Optimal Local Explainer Aggregation for Interpretable Prediction

Qiaomei Li[1], **Rachel Cummings** [2], **Yonatan Mintz** [1]

[1] Department of Industrial and Systems Engineering, University of Wisconsin, Madison
[2] Departments of Industrial Engineering and Operations Research and Computer Science, Columbia Univerisity
qli449@wisc.edu, rac2239@columbia.edu, ymintz@wisc.edu

Abstract

A key challenge for decision makers when incorporating black box machine learned models into practice is being able to understand the predictions provided by these models. One set of methods proposed to address this challenge is that of training surrogate *explainer models* which approximate how the more complex model is computing its predictions. Explainer methods are generally classified as either *local* or *global* explainers depending on what portion of the data space they are purported to explain. The improved coverage of global explainers usually comes at the expense of explainer *fidelity* (i.e., how well the explainer's predictions match that of the black box model). One way of trading off the advantages of both approaches is to aggregate several local explainers into a single explainer model with improved coverage. However, the problem of aggregating these local explainers is computationally challenging, and existing methods only use heuristics to form these aggregations.

In this paper, we propose a local explainer aggregation method which selects local explainers using non-convex optimization. In contrast to other heuristic methods, we use an integer optimization framework to combine local explainers into a near-global aggregate explainer. Our framework allows a decision-maker to directly tradeoff coverage and fidelity of the resulting aggregation through the parameters of the optimization problem. We also propose a novel local explainer algorithm based on information filtering. We evaluate our algorithmic framework on two healthcare datasets: the Parkinson's Progression Marker Initiative (PPMI) data set and a geriatric mobility dataset from the UCI machine learning repository. Our choice of these healthcare-related datasets is motivated by the anticipated need for explainable precision medicine. We find that our method outperforms existing local explainer aggregation methods in terms of both fidelity and coverage of classification. It also improves on fidelity over existing global explainer methods, particularly in multi-class settings, where state-of-the-art methods achieve 70% and ours achieves 90%.

1 Introduction

When applying machine learning and AI models in high risk and sensitive settings, one of the biggest challenges for decision makers is to rationalize the insights provided by the model. In applications such as precision medicine, both prediction accuracy (e.g., anticipated efficacy of treatment) and transparency of how predictions are made are key for obtaining informed consent. However, the models that typically achieve the highest levels of accuracy also tend to be extremely complex, and even machine learning experts describe them as "black boxes" because it is difficult to explain why certain predictions are made (Breiman 2001). One popular approach to resolve this trade off between explainability and accuracy is to extract simple *explainer* models from complex black box models. These models are intended to provide a simplified facsimile of the true model that is more useful for human interpretation of the generated predictions.

Two important widely-used metrics for evaluating explainer models are *fidelity* and *coverage*. Fidelity measures how well the explainer's predictions match the predictions of the original black box model, and coverage measure the fraction of the data universe that is reasonably explained by the explainer model. Explainer methods are generally classified as either *global* or *local*, based on how they trade off between these two quantities. Global explainers attempt to explain the full black box model across the entirety of the data. These models have a hard constraint to provide 100% coverage, often at the expense of fidelity. Local explainers, on the other hand, sacrifice coverage to potentially provide higher fidelity explanations in a smaller region of the data, usually centered around one single prediction.

Recent proposals suggest finding a middle ground between these two extremes by forming global (or near-global) explainers by aggregating local explainer models (Ribeiro, Singh, and Guestrin 2016). This approach would allow the decision-maker to trade off among coverage, fidelity, and explainability: including more local explainers in the aggregate model would improve coverage and fidelity, at the cost of a more complex—and hence less interpretable—aggregate model. However, the problem of computing the best subset of local explainers to explain a given black box model is combinatorial in nature, and hence computationally challenging to solve. All existing methods for building aggregate explainers use only heuristic approaches, and thus do not provide theoretical performance guarantees.

In this work, we present a novel way of constructing provably optimal aggregate explainer models from local explainers. We use an integer programming (IP) optimization

framework that trades off between coverage of the aggregate model and fidelity of the local explainers that comprise the aggregate model. We also propose a local explainer methodology that uses an information filter for feature selection, and is designed for use in aggregation. We empirically validate the performance of this framework in two healthcare applications: Parkinson's Disease progression and geriatric mobility. These experimental results show that our model provides higher fidelity than existing methods. In this application, a clinician would use a black box model for their initial diagnosis of a patient, and then use that patient's data in the particular local explainer selected by our algorithm to understand why the black box model made its prediction.

1.1 Related Work

Our paper builds on previous work in the broader field of interpretable machine learning. The two primary types of interpretable learning include models that are interpretable by design (Aswani et al. 2019), and black box models that can be explained using global explainer (Wang and Rudin 2015; Lakkaraju, Bach, and Leskovec 2016; Ustun and Rudin 2016; Bastani, Bastani, and Kim 2018) or local explainer (Ribeiro, Singh, and Guestrin 2016, 2018) methods.

Models that are interpretable by design are perhaps the gold standard for interpretable ML. However, these models often require significant domain knowledge to formulate and train, and are not suited for exploratory tasks such as the precision healthcare applications we study in Section 4.

Global explainer methodology attempts to train an explainable model (e.g., a decision tree with minimal branching) to match the predictions of a black box model across the entirety of its feature space. While these models provide some understanding on the general behavior of the black box model, if the relationship between features and black-box predictions is too complex, then the global explainer may remove subtleties that are vital for explanation.

Local explainer methods attempt to train simpler models centered around the prediction for a single data point. The most commonly used local explainer methods are Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) and anchors (Ribeiro, Singh, and Guestrin 2018). While local methods cannot validate the full black box model, they are useful for understanding the subtleties and justification for particular predictions. In recent literature several other local explainer methods have been proposed that draw inspiration from this stream (Rajapaksha, Bergmeir, and Buntine 2019; Sokol and Flach 2020; Plumb, Molitor, and Talwalkar 2018).

A third option which has been explored in recent literature, is that of aggregating several local explainer models to form a near-global explainer, as a method for improving the tradeoff between fidelity and coverage. Generally speaking, these methods have a budget for the maximum number of local explainers that can be incorporated into the aggregation and attempt to maximize possible coverage and fidelity within this budget. One method proposed to form such aggregate explainers is the submodular pick method (Ribeiro, Singh, and Guestrin 2016), which computes feature importance scores and greedily selects the features with highest importance. (van der Linden, Haned, and Kanoulas 2019) argue the Submodular Pick Algorithm has its limitations on predicting global behaviors from local explainers, and that the choice of aggregation function for local explainers is important for performance. They introduce the Global Aggregations of Local Explanations (GALE) method, which can be used to analyze how well the aggregation explains the model's global behavior. They compared the performance of global LIME aggregation with other global aggregation methods for binary and multi-class classification tasks, and found that different aggregation approaches performed best in binary and multi-class settings. A recently proposed aggregation method (called GLocalX) hierarchically combines local explainers to from global explainers (GLocalX) (Setzu et al. 2021). This methodology could be incorporated in to our optimization approach as pre-processing.

The methodology we propose in this paper builds on top of these existing explainer aggregation methods. In contrast to existing approaches which are heuristic in nature, we formulate the problem of choosing local explainers for the aggregate as an optimization problem. By doing so, our methods can produce explainer aggregates that provide both higher fidelity and higher coverage than existing approaches. In addition, our formulation includes parameters that allows for a direct tradeoff between coverage, fidelity, and interpretability. We believe this approach is especially appropriate for problems in explainable precision healthcare, where the relationship between diagnostic screening measures and the diagnosis is quite complex, and the model should incorporate the richness of this relationship in its predictions.

We propose a local explainer approach in Section 3 that includes a feature selection subroutine to improve explainability. Prior work on feature selection includes instance-wise feature selection (Chen et al. 2018) and Instance-wise Variable Selection using Neural Networks (INVASE) (Yoon, Jordon, and van der Schaar 2018). These approaches select the important features for each sample point using networks for classification with and without the features. Shapley values have also been used for complex model predictions, such as Shapley Sampling Values (Štrumbelj and Kononenko 2014; Aas, Jullum, and Løland 2019) and Shapley Additive Explanations (SHAP) (Lundberg and Lee 2017; Lundberg et al. 2020), which computes Shapley values and presents the explanation as an additive feature attribution method. In contrast to these methods, our feature selection approach relies on a mutual information filter (Brown et al. 2012) to identify important features. While mutual information has been used in the past for feature selection, we introduce a computationally efficient way to compute this mutual information for the use of training local explainer models.

1.2 Our Contributions

In this paper, we formulate the problem of aggregating local explainers into an aggregate explainer algorithm as a non-convex optimization problem. In particular, we show that this aggregation problem can be written as an integer program (IP), that can be solved effectively using commercial solvers. This formulation is also helpful as it allows us to directly tradeoff coverage and fidelity of the resulting aggre-

gation through parameters of the optimization problem. This approach provides flexibility to the practitioner to adapt the algorithm to the specific needs of her use case. Our approach easily handles multi-class prediction problems that arise in complex application domains such as precision healthcare, as well as the traditionally-studied binary classification.

Additionally, we design a new methodology for training local explainers for effective use in aggregation. Our local explainer algorithm directly computes locally significant features using an information filter, and we are the first to use information filters in local explainers. We introduce a novel computationally efficient algorithm for this filtering step, and our approach results in simpler (i.e., more interpretable) local explainers compared to prior work that used regularization for feature selection.

To validate our results, we compare our optimization based methodology against four other state of the art methods on two real world data sets. Both data sets come from the applications in the healthcare space. The first uses the Parkinson's Progression Marker Initiative (PPMI 2019), where we create explainer methods for a model tasked with screening patients for Parkinson's Disease. The second uses a dataset of Geriatric activity, where we explain the predictions of a model that classifies the physical activity of geriatric patients to prevent falling. Our experiments show that our optimization method outperforms many of the existing explainer methods in terms of fidelity and coverage. In particular, when we examine cases of explaining multi-class model predictions, our explainer method can achieves 90% fidelity at 40-50% coverage, while existing global methods only achieved 70% fidelity, albeit at 100% coverage. Our results show that our approach on the Pareto frontier of the fidelity and coverage tradeoff. Our IP framework outperforms existing aggregation methods in terms of both coverage and fidelity across all potential aggregation budgets (i.e., numbers of local explainers in the aggregate model).

2 Explainer Aggregation Methodology

Explainer models which can generalize to a large portion of the feature space are critical transparency. However, an explainer that is constrained to explain the entire feature space is likely have low fidelity since, by design, the explainer model is less complex than the black box model it is purported to explain. However, simpler models can achieve higher fidelity by attempting to explain the local behavior of the black box model at the cost of lower coverage.

One way to address the tradeoff between coverage and fidelity is to create a near-global aggregate explainer model by combining several local explainer models. Existing approaches have used this idea (Ribeiro, Singh, and Guestrin 2016) by formulating the construction of an aggregate explainer as an optimization problem: maximize coverage of the explainer subject to a constraint on the total number of local explainers included in the aggregate. Solving this optimization problem is conjectured to be computationally intractable (Ribeiro, Singh, and Guestrin 2016), and prior work has only attempted to solve it using heuristics.

In this section, we formulate the problem of constructing the aggregate explainer *from an arbitrary black-box model*, explicitly as an integer linear program that can be solved efficiently using commercial solvers, and allows us to directly trade off coverage and fidelity.

2.1 Mathematical Programming Formulation of Aggregation Problem

To formulate the optimization problem of constructing the aggregate explainer, we must first formally define the concepts of coverage and fidelity.

Let $\mathcal{X} \subset \mathbb{R}^m$ be the feature space that is modeled with a black box function, and let $f : \mathcal{X} \to \mathbb{Z}_+$ be the black box function of interest. Let $\mathcal{L} \subseteq \mathbb{Z}_+$ be the label space in the image of f. We consider our explanation task over a dataset $\mathcal{D}$ containing n ordered pairs (x_i, y_i) for $i \in [n]$, where $x_i \in \mathcal{X}$ are the feature values and $y_i \in \mathcal{L}$ is the class label which has been generated by f. That is, $y_i = f(x_i)$.

Let $g_{i,r} : \mathcal{X} \to \mathcal{L}$ denote a local explainer model that explains the local behavior of the black box function f on inputs within a ball of radius $r \in \mathbb{R}_+$ centered around the point $x_i \in \mathcal{X}$. We use $\mathcal{X}_{i,r} := \{x \in \mathcal{X} : \|x - x_i\| \leq r\}$ to denote the region explained by $g_{i,r}$.

Define an aggregate explainer γ to be a set of local explainers centered around a subset of points in $\mathcal{D}$, where the local explainer for point $x_i \in \mathcal{D}$ has radius r_i.[1] We will refer to a generic local explainer $g \in \gamma$ and corresponding region of explanation $\mathcal{X}_g$.

Using these quantities we define the *coverage of aggregate explainer* γ on data set $\mathcal{D}$ as the total number of points in the data set that are covered by the explanation radius of at least one explainer contained in γ. We denote this as:

$$\text{Cov}(\gamma, \mathcal{D}) = \sum_{x \in \mathcal{D}} \max_{i \in \{i : g_{i,r} \in \gamma\}} \mathbb{1}[x \in \mathcal{X}_{i,r}]. \qquad (1)$$

Next we note that the fidelity of a single local explainer is defined as the accuracy of that explainer with respect to the predicted labels of the black box model. We emphasize that fidelity captures the explainer's ability to replicate the predictions of the black-box model, and rather than ground truth predictive accuracy.

We define the *fidelity of aggregate explainer* γ on data set $\mathcal{D}$ as the minimum of the fidelity obtained by each individual local explainer in γ. We first need to define $\mathcal{D}_g$ as the number of points in the data set contained in the explanation region of g, i.e., $\mathcal{D}_g = \{x \in \mathcal{D} : x \in \mathcal{X}_g\}$. We denote this as:

$$\text{Fid}(\gamma, \mathcal{D}) = \min_{g \in \gamma} \frac{1}{|\mathcal{D}_g|} \sum_{x \in \mathcal{D}_g} \mathbb{1}[g(x) = f(x)]. \qquad (2)$$

While one could instead define the fidelity of γ as the average of the fidelities of its component explainers, our choice to use the minimum fidelity gives a stricter measure of how well the aggregate explainer captures the behavior of the black box model. This stricter measure is more appropriate for the healthcare applications we consider in Section 4,

[1] More generally, any local explainers can be aggregated into γ. However, we assume the the explainer algorithm only has access to points in $\mathcal{D}$, so we restrict ourselves to only considering these points. It is assumed that the radii r_i are parameters of the problem and hence known to decision-maker.

where a minimum standard of care is required. Note also that while we may be interested in the coverage and fidelity of γ across the entirety of $\mathcal{X}$, computing these quantities may be intractable or impossible in practice when $\mathcal{X}$ is not known a priori. Thus we consider these quantities only across an r-ball covering of our dataset.

Let K denote the budget of the maximum number of local explainers that can be contained in γ, and let φ be the minimum fidelity required for the aggregate explainer. Then the problem of computing an aggregate explainer can be formulated as the following optimization problem:

$$\max_{\gamma} \ \{\text{Cov}(\gamma, \mathcal{D}) : \text{Fid}(\gamma, \mathcal{D}) \geq \varphi, |\gamma| \leq K\}. \tag{3}$$

2.2 Reformulation as Integer Program (IP)

As written, optimization problem (3) is not trivial to solve, and could require enumerating all possible subsets γ of local explainers. To address this challenge, we propose reformulating problem (3) as an Integer Program (IP) that can be solved using commercial software. We first define three sets of binary variables w_i, y_j, z_{ij}. Let w_i be a binary variable that is equal to 1 if explainer $g_{i,r_i} \in \gamma$. That is, $w_i = \mathbb{1}[g_{i,r_i} \in \gamma]$. Let y_j be a binary variable that is equal to 1 if point j is covered by the aggregate explainer γ. That is $y_j = \mathbb{1}[x_j \in \cup_{g\in\gamma} \mathcal{X}_g]$. Finally, let z_{ij} be a binary variable that is equal to 1 if explainer $g_{i,r_i} \in \gamma$ covers point x_j. That is, $z_{ij} = \mathbb{1}[x_j \in \mathcal{X}_{i,r_i}]$. We now define the coverage and fidelity of aggregate explainer γ as IPs written in terms of these three sets of variables.

Proposition 1. *Cov$(\gamma, \mathcal{D})$, the coverage of aggregate explainer γ on dataset $\mathcal{D}$, can be expressed with the following set of integer variables and constraints:*

$$\begin{aligned} \textit{Cov}(\gamma, \mathcal{D}) = &\textstyle\sum_{j=1}^{n} y_j, \\ \textit{s.t.} \quad & z_{ij} \leq w_i, \quad i, j \in [n], \\ & y_j \geq z_{ij}, \quad i, j \in [n], \\ & y_j \leq \textstyle\sum_{i=1}^{n} z_{ij}, \quad j \in [n], \\ & \|x_i - x_j\| z_{ij} \leq r_i, \quad i, j \in [n]. \end{aligned} \tag{4}$$

Proof. Recall the definition of $\text{Cov}(\gamma, \mathcal{D})$ as given in Equation (1). We will directly reconstruct this definition using the binary variables defined above. First note that through a simple direct substitution we obtain $\text{Cov}(\gamma, \mathcal{D}) = \sum_{x\in\mathcal{D}} \max_{i\in\{i:g_{i,r}\in\gamma\}} z_{ij}$. Since taking the maximum of binary variables is equivalent to the Boolean OR operator, we see that $y_j = \max_{i\in\{i:g_{i,r}\in\gamma\}} z_{ij}$, which provides us with the first equality. The next two inequalities directly capture that a local explainer g_{i,r_i} can only explain point x_j if g_{i,r_i} is included in γ, which is a standard way of modeling conditional logic in IP (Wolsey and Nemhauser 1999). The next two constraints come from modeling the Boolean OR operator using integer constraints (Wolsey and Nemhauser 1999). The final constraint ensures that a point x_j can only be covered by an explainer g_{i,r_i} if $x_j \in \mathcal{X}_{i,r_i}$, thus preserving the local region of the local explainer. □

Next we consider the minimum fidelity constraint.

Proposition 2. *The constraint Fid$(\gamma, \mathcal{D}) \geq \varphi$ can be modeled using the following set of integer linear constraints:*

$$\begin{aligned} &\|x_i - x_j\| z_{ij} \leq r_i, \quad i, j \in [n], \\ &z_{ij} \leq w_i, \quad i, j \in [n], \\ &\textstyle\sum_{j=1}^{n} (\mathbb{1}_{\{f(x_j)=g_{i,r_i}(x_j)\}} - \varphi) z_{ij} \geq 0, \quad i \in [n]. \end{aligned} \tag{5}$$

While the full proof of Proposition 2 is deferred to Appendix B, we note that the first two constraints ensure proper local behavior of the local explainer as in Proposition 1. The third constraint is derived by analysis of the definition of $\text{Fid}(\gamma, \mathcal{D})$ in Equation (2), dis-aggregating the lower bound constraint across all $i \in [n]$, and re-writing the new lower-bound constraint to remove the $\min$ using properties of z_{ij}.

We can then use these expressions to for coverage and fidelity to re-write our optimization problem as an integer program that can then be solved using commercial solvers.

Proposition 3. *The optimization problem in* (3)*,*

$$\max_{\gamma} \ \{\textit{Cov}(\gamma, \mathcal{D}) : \textit{Fid}(\gamma, \mathcal{D}) \geq \varphi, |\gamma| \leq K\},$$

can be written as the following integer program:

$$\begin{aligned} &\max \textstyle\sum_{j=1}^{n} y_j, \\ \textit{s.t.} \quad & z_{ij} \leq w_i, y_j, \quad i, j \in [n], \\ & y_j \leq \textstyle\sum_{i\in\mathcal{X}} z_{ij}, \quad j \in [n], \\ & \|x_i - x_j\| z_{ij} \leq r_i, \quad i, j \in [n], \\ & \textstyle\sum_{j=1}^{n} (\mathbb{1}_{\{f(x_j)=g_{i,r_i}(x_j)\}} - \varphi) z_{ij} \geq 0, \quad i \in [n], \\ & \textstyle\sum_{i\in\mathcal{X}} w_i \leq K, \\ & y_j, w_i, z_{ij} \in \{0, 1\} \quad i, j \in [n]. \end{aligned} \tag{6}$$

Proof. The objective function and first four constraints come directly from Propositions 1 and 2. The next constraint comes using the definition of w_i and direct substitution to obtain that $|\gamma| = \sum_{i\in=1}^{n} w_i$, which is then used to rewrite the budget constraint from (3). The final constraint ensures that our new variables are binary integers. □

3 Aggregate-Designed Efficient Local Explainer

While our main contribution in this paper is the local explainer aggregation methodology, we have additionally designed a new methodology for training local explainers for effective use in aggregation. The key to our methodology is ensuring that local explainers only focus on the most relevant features in the particular region they are designed to explain. In contrast to previous methods that proposed the use of regularization to achieve this goal, we propose directly computing locally significant features using an information filter. Computing such filters are generally computationally expensive and requires the use of numerical integration; however, we introduce an efficient algorithm for filtering out less significant features. This methodology allows us to train local explainers that are significantly less complex than those that use regularization, with better fidelity for their specified region. In this section we present an

overview of our methodology and highlight key results. Further details on the technical specifics of this methodology are deferred to the appendix.

3.1 Local Explainer Overview and Training Procedure

Our local explainer training methodology is formally presented in Algorithm 1. We give a brief overview of its operations here, and defer full details to Appendix D. The algorithm takes in hyper-parameters including number of points N to be sampled for training the explainer, a distance metric d, and a radius r around the point $\bar{x}$ being explained. First the algorithm samples N points uniformly from within a r radius of $\bar{x}$; we call this set of points $T(\bar{x})$. Depending on the distance metric being used this can often be done quite efficiently, especially if the features are binary valued or an ℓ^p metric is used (Barthe et al. 2005). Then using the sampled points, the algorithm uses the Fast Forward Feature Selection (FFFS) algorithm as a subroutine (discussed and formally presented in Appendices G), which uses a mutual-information-based information filter to remove unnecessary features and reduce the complexity of the explainer model. The FFFS algorithm uses an estimate of the joint empirical distribution of $(T(\bar{x}), f(T(\bar{x}))$ to select the most important features for explaining the model's predictions in the given neighborhood using tree traversal. We denote this set of features $\hat{\Phi}$. Then, using these features and the selected points, the local explainer model g is trained by minimizing an appropriate loss function that attempts to match its predictions to those of the black box model. In principle, a regularization term can be added to the training loss of explainer g. However, in our empirical experiments (presented in Appendix E), we found that FFFS typically selected at most five features, so even the unregularized models where not overly complex.

Algorithm 1: Local Explainer Training Algorithm

Require: sampling radius r, number of sample points N, black box model f, data point to be explained $\bar{x}$, and loss function L for the explainer model $(\bar{x}, \bar{y})$

Initialize $T(\bar{x}) = \emptyset$
for $j = \{1, ..., N\}$ **do**
  Sample $x \sim U(\mathcal{B}(\bar{x}, r, d))$
  $T(\bar{x}) \leftarrow T(\bar{x}) \cup x$
end for
Obtain $\hat{\Phi}(\bar{x}) = \text{FFFS}(T(\bar{x}), \Phi, f)$
Train $g = \arg\min_{\hat{g} \in \mathcal{G}} \{\sum_{x \in T(\bar{x})} L(f(x) - \hat{g}(x[\hat{\Phi}]))\}$
return g =0

4 Experimental Results

In this section we compare the performance of our IP method against five state-of-the-art explainer methods. We consider two local explainer aggregation methods—Submodular Pick and Anchor Points (Ribeiro, Singh, and Guestrin 2016, 2018)—and three global explainer methodologies—interpretable decision sets (Lakkaraju, Bach, and Leskovec 2016), active learning decision trees (Bastani, Bastani, and Kim 2018), and naive decision tree global explainers (Friedman, Hastie, and Tibshirani 2001).

We compare these methods in both coverage and fidelity across two different datasets. These datasets are the Parkinson's Progression Marker Initiative (PPMI) data set, where we generate explainers for a black box model aimed at predicting Parkinson's Disease (PD) progression subtypes, and a Geriatric activity data set (Torres et al. 2013) where we generate explainers for a model that classifies the movement activities of geriatric patients based on wearable sensor data. We split each dataset, using 80% for training and 20% as a holdout test set, and we apply 10-fold cross validation. One important feature of both these datasets is that they enable multi-class classification. Our experimental results show that our proposed optimization framework is better suited to these multi-class settings than existing state-of-the-art methods.

In addition to measuring the performance of our local aggregation methodology on different data sets and classification tasks, we also compare the performance of our information-filter-based decision-tree local explainer and LIME (Ribeiro, Singh, and Guestrin 2016) in the aggregation framework. We also measure performance for each of the aggregation-based methods under varying budgets of component local explainers. This budget is an informal measure of simplicity and interpretability, where aggregating fewer local explainers leads to a more interpretable aggregate explainer, but may sacrifice fidelity and/or coverage. Our results show that our methodology outperforms existing techniques in terms of fidelity and coverage, especially in the multi-class case.

4.1 PD Progression Cluster Classification

For our first set of experiments we used the PPMI data set to classify the disease progression of different patients into several subtypes based on screening measures. The PPMI study was a long run observational clinical study designed to verify progression markers for PD. To achieve this aim, the study collected data from multiple sites and includes lab test data, imaging data, and genetic data, among other potentially relevant features for tracking PD progression. The study includes measurements of all these features for the participants across 8 years at regularly scheduled follow up appointments. The complete data set contains information on 779 patients, and included 548 patients diagnosed with PD or some other kind of Parkinsonism and 231 healthy individuals as a control group. For our analysis we will focus on the first seven visits of this study which correspond to a span of approximately 21 study months, since these visits were conducted relatively close together temporally.

The classification task considered was the disease progression of the patients, and we performed a cluster analysis to generate labels, detailed in Appendix C. Our analysis identified four different subtypes of disease progression, corresponding to different trajectories of the diagnostic measurements' evolution over time. We also included one additional subtype corresponding to patients who did not have PD. Appendix C presents a full description of these subtypes

and their identification in the data.

As our black box model, we trained a random forest model to predict the progression subtype of a patient based on measurements taken during the baseline appointment and follow ups. We considered two different prediction tasks: (1) a binary prediction task to predict whether or not an individual has PD; (2) a multi-class prediction task to predict one of the five identified PD progression subtypes. Further details on the construction of the black box model and its performance on these tasks are given Appendix E.

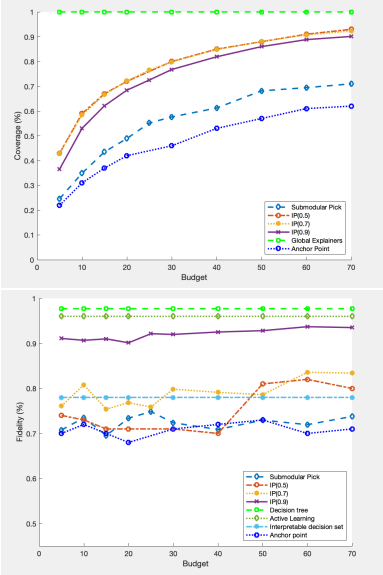

Figure 1: 2-class fidelity (bottom) and coverage (top) plots for various global explainers for a random forest model trained on the PPMI data set. The x-axis corresponds to the number of constituent local explainers that are used by the aggregation methods.

We used each of the explainer methods presented above to explain the predictions made by these random forest models, and measured coverage and fidelity of these explainers. Coverage and fidelity for the binary prediction task are shown in Figure 1, and similar plots for the multi-class prediction task are shown in Figure 2.

Figures 1 and 2 show that for both prediction tasks, our optimization-based aggregation algorithm obtains a higher level of coverage than both Anchor points (Ribeiro, Singh, and Guestrin 2018) and Submodular Pick methods (Ribeiro, Singh, and Guestrin 2016) across all possible local explainer budgets. Note that when comparing coverage, global explainers are constrained to always achieve 100% coverage.

In terms of fidelity, Figure 1 shows that across fidelity lower bounds of 0.7 and 0.5, our methodology performs comparably with the other aggregate explainer methods and with the explainable decision set method. When increasing our fidelity lower bound to 0.9, our method significantly outperforms these methods. This shows that the fidelity lower bound parameter φ in our framework allows for higher fidelity explainers given proper tuning.

In the binary case our methodology does not outperform active learning and naive decision tree in terms of fidelity or

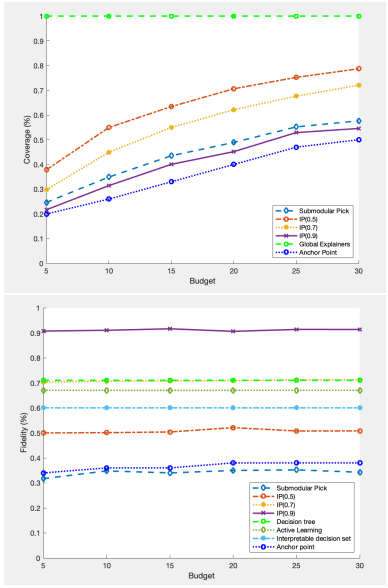

Figure 2: 5-class fidelity (bottom) and coverage (top) plots for various global explainers for a random forest model trained on the PPMI data set. The x-axis corresponds to the number of constituent local explainers that are used by the aggregation methods.

coverage; however, when considering the multi-class setting of Figure 2, we see that our framework allows for significantly higher fidelity explanations. In particular, while active learning and naive decision trees achieve a fidelity of approximately 0.7 our optimization based global classifier with $\varphi = 0.9$ can achieve a fidelity of 0.9 in this case. While this is a significant increase, it does come with a cost for the coverage, as the explainer with this high fidelity only covers 40–50% of the data, as compared to the global explainer methods of active learning and naive decision tree which cover 100% of the data.

Our methodology allows for greater flexibility in terms of trading off explainer coverage and fidelity, especially in this multi-class case. In contrast, the pure global explainer methods do not allow for this trade-off by ensuring a hard constraint of 100% coverage, which results in low fidelity explainers. Since our methodology outperforms existing aggregation methods, this indicates that using IP allows us to navigate the fidelity and coverage tradeoff more efficiently.

Empirical evaluation of our local explainer's performance compared with other local explainer methods, when used in the aggregate explainer are given in Appendix F. We find that our local explainer methodology outperforms LIME in both fidelity and coverage.

Figure 3 shows the Pareto frontier of the tradeoff between coverage and fidelity for the binary class prediction task. One advantage of our approach is that we allow a tunable tradeoff between the coverage and fidelity—corresponding to the three curves in the figure—while the other methods do not provide this option—corresponding to only a single point for the other methods. We see that our approach yields higher fidelity and higher coverage than most of the other

local explainers, although there is less of a clear advantage of our proposed method compared to the global explainers. Figure 3 shows the tradeoff between the coverage and fidelity for the multi-class setting. In this setting we again see that our proposed local explainer provides better coverage and higher fidelity than other local explainers. In addition, our method also provides significantly higher fidelity than the global explainers at the expense of coverage.

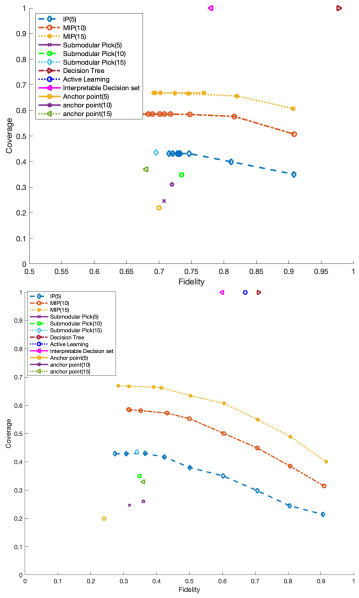

Figure 3: Pareto frontier for the tradeoff between fidelity and coverage on PPMI for binary (top) and multiclass (bottom, 5 class) classification task. The x-axis corresponds to fidelity and the y-axis corresponds to coverage.

4.2 Geriatric Activity Classification

For the second set of experiments we used a data set of Geriatric Activity based on the study conducted by (Torres et al. 2013). The main goal of this study was to provide ways of potentially reducing the likelihood of falls for geriatric individuals by classifying their activities when transferring beds. Generally, the highest risk for geriatric patients to fall is when getting out of bed so various sensors were deployed to detect whether an individual was attempting to leave their bed and detect other potentially risky activity. For this particular study, the authors used a novel wearable and environmental sensor which they validated with 14 individuals aged 66–86. The goal was to use this sensor data to classify between three different activities, namely laying in bed, sitting in the bed, and getting out of the bed. To generate the data set, each of the participants was asked to perform a random set of five activities which ranged between the three potential activity classes.

Much like in the case of the PPMI data set, we trained a random forest model to classify between the various activity classes that we used to extract global explainers. However, unlike the PPMI experiments, since there was no straight forward way to convert the multiclass classification task of detecting the different activities into a binary classification task we only performed the experiments for the multiclass case. The results for all explainer methods can be seen in Figure 4. Much like in the case for the PPMI data set, we note our methodology outperforms other aggregation based global explainers with respect to coverage across all budgets and fidelity lower bounds; however, it is still not obtaining 100% coverage like the pure global explainer methodologies. In terms of fidelity, much like in the multiclass case of the PPMI data, our methodology outperforms all other global explainers, with active learning being close to on par with our performance. This further suggests that using this form of optimization based local explainer aggregation is well suited to explaining multiclass predictions regardless of the underlying data set. Figure 5 shows the Pareto frontier of the tradeoff between coverage and fidelity across different explainers. Our methodology outperforms all local explainers in both metrics and all global explainers in terms of fidelity at the cost of decreased coverage.

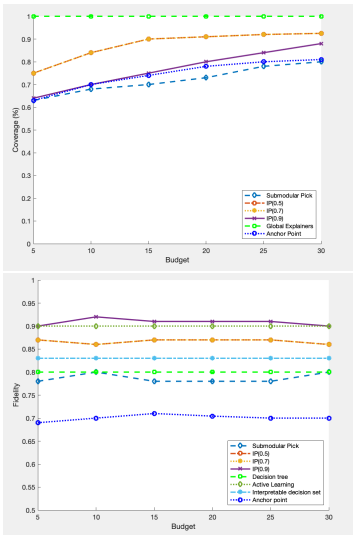

Figure 4: 5-class fidelity (bottom) and coverage (top) plots for various global explainers for a random forest model trained on the Geriatric Activity Dataset. The x-axis corresponds to the number of constituent local explainers used by the aggregation methods.

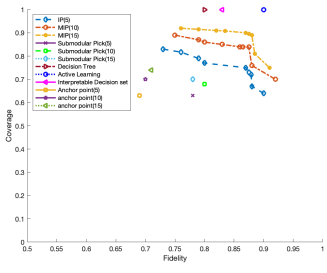

Figure 5: Pareto frontier for the tradeoff between fidelity and coverage on Geriatric Activity Dataset. The x-axis corresponds to fidelity and the y-axis corresponds to coverage.

References

Aas, K.; Jullum, M.; and Løland, A. 2019. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv preprint arXiv:1903.10464*.

Aswani, A.; Kaminsky, P.; Mintz, Y.; Flowers, E.; and Fukuoka, Y. 2019. Behavioral modeling in weight loss interventions. *European Journal of Operational Research*, 272(3): 1058–1072.

Barthe, F.; Guédon, O.; Mendelson, S.; and Naor, A. 2005. A probabilistic approach to the geometry of the ℓ_p^n-ball. *The Annals of Probability*, 33(2): 480–513.

Bastani, H.; Bastani, O.; and Kim, C. 2018. Interpreting predictive models for human-in-the-loop analytics. *arXiv preprint 1705.08504*.

Bhat, S.; Acharya, U. R.; Hagiwara, Y.; Dadmehr, N.; and Adeli, H. 2018. Parkinson's disease: Cause factors, measurable indicators, and early diagnosis. *Computers in Biology and Medicine*, 102: 234–241.

Breiman, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3): 199–231.

Brown, G.; Pocock, A.; Zhao, M.-J.; and Luján, M. 2012. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(Jan): 27–66.

Chen, J.; Song, L.; Wainwright, M. J.; and Jordan, M. I. 2018. Learning to explain: An information-theoretic perspective on model interpretation. *arXiv preprint arXiv:1802.07814*.

Fereshtehnejad, S.-M.; and Postuma, R. B. 2017. Subtypes of Parkinson's disease: What do they tell us about disease progression? *Current neurology and neuroscience reports*, 17(4): 34.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1675–1684.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1): 56–67.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov): 2579–2605.

Marek, K.; Jennings, D.; Lasch, S.; Siderowf, A.; Tanner, C.; Simuni, T.; Coffey, C.; Kieburtz, K.; Flagg, E.; Chowdhury, S.; et al. 2011. The Parkinson progression marker initiative (PPMI). *Progress in neurobiology*, 95(4): 629–635.

Martínez-Martín, P.; Gil-Nagel, A.; Gracia, L. M.; Gómez, J. B.; Martinez-Sarries, J.; Bermejo, F.; and Group, C. M. 1994. Unified Parkinson's disease rating scale characteristics and structure. *Movement Disorders*, 9(1): 76–83.

Martinez-Martin, P.; Rodriguez-Blazquez, C.; and Forjaz, M. J. 2017. *Rating scales in movement disorders*. Elsevier.

MATLAB. 2010. *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc.

Nasreddine, Z. S.; Phillips, N. A.; Bédirian, V.; Charbonneau, S.; Whitehead, V.; Collin, I.; Cummings, J. L.; and Chertkow, H. 2005. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4): 695–699.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; and Dubourg, V. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct): 2825–2830.

Plumb, G.; Molitor, D.; and Talwalkar, A. S. 2018. Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems*, 2515–2524.

PPMI. 2019. Parkinson's Progression Markers Initiative.

Rajapaksha, D.; Bergmeir, C.; and Buntine, W. 2019. LoRMIkA: Local Rule-based Model Interpretability with k-optimal Associations. *arXiv preprint arXiv:1908.03840*.

Rao, S. S.; Hofmann, L. A.; and Shakil, A. 2006. Parkinson's disease: Diagnosis and treatment. *American Family Physician*, 74(12): 2046–2054.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI '18, 1527–1535.

Setzu, M.; Guidotti, R.; Monreale, A.; Turini, F.; Pedreschi, D.; and Giannotti, F. 2021. GLocalX-From Local to Global Explanations of Black Box AI Models. *Artificial Intelligence*, 294: 103457.

Siderowf, A. 2010. *Schwab and England activities of daily living scale*, 99–100. Elsevier.

Sokol, K.; and Flach, P. 2020. LIMEtree: Interactively Customisable Explanations Based on Local Surrogate Multi-output Regression Trees. *arXiv preprint arXiv:2005.01427*.

Štrumbelj, E.; and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3): 647–665.

Torres, R. L. S.; Ranasinghe, D. C.; Shi, Q.; and Sample, A. P. 2013. Sensor enabled wearable RFID technology for mitigating the risk of falls near beds. In *2013 IEEE International Conference on RFID (RFID)*, 191–198. IEEE.

Ustun, B.; and Rudin, C. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3): 349–391.

van der Linden, I.; Haned, H.; and Kanoulas, E. 2019. Global aggregations of local explanations for black box models. *arXiv preprint arXiv:1907.03039*.

Wang, F.; and Rudin, C. 2015. Falling rule lists. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, AISTATS '15, 1013–1022.

Wolsey, L. A.; and Nemhauser, G. L. 1999. *Integer and Combinatorial Optimization*. Wiley.

Yoon, J.; Jordon, J.; and van der Schaar, M. 2018. INVASE: Instance-wise variable selection using neural networks.