GAPS

Generality and Precision with Shapley Attribution

Brian Daley
Computer Science Dept.
Columbia University
New York City, U.S.
brian.daley@columbia.edu

Qudrat E Alahy Ratul
Computer Science Dept.
Boise State University
Boise, United States
qudratealahyratu@u.boisestate.edu

Edoardo Serra
Computer Science Dept.
Boise State University
Boise, United States
edoardoserra@boisestate.edu

Alfredo Cuzzocrea
iDEA Lab
University of Calabria
Rende, Calabria, Italy
alfredo.cuzzocrea@unical.it

Purpose

- Machine-learning can still be somewhat of a "black-box"
- Must understand decision-making process to facilitate trust
  - Undesirable techniques in classification
- Dire consequences in major decisions
- Want to be able to explain ML models' classifications without compromising performance
- As ML use expands, Explainable AI (XAI) fields grow to harbor trust
- Examples include U.S. Defense Advanced Research Projects Agency (DARPA)
- Other countries such as the UK, France, and Portugal following suit
- EU statements of the importance of ML understandability

(a) Husky classified as wolf (b) Explanation

Image Source: Ribeiro 2016

Motivation

- One way to achieve explainability is through local model-agnostic methods
- Local interpretation methods seek to explain individual instances
  - Generate importance values for features
- Two of the most famous examples are LIME and SHAP
- A recent experiment proved that these models have low generality and precision scores
  - Previously only usable for rule-based explanation models
- Goal for this research is to create a local model-agnostic explainability model that improves these evaluation metrics
- Important for explanations to be accurate to maintain transparency of machine learning to humans

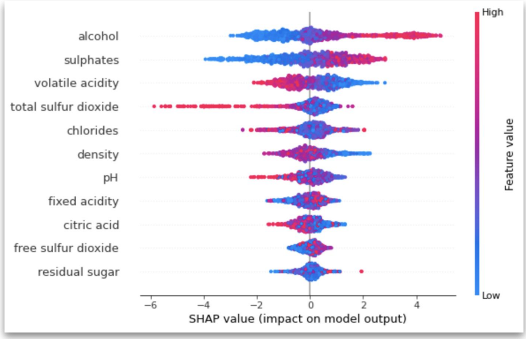
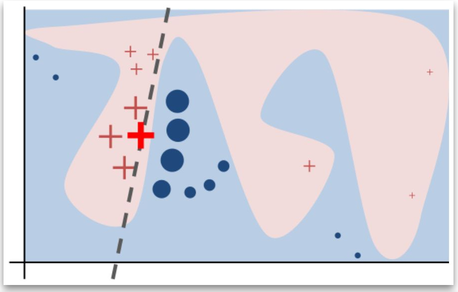


Image Source: Radečić 2020

BOISE STATE UNIVERSITY

Existing Attribution Methods

- LIME modifies the feature values of an instance slightly and observes changes in classification
  - Perform local sensitivity analysis based on small perturbations
  - Generates neighbors and weights based on distance
  - LIME creates simple linear abstractions close to the instance
- SHAP determines feature contributions using coalitional game theory
  - Generate coalition of whether or not a feature is present
  - Mean of all feature values if not present in coalition
  - Payouts are generated and fitted to a linear model
  - KernelSHAP utilizes random set of samples, improving runtime



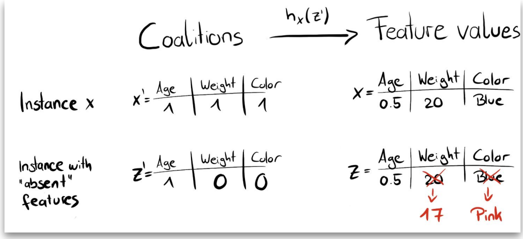


Image Source: Molnar 2022

Evaluation Metrics

- Generality and precision are used in rule-based explanations for evaluation
  - Precision: A rule with one classification should not have the opposite classification with the same rule
  - Generality: A rule with one classification should explain other instances of the same class
- Reverse precision measures the percent of instances with the same top features of an instance of the opposite class
- Generality measures the number of top features in common with an instance belonging to the top neighbor instances

$$avgRP^k(I_a) = \frac{\sum_{x \in I_a} RP^k(x, att_x)}{|I_a|}$$

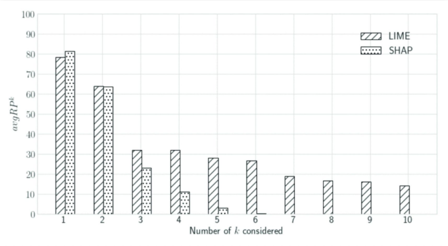
$$RP^k(x, att_x) = \frac{|\{\hat{x} | \hat{x} \in I_{\neg a}, sel(S^x_{at}, x) = sel(S^x_{at}, \hat{x})\}|}{|I_{\neg a}|}$$

$$agg(\{common_k(att_x, att_{\hat{x}}) | \hat{x} \in topNeighbour_h(x, I_a)\})$$

$$common_k(att_{x_1}, att_{x_2}) = |top_k(att_1) \cap top_k(att_2)|$$

GAPS

- Generality and Precision Shapley Attribution
  - Goal: Increase the precision and generality scores of LIME and SHAP
- Generate coalitions of present features with randomly generated binary vector
  - Present features unchanged, perturbed features on the normal curve with the mean being the feature value & SD from all values from feature
- Input neighbors into model that was used for classification
  - In this experiment, we use a Random Forest Classifier
- Neighbors of the same class belong to N(x,s,a) and opposite class belong to N(x,s,¬a)



$$f(s,x) = \left[\begin{array}{c} E_{l\sim m(s,x)}[c(l)] + \\ + \sum_{z\in N(x,s,a)} \frac{\lambda_G c(z)}{|N(x,s,a)|} + \sum_{z\in N(x,s,\neg a)} \frac{\lambda_P (c(z)-1)}{|N(x,s,\neg a)|} \end{array} \right]$$

Image Source: Ratul 2021

GAPS Cont'd

$$\mu(S) = \frac{|F|-1}{\binom{|F|}{|S|}|S|(|F|-|S|)}$$

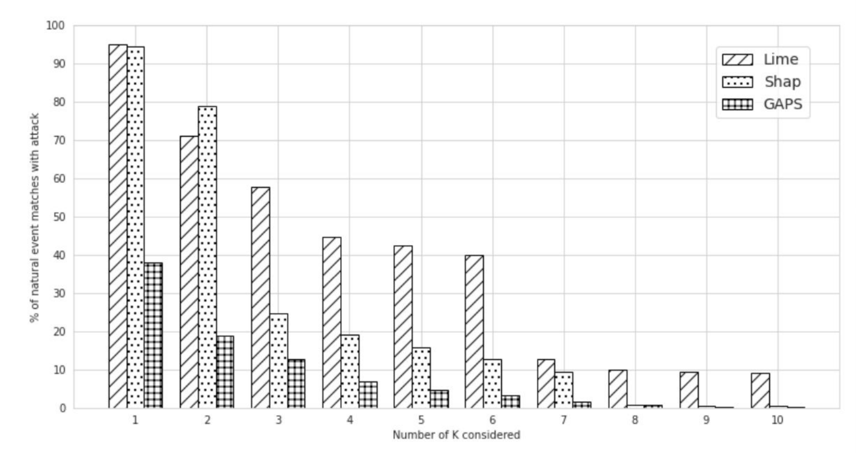
- λ_G and λ_P scale prevalence of generality and precision scores
- Find the confidence of the classification for each neighbor *z*
- Sum confidence of each neighbor times coefficient over the number of neighbors belonging to each class
- Add the expected value of the confidence of many randomly generated neighbors from the coalition
  - Features not in the coalition are from a randomly selected feature
- Like LIME and SHAP, pass the coalitions and rewards into a linear model with Kernel from KernelSHAP
- Coefficients from the linear model are then treated as the importance values

$$f(s,x) = \left[\begin{array}{c} E_{l \sim m(s,x)}[c(l)] + \\ + \sum_{z \in N(x,s,a)} \frac{\lambda_G c(z)}{|N(x,s,a)|} + \sum_{z \in N(x,s,\neg a)} \frac{\lambda_P (c(z)-1)}{|N(x,s,\neg a)|} \end{array} \right]$$

Image Source: Ratul 2021

Experimental Findings

- Used “UNSW-NB15” dataset which measures raw network packet data
- Classified as real normal network behavior and synthetic attacks
- Lower reverse precision scores, higher generality scores than LIME and SHAP
- The GAPS attribution methods better fit precision and generality evaluations

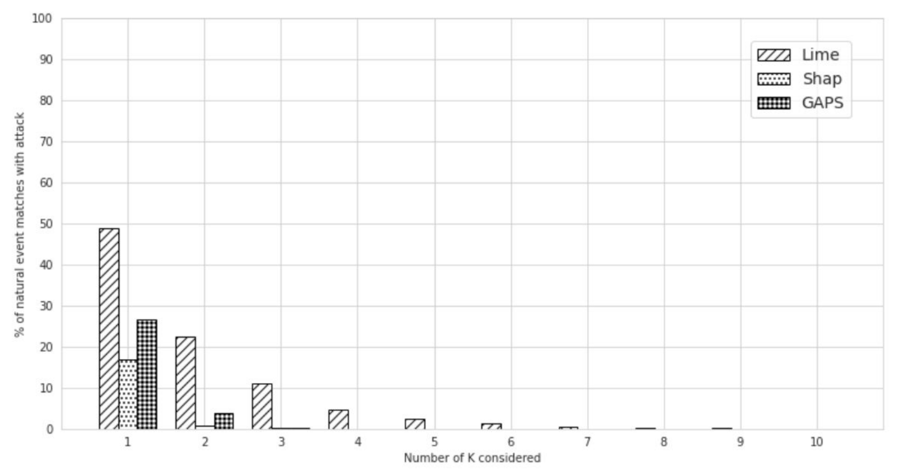


		Mean GAPS intersection size		
		Max	**Mean**	**Min**
No of Neighbors (h)	**No of Features (k)**			
1	**1**	1.00	0.68	0.00
	5	5.00	3.91	0.00
	10	10.00	8.27	1.00
5	**1**	1.00	0.65	0.00
	5	5.00	3.70	0.00
	10	10.00	7.95	1.00
10	**1**	1.00	0.62	0.00
	5	5.00	3.61	0.30
	10	10.00	7.82	1.30

		Mean LIME intersection size			Mean SHAP intersection size		
		Max	**Mean**	**Min**	**Max**	**Mean**	**Min**
No of Neighbors (h)	**No of Features (k)**						
1	**1**	1.00	0.37	0.00	1.00	0.84	0.00
	5	5.00	1.83	0.00	5.00	4.11	0.00
	10	10.00	5.46	2.00	10.00	8.64	3.00
5	**1**	0.20	0.01	0.00	1.00	0.84	0.00
	5	3.20	1.81	0.80	5.00	3.91	0.80
	10	7.40	5.44	4.00	10.00	8.52	3.00
10	**1**	0.20	0.01	0.00	1.00	0.84	0.00
	5	3.10	1.81	0.50	5.00	3.83	0.90
	10	7.20	5.44	2.50	10.00	8.42	4.20

Experimental Findings Cont'd

- Used "ICS: Power System" dataset which measures power system disturbance
- Also classified normal network behavior and network attacks
- Higher performance metrics than LIME in some settings and lower than SHAP



		Mean GAPS intersection size		
		Max	**Mean**	**Min**
No of Neighbors (h)	**No of Features (k)**			
1	**1**	1.00	0.15	0.00
	5	4.00	0.76	0.00
	10	7.00	1.82	0.00
5	**1**	0.80	0.16	0.00
	5	2.00	0.78	0.00
	10	4.00	1.87	0.20
10	**1**	0.60	0.16	0.00
	5	1.90	0.79	0.00
	10	3.70	1.86	0.30

		Mean LIME intersection size			Mean SHAP intersection size		
		Max	**Mean**	**Min**	**Max**	**Mean**	**Min**
No of Neighbors (h)	**No of Features (k)**						
1	**1**	1.00	0.01	0.00	1.00	0.39	0.00
	5	2.00	0.22	0.00	5.00	2.09	0.00
	10	4.00	0.83	0.00	10.00	4.41	0.00
5	**1**	0.20	0.01	0.00	1.00	0.34	0.00
	5	1.00	0.22	0.00	4.20	1.69	0.00
	10	2.20	0.85	0.00	9.20	3.65	0.40
10	**1**	0.10	0.01	0.00	0.90	0.32	0.00
	5	0.80	0.21	0.00	3.70	1.57	0.00
	10	1.90	0.87	0.20	8.40	3.36	0.40

Conclusions

- Explainability is an enormously important aspect of machine learning
  - Trust between humans and machines is vital
  - Avoid undesirable techniques in classification
- Algorithms to explain a single instance exist such as LIME and SHAP
  - Poor precision and generality scores upon examination
- GAPS works to increase these performance metrics without compromising accuracy
- GAPS shows great promise as an attribution method for local model-agnostic explainability as it had higher generality and precision than LIME and SHAP
  - Further research is necessary, as in another dataset, “ICS: Power System,” GAPS outperformed LIME, but not SHAP, possibly due to adjusting weighting coefficients

Thank you for your time!

Questions/Feedback?

edoardoserra@boisestate.edu