we use a triangular kernel:

\[
\alpha(u) = \max\left(0, 1 - \frac{|u|}{\tau R_t}\right) \quad (6)
\]

where \( \tau \) is a configurable parameter impacting the length of the fade, and \( R_t \) the range of the marginal distribution of feature \( t \).

The result is shown in Figure 5A, which depicts the same data as in Figure 2b. Note that towards the end of the feature range, Figure 2b shows a third bump in feature contribution. This bump is not visible in Figure 5A with line fading. This shows that the effect was extrapolated from out-of-distribution data.

Additionally, the original data points can be shown to further enable the identification of out-of-distribution effects (Figure 4a).

Our implementation (shown in Figure 5) contains two views. The model view shows small multiples of GCV plots for all features (Figure 5A). This enables data scientists to determine which features are used by the model, and what values play an important role in
predictions. The y-axis is shared across all plots for easy comparison. Line fading can be customized by configuring the fading parameter \( \tau \) on-the-fly, and an option is provided to average all local polylines (similar to global PDPs). Selection is enabled by lasso brushing [17]: dragging a line in the plot will select all polylines intersecting that line, revealing clusters in the feature contribution vectors. This selection is linked to all other GCV plots as well as the data view.

The data view (Figure 5B) contains a list of histograms to show the original data distributions. The distribution of the selected instances in the model view is highlighted in blue. In the example, the data view shows that the selected cluster in the model view (for which the feature ‘pH’ = 3.05 is important to the predictions) corresponds with data instances with high alcohol content. The x-axis of the histograms can also be brushed to selected instances with specific feature values, and to highlight lines in the GCVs.

6 DISCUSSION AND FUTURE WORK

Our proposal can answer many questions about the model that remain challenging to answer with traditional techniques.

First, an expert can check feature contribution and the relationship with feature value at a single glance. In prior work this could only be done with separate visualizations of feature contribution and partial dependence based plots. We showed these are difficult to compare (it requires estimating the slope), and may not show consistent results, as they encode different information.

Next, patterns (or ‘strategies’) can be spotted that would otherwise remain hidden. For instance, in Figure 2b two distinct clusters of lines are visible. In addition, linking with the data view helps to ascertain what constitutes this strategy (e.g., alcohol contents).

Finally, our approach enables the validation of the (un)certainty of contribution through line fading. Comparing Figure 2b and Figure 5A shows the effect of the increase of contribution towards the end of the feature range was an artifact of out-of-distribution data.

However, the current implementation has a few limitations. First, much computation is needed to obtain these curves: our examples of the Wine Quality dataset took 5 minutes (on AMD Ryzen 5 3600X); it will take longer for larger datasets and more complex models. Hence, computing these plots on-the-fly is not possible. We address this by caching the results in our implementation. The optimization of current implementations of feature contribution methods for large datasets is an interesting direction for future research.

Next, even though we can visually represent many features there is a practical limitation on the number of features that can be shown. Hence the plots are best applicable to data sets with at most 10-20 features. In addition, for the plots to be interpretable we rely on a dataset that has features with inherent meaning.

Finally, we intend to follow up with an expert-based evaluation to confirm that the plots are beneficial in real-world use cases.

7 CONCLUSION

We have presented Local Contribution Value (LCV) plots, a novel way of conveying feature contribution as a function of feature values. This was previously only possible by combining multiple views, or by fallibly estimating the slope of partial dependence curves, which is challenging and subject to errors. Furthermore, we introduced Global Contribution Value (GCV) plots to show a comprehensive overview of the full model behavior. These plots are information dense and enable novel insights into a model. We have tackled uncertainty of the sensitivity analysis by interactively fading out lines, enabling the validation of patterns for real data, and empower an analysis workflow with linked views.

The proposed visualizations provide data scientists with an in-depth view of the role of a feature in predictions, and enable model diagnosis, refinement, decision support and justification use cases commonly driven by model interpretability [3].

ACKNOWLEDGMENTS

This work is part of the research programme Commit2Data, specifically the RATE Analytics project with project number 628.003.001, which is financed by the Dutch Research Council (NWO).
REFERENCES


SUPPLEMENTARY VIDEO

https://youtu.be/BRKthx-Wk7Q