GUILHERME HENRIQUE RESENDE DE ANDRADE

INVESTIGATION ON THE USE OF QUANTITATIVE INFORMATION FLOW FOR EXPLANATIONS OF MACHINE LEARNING MODELS

Scientific monograph presented as a requirement for concluding the Undergraduation in Information Systems of the Federal University of Minas Gerais.

Advisor: Flavio Vinicius Diniz de Figueiredo

Belo Horizonte

October 2020

To my family.

Acknowledgments

Since the beginning of the undergraduate program, my life has changed drastically. Every moment of joy, sadness, anguish, doubt, and confidence led me to make decisions that brought me where I am now. So, before anything else, I express my wonder about the effects of randomness in our lives.

I also want to thank my family which, even though sometimes in strange ways, helped me overcome every barrier in order to achieve my goals. I know they will always rejoice with my achievements, and for that I am grateful. This conquest is just a small way for me to give something in return for everything you have done for me. Love you.

Next, I want to pay my gratitude to all my friends and my colleague from "Mestres dos Métodos". You are among the most important things that university gave me. Without you, this whole process would be way heavier than it was. Every conversation at the canteen, aisles, and CRC helped to cheer me up and project me toward my goals.

To my teachers, I express my complete gratitude and admiration. Especially, to Gisele Lobo Pappa, Fabrício Murai, Mario Sérgio Alvim, Heitor Ramos, Flávio Figueiredo, André Hora, Rodrygo Santos, Vinicius Santos, and Pedro Olmo who still do their jobs with dedication, good-will, and empathy.

Lastly, but not least, I would like to thank every friend who has passed but filled my life with memories and experiences, and also every person that once was part of my life. Unfortunately, none of you will ever read this, but I know you all contributed to this to happen.

"The scenes in our life resemble pictures in a rough mosaic; they are ineffective from close up, and have to be viewed from a distance if they are to seem beautiful"
(Schopenhauer, Arthur)

Abstract

In this work, we investigate the use of the framework Quantitative Information Flow (QIF) in explaining predictions of machine learning models. This project follows a quantitative analysis contrasting QIF and Shapley Additive Explanations (SHAP) importance coefficients as a way to encourage or discourage QIF's usage.

Palavras-chave: Information Theory; Game Theory; Machine Learning; Interpretability; Quantitative Information Flow; Shapley Additive Explanations.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Since the advent of modern computers, humans have been trying to automate tasks with what nowadays we call Machine Learning approaches. In the beginning, due to computation and memory constraints, the proposed solutions - as well as their outputs - were simple and intuitive to human cognition (Rosenblatt [1958]; Cover and Hart [1967]). However, the growth in computation and memory over the years leveraged more complex and computationally expensive solutions which, consequently, led to considerable damage in model interpretability (Cortes and Vapnik [1995]; Hochreiter and Schmidhuber [1997]; Chen and Guestrin [2016]).

Machine Learning (ML) models are built over a fundamental resource, that is, data. Considering data is, essentially, a result or description of a process, and is influenced by humans either directly or indirectly, it is extremely prone to carry many of the human biases. As shown in Figure 1.1, the environment is strictly influenced by humans, and both humans and the environment act directly in the generation of data.

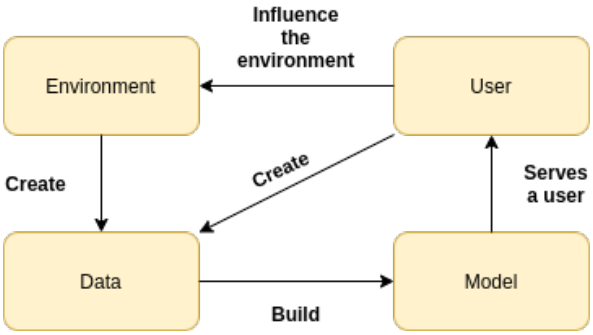

Figure 1.1: Data generation cycle

In contemporary society, it is not unusual to witness/know about discrimination and racism cases (Yong [2018]; Kleeman [2019]; Hutchinson et al. [2020]). The

motivations of such events range from financial to geographic issues. This kind of human behavior explicit the real need to become aware of what extent the data used in predictive models are biased towards negative/misleading patterns.

For example, financial institutions often establish the level of interest rates depending directly on a person's characteristics (e.g earning and indebtedness patterns, assets, salary, and so on). After a series of analyses, the institution decides whether or not to grant the loan to the client. Nevertheless, there might be a misleading consensus among the institution employees towards the reproval of loan grants for people from a certain low-income neighborhood. When building a predictive model to automate the concession task, the discriminative behavior is probably embodied by the model, resulting in loan denials based not on financial issues, but on geographic issues (not directly related to whether or not someone will pay a borrowing).

In the previous example, some might argue that the problem could be solved by simply removing geographic features. Such an approach could substantially reduce the bias in the data, and consequently, in the model. However, oftentimes features present a considerable level of mutual information with the removed variables and, hence, serve as proxies to the deleted properties. Therefore, we must first comprehend the decision process of the models and then work towards stronger approaches not based solely on features removal.

Grounded on the importance of explanation approaches, this work aims to investigate the use of a method taken from the realm of Information Theory, namely, Quantitative Information Flow (QIF) (Denning [1982]; Gray III [1992]), to explain the predictions of machine learning models. Our investigation intends to generate importance coefficients based on the amount of information flowing from the input features to the model predicted output. The calculated coefficients will then be empirically validated by being contrasted with those generated by a consolidated explanation approach based on a Game Theory method, Shapley Additive Explanations (SHAP) (Lundberg and Lee [2017]).

Chapter 2

Related Works

The conjecture and development of this work rely majorly on two approaches. The first one, Shapley Additive Explanations (SHAP) inspired on a Game Theory framework, and the second one, Quantitative Information Flow (QIF), based on an Information Theory framework.

2.1 Shapley Additive Explanations (SHAP)

Usually, the best way to explain a model is the model itself. However, oftentimes models' complexities become unsustainably high, forbidding its comprehension. Aiming at such scenarios, SHAP tries to develop an additive auxiliary model to mimic the performance of the original solution. The auxiliary model, in this case, is a linear model which considers the Shapley Values as its multiplicative coefficients.

2.1.1 Shapley Values

In Game Theory, a wide area dedicated to the study of strategic actions players can perform to increase their chance of profit/earnings, it is often necessary to know what is the expected value or return a given player can get from a specific situation. In Coalitional (also known as Cooperative) games, the figure is no different, however, the value or return of a player can change considerably according to the group in which he/she is inserted.

In real-world scenarios, the interest resides in finite cooperative groups. In this work, we are specifically more interested in games that agree with the following properties:

- The group aggregates a finite number of players;

- the game is cooperative, i.e agreements are supposed and encouraged to happen;
- there is a medium of exchange that flows freely and in unlimited amounts from one player to another;
- the game is accurately described by a characteristic function.

Considering the aforementioned statements, we can define a coalition as a group S comprising N players. The worth of the entire group is defined by a characteristic function v that maps a group of players S to its respective maximal total payoff $v(S) \in \mathbf{R}$.

The definition of a characteristic function allows us to search for the best coalition. Aiming to calculate the worth of an entire coalition, as well as the contribution from each one of its players, Lloyd F. Shapley proposed an approach based on the following assumptions:

- Equal Treatment: if one player can be replaced by another without impact in the worth of the coalition, then they must have the same assigned value;
- Null Player: a player which causes no changes in the worth of the coalition;
- Pareto Optimality: the sum of the values of all players must be equal to the maximal amount they can jointly get;
- Additivity: the value of the sum of two games, is the sum of the values of the two games.

The definitions above ensure a unique payoff vector of values named Shapley Values (Shapley [1953]). In the resulting solution, the order in which players appear does not imply any change in the final worth of the coalition. Hence, each one of the possible permutations of players has the same value.

As described by the Pareto Optimality, the worth of a player for a cooperative group equals its marginal contribution to the very group. Based on the previous premise, we calculate the worth of a player for a given game with the following equation:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \tag{2.1}$$

where S is a coalition not comprising i, and N the set of all players. The Equation (2.1) can be interpreted as the average contribution of a player i over all the possible coalitions from N.

2.2 Quantitative Information Flow (QIF)

Information Theory is a vast area interested in the specificities of topics such as storage, measurements, and communications. When developing information systems, developers and institutions take security as a valuable resource. However, the usual path followed by them is to patch leakages and vulnerabilities as they appear on the course of development, implementation, or production. Although the previous approach is capable of solving the problems, it does not prevents leakage, nor guarantees that there are no more possible vulnerabilities in the system at hand.

There are many ways in which information can be uncovered and stolen by malicious users, one of them is through information flow. Essentially, information flow can be defined as the transfer of information from a source (sender) to a target (receiver). When dealing with real-world problems, leakage is an inevitable fate and, instead of trying to avoid it by all means, a preferable path is to quantify the information flow and its respective possible damages.

Quantitative Information Flow is a framework based on measurements from a secret's leakage. The measurements - which can be made with a series of metrics - are performed before and after a system is run, and the difference between the two values (a priori, and a posteriori) is said to be the amount of information that flowed from the system being analyzed.

Since QIF is capable of measuring the amount of information flowing from a system, and the importance of each variable to the leaked quantity is also known, we are able to decide whether or not the leakage is acceptable, and what can be done to reduce or restrain the flow.

Chapter 3

Datasets

Considering the quantitative experimental characteristic of this work, we have chosen a series of well-known datasets currently available at Dua and Graff [2017]. Initially, for simplicity's sake, we opted for binary classification problems, or ones easily mapped to a binary output. We have also considered a multiclass image classification problem LeCun and Cortes [2010] as a way to present a visual human-intuitive interpretation for the explanation.

3.1 Adult

The data was extracted by Barry Becker from the 1994 Census database. It has a binary target variable telling whether or not a given person makes more than US$ 50.000 a year. When positive, we assign the value 1.

3.2 Bank Marketing

This dataset records information related to a Portuguese financial institution marketing campaign. It presents a binary target variable which is positive(1) whenever a client signed for a term deposit.

3.3 Breast Cancer

The records present cell properties drawn from breast mass image exams. It has a binary target variable which tells if a given diagnosis regards a benign or malignant

tumor. To what concerns modeling, the presence of a malignant tumor is assigned to the value 1.

3.4 Heart Disease

This dataset comprises data from four sources, namely, Hungarian Institute of Cardiology, University Hospital from Zurich, University Hospital from Basel, and V.A. Medical Center from Long Beach and Cleveland Clinic Foundation. Every record describes a patient's characteristics, and the binary target variable points to the presence - represented by the value 1 - of heart disease or not.

3.5 Wine Quality

This dataset presents both red and white variants of a Portuguese wine brand. The wine quality is graded between [0-10] and is considered to be good (1) if it has a score bigger than 6, otherwise, the target variable is set to 0.

3.6 MNIST

This dataset regards a multiclass problem and consists of many handwritten, size-normalized, and centralized digits. Every record represents a digit in [0-9], and every feature is a pixel.

Chapter 4

Experiments

The experiments were performed on a personal computer with a Unix based operating system. The machine has a 1 TB Hard Drive, together with 128 GB SSD, 16 GB RAM, Intel Core i7, and a GeForce MX150 graphic card. All experiments shown below can be replicated by setting the Numpy random seed to 1.

4.1 Input Modelling

Considering all aforementioned datasets consists of both continuous and categorical variables, and the categorical values can be either binary or multi-valued, we assume the following modeling standard:

- Continuous: values were discretized in intervals (bins) with increasing importance comprising a series of records.
- Binary: the unique values are sorted in ascending order, and then are assigned to 0 and 1, respectively.
- Multi-Valued
  - Implicit Hierarchy: the values were left unchanged.
  - No Hierarchy: we considered transforming each one of the values to a column and then placing a 1 whether the record contains the property, and 0 otherwise.

4.2 Experimental Setup

When dealing with machine learning problems we must - ideally - choose a parameter combination that leads to the best performing model for the dataset in hand. However, seeking the best model considering pre-established train and test sets can potentially lead to misleading conclusions due to data overfitting, i.e. we could be actually optimizing our test set performance instead of our generalization ability.

As a way to reduce the chance of overfitting, we consider performing a K-Fold-1-Out-Cross-Validation over the training set as a way to find the best model. The experiments consider the average performance of XGBoost (Chen and Guestrin [2016]) over K folds when varying the **maximal depth** and the **number of estimators** according to Table 5.1. The model with the best performance is considered to be the final model. In the case of a tie, the model with the simplest combination of parameters is chosen.

	Values
Maximal Depth	2, 4, 8, 16
Number of Estimators	16, 32, 64, 128, 256

Table 4.1: Set of values for each parameter. All combinations of the values above are tested.

Chapter 5

Analysis

In this chapter, we present analyses of QIF and SHAP outputs regarding the best model predictions for each dataset. We develop analysis concerning metrics, correlation, execution time, visual intuition for image explanations.

The main difference between the approaches lies in the use scenarios. Since QIF measures the amount of information flowing from a feature to the model output, the explanation is directly related to the class being analyzed, instead of individual records. Hence, QIF shows, in reality, the importance of each feature in the prediction of a given class, whereas SHAP gives explanations for predictions of single records. As a way to compare the explanation coefficients from both models, we get the average explanation vector from all outputs generated by SHAP for a given class.

5.1 Metrics

In real-world problems, before start developing a solution, we must define which metric best fits our necessities. What we expect with this, is to make sure that the approximation built by the model truly reflects the data generator function.

We defined two metrics to be optimized by our model, namely, F1 and the Area Under ROC Curve (AUC). As shown in Table 5.1, all models have reached considerably good performances in all six datasets. Hence, we can expect models to be pretty confident in their outputs, fairly approximating the data generator function.

	Parameters		Metric	
Datasets	Maximum Depth	Number of Estimators	F1-Score	AUC
Adult	4	128	0.63	0.89
Bank Marketing	4	128	0.44	0.90
Breast Cancer	2	128	0.95	0.99
Heart Disease	2	64	0.85	0.89
Wine Quality	8	128	0.81	0.82
MNIST	8	128	0.97	-

Table 5.1: Parameter combination that results in the best performance, and its respective F1 and AUC scores. Since AUC is not defined for multiclass problems, it cannot be applied to MNIST. Note that both F1 and AUC scores are the average over 5 folds in cross validation.

5.2 Correlation

As cited in Chapter 1, our purpose was to perform comparisons between the coefficients generated by QIF and SHAP in order to validate the use of the framework Quantitative Information Flow to explain machine learning models. We compared the coefficients outputted for each dataset with two correlation approaches, i.e. Kendall and Pearson. The results are presented in Table 5.2.

Despite the correlation coefficients not being negligible in some cases, the methods do not seem to have a strong correlation among themselves. The results are not yet conclusive, therefore, more experimentations - quantitative and qualitative - may be required to statistically and practically reject the use of QIF in such scenarios.

	Correlation	
Datasets	Kendall	Pearson
Adult	-0.40	-0.79
Bank Marketing	-0.21	0.02
Breast Cancer	-0.22	0.14
Heart Disease	0.10	0.11
Wine Quality	0.38	0.71
MNIST	0.39	0.15

Table 5.2: Kendall and Pearson correlations of QIF and SHAP coefficients for each dataset. Since MNIST is a multiclass problem, its coefficients were calculated as the average value over the classes' coefficients.

5.3 Execution Time

When proposing a new approach, it is interesting to be aware of properties such as time complexity. Our QIF implementation[1] applies the calculation of the joint distribution between a feature $X[j]$ and the corresponding target variable Y. Roughly speaking, the joint distribution is calculated by building an index from the Cartesian product between the unique values from the feature in hand ($U_{X[j]}$), and the target variable (U_Y), what results in a $O(\|U_{X[j]}\|\|U_Y\|)$ time complexity, and counting the co-occurrences over all records with a cost of $O(N \log N)$, where N is the number of records in the data. Nonetheless, it is not impossible to have a scenario where $\|U_{X[j]}\| = \|U_Y\| = N$, hence, the time complexity is asymptotically dominated by $O(N^2)$. Considering this calculation is done for each feature in X, the final time complexity is $O(MN^2)$, where M is the number of features.

As we can see in Table 5.3, our approach performs poorly in comparison to SHAP in every dataset. The worst time performance is with MNIST, due to its relatively high dimensionality.

	Time	
Datasets	SHAP	QIF
Adult	1min 10s	3min 1s
Bank Marketing	58.4 s	1min 34s
Breast Cancer	74.2 ms	358 ms
Heart Disease	440 ms	1.36 s
Wine Quality	4.47 s	4.83 s
MNIST	7min 53s	34min 58s

Table 5.3: Execution time for each method per dataset.

5.4 Image Explanations

When interpreting machine learning models' outputs, we often are incapable of telling whether or not the explanations make sense. However, for some cases, there is a way to confirm the correctness of the approach. Those cases are often related to images and texts, which are intuitive to the human cognition.

As aforementioned, the main difference between SHAP and QIF is that the first approach gives explanations to a single record, whereas the last gives explanations for the whole class. In Figure 5.1 we are able to see the explanations given be QIF to

[1] https://github.com/FelipeGiori/qif

each target class showing that the method seems to be able to fairly outline the main characteristics from each digit.

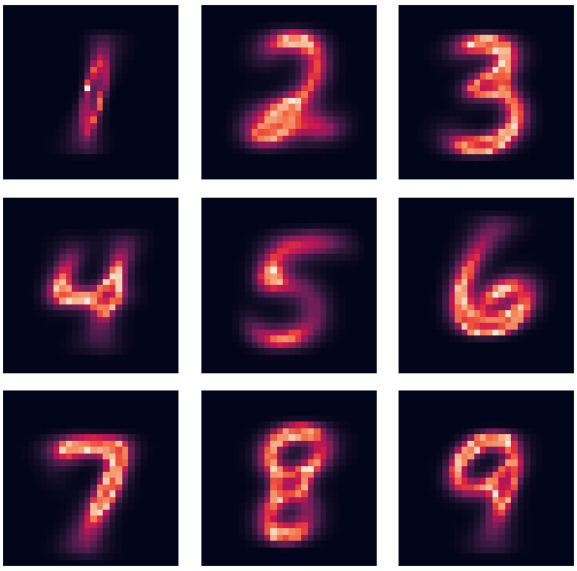

Figure 5.1: QIF coefficients by class.

On the other hand, in Figure 5.2 we can see the SHAP explanation given to a single record. Unlike QIF explanations, which present the overall properties from the classes, with SHAP it is possible to see specificities from the input, such as personal handwriting style.

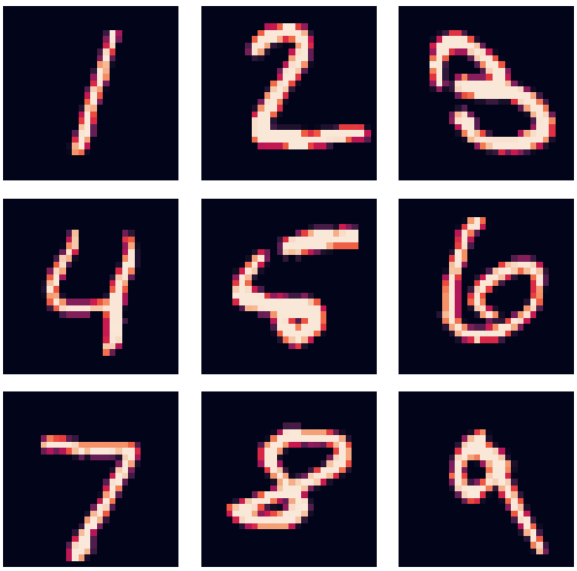

Figure 5.2: SHAP coefficients from a unique record by class

Even though QIF explanations made sense when applied to MNIST dataset, they are probably not useful if applied to different scenarios where the important characteristics are not always placed in the same spot (set of pixels), e.g. cat images where the location of the cat can vary and the final intention is to classify the image between a cat and not a cat.

Chapter 6

Conclusion

In this work, we scrutinized the use of the framework Quantitative Information Flow when applied to understand the decision process of machine learning models. As shown in Chapter 6, the correlation between SHAP and QIF coefficients on the datasets presented - although not negligible - does not clearly encourage the use of QIF to explain models' predictions. One possible reason can be the fact that SHAP focus on single explanations, whereas QIF performs explanations from the global perspective (what consequently results in a small correlation value).

We also explored characteristics such as execution time. With those experiments we were able to see that QIF's executions always lasted longer than SHAP's, increasing the discrepancy as the number of features increase. Since QIF theoretically adds a quadratic complexity for every new feature, it will have problems dealing with high dimensional data.

When applied to images, QIF's explanations appeared to be assertive in pointing out the important pixels for each class. Nonetheless, considering QIF relies on the distribution of the values of each pixel, its explanations would only make sense when applied to cases where the important items for predicting a class are placed at the same locality. For example, consider the problem of classifying whether or not an image contains a cat. If for the vast majority of the data, the cats are placed in the center of the image, and the model takes that as an important component for the prediction, the coefficients outputted will comprise that confidence. On the contrary, if the cats are placed in many locations, the model will pay attention to different parts of the image at different moments, and the final coefficients will point out to the intersection of pixels that the model considered the most.

The results found here do not invalidate the use of Quantitative Information Flow to explain models' decisions, however, they also do not encourage its use. Given the

natural disposition of QIF to explain the overall importance of variables for each class, we left as future work the qualitative investigation of those scenarios, as well as more quantitative experiments in order to reach a statistical validation or invalidation of the framework's use for sucha a purpose.

Bibliography

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785--794.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273--297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21--27.

Denning, D. E. R. (1982). *Cryptography and data security*, volume 112. Addison-Wesley Reading.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Gray III, J. W. (1992). Toward a mathematical foundation for information flow security. *Journal of Computer Security*, 1(3-4):255--294.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735--1780.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

Kleeman, J. (2019). Snl producer and film-maker are latest to accuse youtube of anti-lgbt bias.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

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

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307--317.

Yong, E. (2018). A popular algorithm is no better at predicting crimes than random people.