AFRL-AFOSR-UK-TR-2019-0011

Who's behind these predictions? Reconciling transparency and privacy in machine learning

Maria Jose Ramirez Quintana
UNIVERSIDAD POLITECNICA DE VALENCIA
CAMINO VERA 14
VALENCIA, 46020
ES

02/14/2019
Final Report

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
14-02-2019	Final	15 Sep 2017 to 14 Sep 2018

4. TITLE AND SUBTITLE
Who's behind these predictions? Reconciling transparency and privacy in machine learning

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-17-1-0287

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Maria Jose Ramirez Quintana

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSIDAD POLITECNICA DE VALENCIA
CAMINO VERA 14
VALENCIA, 46020 ES

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
EOARD
Unit 4515
APO AE 09421-4515

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-UK-TR-2019-0011

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The PI completed this project investigating the tradeoff between external query and response access to machine learning (ML) models while preserving security and confidentiality. The objective was to use queries and responses from an unknown ML model to discover the model family (neural net, decision tree, support vector machine, genarlized ilnear model, etc) of the unknown model. This represents a security risk as attackers can leverage known aspects of a ML model to inject adversarial examples during training and/or model deployment. The main idea was to train surrogate models based on the inputs and outputs of the unknown model. The PI then employed a dissimilarity measure (as specified in the attached final report) to determine which surrogate model best matched input-output data from the unknown model. The results show such an approach has potential as they were able to correctly identify model families much better than random chance, but there remains much room for further investigation. More specifics may be found in the attached final report and the technical papers reference therein.

15. SUBJECT TERMS
machine learning, reverse engineering, model family, characterization, active learning, item response theory, membership queries

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			PETERSON, JESSE
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER *(Include area code)* 314 235 6292

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Final Performance Report

FA9550-17-1-0287

Who's behind these predictions? Reconciling transparency and privacy in machine learning

Dr. María José Ramírez Quintana

Period of Performance:
15 SEP 2017 - 14 NOV 2018

Contents

Problem Statement

The increasing ubiquity of machine learning (ML) models in devices, applications,and assistants, which replace or complement human decision making, is prompting users and other interested parties to model what these ML models are able to do, where they fail, and whether they are vulnerable. On many occasions, we can only interact with the ML models by querying them through a public interface deployed with the aim to preserve model confidentiality. But a general question arises: Can we characterize the tension between external query access and confidentiality in machine learning models while preserving the transparency and intelligibility for the internal trustful users? In this project we study the extraction of some model characteristics from black-box models using queried input-output pairs.

Summary

In this report we summarize the research carried out along the developing of the project. More concretely, we have been working on three research issues: to characterize how ML models could be modelled taking the interests of users into account, to identify the family of black box ML models, and to generate a new hierarchy of machine learning techniques according to the similarity of the behaviour of the models constructed by them. The two first topics were developed during the first year of the project, whereas the last one was performed during the two-months project extension granted for this purpose. Related to the first topic, we have been exploring some scenarios for (partially) capturing machine learning model behaviour taking into account the aspect in that the users are interested in. For the second research topic, we have developed two approaches based on machine learning to identify the family of black box models. Basically the idea is to generate queries that are answered by the model (which acts as an oracle) and then to use these examples as the training dataset (we denote as the surrogate dataset) for learning different surrogate machine learning models. Then properly evaluating the behaviour of the surrogate models is enough to identify the oracle's family. Given that our method for family identification is based on the behaviour of the models, in the third research topic we have been investigating the construction of new families of machine learning techniques based on the kappa metric.

List of Participants

- Raul Fabra Boluda
- Dr. Cesar Ferri Ramírez
- Dr. José Hernández Orallo
- Dr. Fernando Martínez Plumed
- Dr. María José Ramírez Quintana (PI)

1 Introduction

Machine Learning (ML) is being increasingly used in confidential and security-sensitive applications deployed with publicly-accessible query interfaces, e.g., FICO or credit score models, health, car or life insurance application models, IoT Systems Security, medical diagnoses, facial recognition systems, etc. However, because of these public interfaces, an attacker can query the model with special chosen inputs, get the results and learn how the model works from these input-output pairs –using ML techniques. This corresponds to the typical adversarial machine learning problem. In an attack scenario, the attacker can take advantage of the knowledge of the type of learning technique (the ML family) the attacked model was derived from (and, in some cases, the true data distribution used to induce it) in order to explore intrinsic flaws and vulnerabilities. However, on many occasions, we can only interact with the ML models by querying them as a black box since they may have been generated by a third party or may be too complex to understand.

One of the main reasons for not having *general* techniques for exploiting black-box models may be due the intrinsic differences between ML techniques: different models constructed using different ML techniques might disagree not only on decision boundaries but also on how they extrapolate on areas with little or no training examples. Figure 1 shows how models differ on boundaries and extrapolation. On the top-left corner we see the class distribution for a Hypothetical2-dimensional data set (original data) with 4 class labels. As can be seen data concentrates on the top and the right hand side of the plot and it presents an empty area. We have used this dataset to learn 7 different models: a decision tree, a Naïve Bayes and a Support Vector Machine (with a radial basis kernel) (on the top raw), and a logistic regression model, a KNN model with k=11, a neural network and an ensemble (Random Forest). More or less all the models behave well in the dense zones but extrapolates very differently in the empty area. We may say that these less dense zones are those more likely to contain vulnerabilities. Hence, for many attacks it is more important to know what the model looks like than its full semantics.

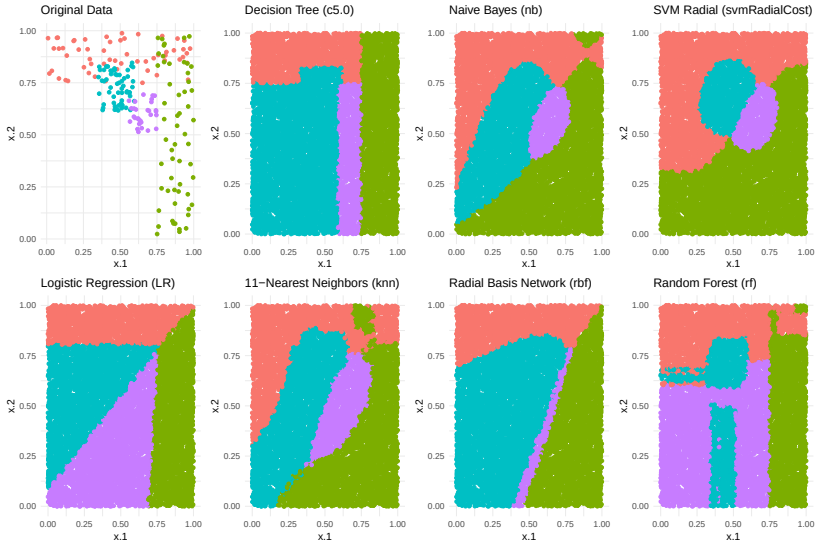

Figure 1: Behaviour of different models learned over the same set of data (shown on the top-left plot). From left to right and from top to bottom: a decision tree, Naïve Bayes, a SVM, logistic regression, 11-Nearest Neighbour, Neural Network, and Random Forest.

2 Methods, Assumptions and Procedures

In this section, we present the main concepts of our approaches for modelling ML models, identifying ML model family, and for defining ML techniques families based on model behaviour.

2.1 Towards Contextual Abstractions for Modelling ML Models

Under the view that ML models are complex behavioural systems, the understanding of how a model behaves can be done looking at different aspects: its intended vs actual behaviour,the relation between inputs and outputs, and understanding the whole model or some of its decisions. In [2] we claim that a more contextual abstraction is necessary in order to describe the behaviour of the model in simpler, more comprehensible, or more understandable terms according to a given context. Given a model M, the basic idea is to build a theory or model A able to explain part of the behaviour of M (the part established or defined by the context).

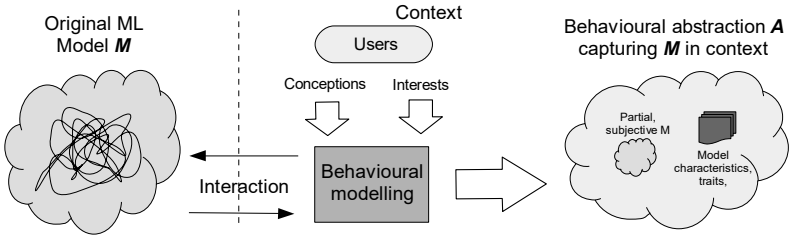

Figure 2: A given model M (black box) that some potential users want to understand better. These users want an abstraction that is customized to their context: their conception of the domain and their interests. This model of the model (the abstraction A) captures behaviour that is very partially, subjective to the context, usually in terms of a set of characteristics and traits.

Figure 2 shows how an abstraction A is built from an original model M according to a certain context, which is given by our interests and conceptions. The procedure for building A must go beyond the definition of a sophisticated utility function that is followed by an optimization process on the model M. We claim that building A requires a re-modelling process that must include regularization terms, the abstract model representation (i.e., features and combinations of those features), and always giving priority to those features that are easier to understand, fairer, leading to more stable models and better calibration. The context (conceptions and interests) must be the drive for abstraction. Only in this way can one ignore the irrelevant (uninteresting) details for the model, thereby overhauling the notion of overfitting. In [2] we present some illustrative examples of how modelling ML models could be done for different applications.

2.2 Model Family Identification

In [4] we present two methods for the identification of the family of a black box ML model (assuming that neither the model nor the original data used to train the model are available). Given a black box model O, the starting point is to generate input examples (queries) to be labelled by O (that acts as an oracle). From this dataset SD (surrogate dataset), we can learn different models A_i (surrogate models), as showing in Figure 3.

As we are interested in identifying the family of the oracle, we can use the Cohen's kappa coefficient to measure the agreement among the oracle and the surrogate models. Our first method for family identification consists on assigning to the oracle the family of the surrogate model that offers the highest kappa. It is

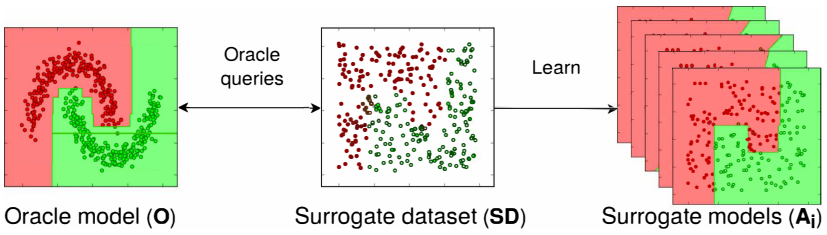

Figure 3: Black box models (oracles), trained over an unknown original dataset, are used to label synthetic surrogate datasets (generated following specific query strategies), which are then used to train surrogate models.

based on the following rational: both models achieve a high degree of agreement to explain SD probably because they belong to the same family.

In order to identify such as surrogate model we proceed as follows: given an oracle model O, we evaluate (with the kappa measure) its associated surrogate dataset SD with different models A_i, so that each A_i belonging to different model families, and then we select as output the family of the surrogate model with highest kappa value. Analternative

Our second approach consists in learning a meta-model for predicting the family of an oracle from a collection of meta-features (based on the κ value of the surrogate models) that abstractly describe the oracle. More concretely, given an oracle O belonging to a learning family $y \in N$, we represent it as the tuple $\langle \kappa_1(SD), \kappa_2(SD), \ldots, \kappa_N(SD) \rangle$, where $\kappa_i(SD)$ is the kappa value for the surrogate model A_i and dataset SD. If we collect M datasets $D_1, D_2, \ldots, D_M$, for each one we learn N oracles O_j (an oracle per family), and for each pair O_j and D_i we generate a surrogate dataset SD_i, we can build a new dataset where each row represents one oracle and dataset, each column is a kappa measure of the surrogate models and the output Y is the model family.

$$\begin{gathered} \langle \kappa_1(SD_1), \kappa_2(SD_1), \ldots, \kappa_N(SD_1) \rangle \\ \vdots \\ \langle \kappa_1(SD_M), \kappa_2(SD_M), \ldots, \kappa_N(SD_M) \rangle \end{gathered}$$

To validate our proposals, we have performed some experiments (a detailed description can be found at [4]). For the experiments, we have considered 11 learning families extracted from the categorization of learning techniques proposed in [1]. They include all the well-known learning techniques such as ensembles, decision trees, support vector machines, etc. In order to generate a dataset of meta-features that is large enough to evaluate our second approach, we employed $D = 25$ dataset. For each dataset, we trained $N = 11$ models (oracles) belonging to the different families introduced above (we learned $D \times N = 25 \times 11 = 275$ oracle models). For each of these oracles, we generated a surrogate dataset SD that we use to learn and evaluate the surrogate models A_i, belonging to the same N model families. Every example of SD was generated at random following the uniform distribution.

2.3 A Behavioural Definition of Family of Machine Learning Techniques

Regarding the selection of techniques to generate the surrogate models, we have focused on the machine learning families proposed in the literature [1, 5]. Thus, for the experiments performed in the previous section, we have considered a set of ML techniques grouped in terms of their learning strategy. As our approaches for family identification are based on the use of the kappa statistic, We present in this section an extensive experimental evaluation we carried out to establish a novel taxonomy of learning families based on the inter-rater agreement of different learning techniques.

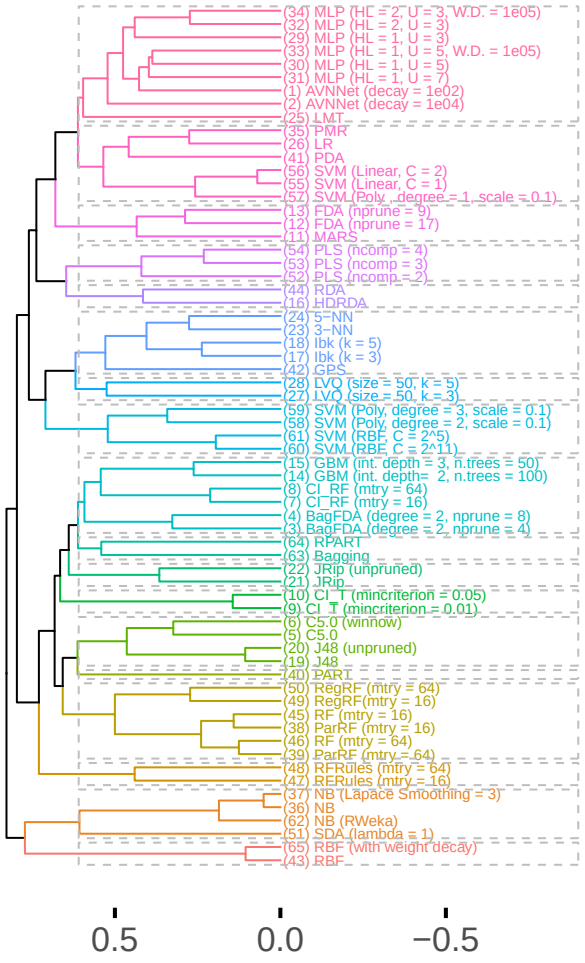

Figure 4: Dendrogram constructed using 64 datasets and 65 ML techniques, applying a hierarchical clustering algorithm using kappa values as the linking criterion.

The taxonomy was constructed applying a hierarchical clustering algorithm using a version of a distance matrix in that the cells contains kappa values (that is the values in the matrix belong to the interval $[-1..1]$). More concretely, for the experiments we used 64 datasets from different well-known dataset repositories, and 65 learning techniques available in the CARET[1] and RWeka[2] packages. For each dataset we created an artificial test set to be labelled by the different learned models. In order to cover all the domain space (dense and sparse zones) we added the training set to the test set. The distance matrix M (that is used by the hierarchical algorithm to group the techniques) is based on kappa. Thus, M is a triangular matrix of size $L \times L$, where L is the number of learning techniques (that is 65). Each cell M_{ij} represents the average kappa (for all datasets) obtained by comparing the outputs of models i and j. Figure 4 shows the dendrogram generated by the hierarchical clustering algorithm. The tree structure allows us to explore different groupings of the learning techniques by cutting the tree at a different linking values (showed in the X-axis). For instance, in this figure the groups generated when the dendrogram is cut at a kappa value

[1] `http://topepo.github.io/caret/index.html`
[2] `https://www.rdocumentation.org/packages/RWeka/versions/0.4-39`

slightly greater than 0,5 are highlighted by gray dashed lines. It results in 18 groups (or ML families) marked in different colours. Regarding the dendrogram, the choice of this cut value seems reasonable and a good tradeoff between the number of groups obtained and the distance at which the groups have been created.

3 Results and Discussion

The main results obtained from the studies described at section 2 are summarized as follows:

- Maximum kappa approach for ML family identification: Table 1 shows the confusion matrix for the experiments performed to evaluate our first approach. we observed that some families (such as decision trees) are well identified whereas the method performs badly at identifying other families (such as logistic regression and neural networks). The overall accuracy obtained for this method 30%. As it is an 11-class classification problem the random baseline stays around 9% so this approach improves the random baseline significantly.

Table 1: Confusion matrix (Real vs. Predicted Class) obtained for the maximum kappa approach for ML family identification.

Family	DA	EN	DT	SVM	NNET	NB	NN	GLM	PLSR	LMR	MARS
DA	**1**	0	0	7	0	0	1	14	0	1	1
EN	0	**7**	2	4	0	0	1	1	0	6	4
DT	0	7	**16**	0	0	0	0	1	0	0	1
SVM	0	1	2	**5**	0	0	2	0	0	12	3
NNET	0	0	0	0	**2**	0	0	8	0	15	0
NB	2	1	1	0	2	**1**	0	4	0	1	13
NN	0	5	4	11	0	0	**4**	0	0	1	0
GLM	0	2	0	1	0	0	0	**11**	0	10	0
PLSR	1	0	0	0	2	0	0	18	**0**	3	1
LMR	0	0	0	0	0	0	0	2	0	**22**	1
MARS	0	2	3	1	0	0	0	0	0	4	**14**

- Meta-model approach for ML family identification: Table 2 shows the confusion matrix for the experiments performed to evaluate our second approach. We observed a general improvement in the results with respect to the previous approach as reflected by an overall accuracy of 56%. Now all the families are better identify although it still remains some families that are difficult to distinguish (as discriminant analysis and principal components regression, or linear model and neural network). In general terms, the results show that, although this is a particularly complex problem, the use of dissimilarity measures to differentiate ML families from one another seems an effective approach.

Table 2: Confusion matrix (Real vs. Predicted Class) obtained for the meta-model approach for ML family identification.

Family	DA	EN	DT	SVM	NNET	NB	NN	GLM	PLSR	LMR	MARS
DA	**9**	1	1	0	1	2	0	0	10	0	1
EN	1	**12**	4	3	0	1	2	1	0	1	0
DT	0	1	**21**	0	0	1	0	0	0	0	2
SVM	1	5	1	**10**	0	0	4	1	0	3	0
NNET	0	0	0	0	**19**	0	0	5	1	0	0
NB	5	1	0	0	0	**13**	1	1	3	0	1
NN	1	2	1	1	0	0	**19**	0	0	0	1
GLM	0	1	0	0	10	1	0	**9**	1	1	1
PLSR	4	0	0	0	2	2	0	0	**15**	1	1
LMR	0	1	2	2	2	0	1	2	1	**13**	1
MARS	1	0	2	0	0	1	1	2	0	3	**14**

- Taxonomy of ML techniques based on model behaviour: regarding the 18 families identified in figure 4, we observe that some of them corresponds to a family of algorithm that belong to the same learning strategy, which means that the behaviour of the models generated with these techniques behave similarly. This is, for instance, the cases of decision trees (the light green group containing several implementations of the C5.0 and J48 algorithms), and neural networks (the red group that comprises Multi-Layer Perceptron algorithms, under different configurations, and the Average Neural Networks techniques). However, other groups are conformed by different learning techniques but with similar behaviour. For instance, the family highlighted in pink includes the linear techniques, i.e. Logistic Regression-based methods, SVM with linear kernels and the Penalized Discriminant Analysis technique. Some configurations of the Random Forest algorithm (concretely, Random Forest Decision Based Rules) form its own group (showed in light brown) separated from other ensemble techniques (the group above it in the dendrogram coloured in green). We are preparing a report ([3]) with a more detailed analysis of the families showed in figure 4 as well as the analysis of the taxonomies we obtained when consider the binary datasets and the multiclass datasets separately.

4 Conclusions

With respect to the first research line, we have outlined the possibilities of using machine learning to model partial accounts of the behaviour of an existing model. Terms like interpretability, transparency, trust, or fairness require understanding machine learning models as being complex behavioural systems for which further abstractions are necessary.

Regarding the problem of ML family identification of a black box model, we have defined two approaches based on dissimilitary measures such as the Cohen's kappa coefficient. The second approach, which performs better, consists in using the kappa values as meta-features for representing the black box model, and learning a meta-model that predicts the learning family of unseen black box models. An alternative approach that we thought would be interesting to be explored is to use soft classifiers as surrogate models. These models estimate the membership probability for each example. This would allow us to measure the degree of agreement between the oracle and the surrogate model using evaluation metrics such as RMSE or MAE instead of using the kappa statistic.

Finally, as our approaches for model identification rely on the kappa measure, we have been completing a novel taxonomy of learning families based on the inter-rater agreement of different techniques. To do this, we have been performing an extensive experimental study, using a large collection of both datasets and machine learning techniques to construct a set of machine learning families as complete as possible. We think that our method for family identification may work more accurately over the new set of learning families because they are based on the same principle: the degree of agreement between models.

References

[1] Manuel Fernández Delgado, Eva Cernadas, Senén Barro, and Dinani Gomes Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[2] R. Fabra, C. Ferri, J. Hernández-Orallo, F. Martínez-Plumed, and M.J. Ramírez-Quintana. Modelling machine learning models. *Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE)*, 44:175–186, 2018.

[3] R. Fabra, C. Ferri, J. Hernández-Orallo, F. Martínez-Plumed, and M.J. Ramírez-Quintana. A novel taxonomy of Machine Learning Techniques based on model agreement. Technical report, (in preparation), 2019.

[4] Raül Fabra-Boluda, Cèsar Ferri, José Hernández-Orallo, Fernando Martínez-Plumed, and María José Ramírez-Quintana. Identifying the machine learning family from black-box models. In *Proc. of CAEPIA: Advances in Artificial Intelligence*, volume 11160 of *LNCS*, pages 55–65. Springer International Publishing, 2018.

[5] Fernando Martınez-Plumed, Ricardo BC Prudêncio, Adolfo Martınez-Usó, and José Hernández-Orallo. Making sense of item response theory in machine learning. In *Proceedings of 22nd European Conference on Artificial Intelligence (ECAI), Frontiers in Artificial Intelligence and Applications*, volume 285, pages 1140–1148, 2016.