ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Interpretation of Model Agnostic Classifiers via Local Mental Images

Aluizio Lima Filho[1], Gabriel P. Guarisa[1], Leopoldo A.D. Lusquino Filho[1],
Luiz F. R. Oliveira[1], Carlos A. N. Cosenza[3],
Felipe M. G. França[1] and Priscila M. V. Lima[1,2] *

1–PESC/COPPE, 2–NCE, 3–PEP/COPPE
Universidade Federal do Rio de Janeiro, RJ, Brazil

Abstract. Although successful black-box learning models have been created, understanding what happens when a machine produces a classification response is still a challenge. This work introduces FRWI – Fuzzy Regression WiSARD Interpreter, a novel fuzzy rules-based algorithm that is capable of interpreting the responses of black-box classifiers via the production of local mental images from a WiSARD n-tuple classifier. FRWI is compared with LIME – Local Interpretable Model-Agnostic Explanations, a pioneering agnostic classification interpreter model. To make a quantitative evaluation of interpretable models, a new metric – Interpretation Capacity Score – is proposed. Using this metric, it is shown that FRWI surpasses LIME in producing coherent interpretations.

1 Introduction

The need to interpret responses from learning models gets higher in different situations [1]. Questions arise such as: how the models make the decision in the classification, or when to trust its process, and when not to do so. One way to answer the first question is to show what is relevant to the model. LIME [2] – Local Interpretable Model-Agnostic Explanations – was developed with the motivation to clarify such relevance. There are other interpreter models focused on DNNs, like Gran-Cam [3], that were later introduced in the literature. How ever, LIME does not have feasible interpretation capacity for all learning models, due to interpretable models have scenarios where they work better as learning models. Experimental tests were performed utilizing LIME to explain decisions made by following classifiers: WiSARD [4], Linear model [5] and Random Forest model [6] trained with images data sets. It will be shown that results will select too much in the image as relevant, and it will not let it clear what is happening inside the classifier.For that reason, the idea of creating a degree of relevance for each pixel in the image came as an alternative to interpret the responses of black-box classifiers more feasible. This work introduces FRWI – Fuzzy Regression WiSARD Interpreter, a WiSARD n-tuple classifier that produces local mental images, via a fuzzy rules-based algorithm, as an interpretation of the responses of black-box classifiers. To compare the interpretation capacity of both LIME and FRWI models, the Interpretation Capacity Score metric is defined.

*This work was partially supported by NGD Systems, Inc./COPPETEC grant PESC21713; CAPES, CNPq and FAPERJ, Brazilian research agencies.

The remainder of this article is organized as follows: Section 2 presents the WiSARD model together with the concepts of mental images, interpretation and the LIME mechanism. In Section 3, the new classifier interpreter and corresponding algorithm are proposed. Section 4 provides an evaluation of the weightless and LIME under the light of a novel metric. Section 5 concludes the paper summarising its contribution and stating ongoing and possible future works.

2 Models and related subjects

2.1 WiSARD and mental images

WiSARD [4] is a class discriminator oriented n-tuple classifier, where each discriminator is composed of RAM pieces, called here of neurons. Training in WiSARD consists only in write into memory, while classification is read from memory. DRASiW [7], can generate a visualization of the learned patterns through mental images, by combining all information passed to the WiSARD in the training process. It does this by reading the content of all RAMs and generating a mental image by the discriminator. This is a reverse process on the mapping function, that directs the content of RAMs to the input structure and therefore creating a superposition of all binary training data.

2.2 Interpretation, local mental image and LIME

The kind of interpretation focused on this work is the interpretation of the classifier's answers, where the answer of a classifier is analysed by an interpretable model to define what it is relevant to the classifier. And therefore making feasible for humans to understanding the classifier's behaviour. The interpretation generated is a local mental image, where the interpreter stimulates the classifier with a permutation of an input image with a local context. The LIME [2] is a model capable to interpret some learning models through its answers. Given an input and a classifier, it determines what is relevant to the classifier. This is made through generating permutations of the input. Each permutation is given to the classifier, which in turn returns a probabilities vector of classes. With this, LIME can figure out which regions of the image are relevant to the classifier using a linear model. This output from LIME is a local mental image.

3 A new classifier interpreter

3.1 Generation of local mental images

The local mental image tries to reveal the regions relevant to the classifier and with which degree. To reach this target the Fuzzy Regression WiSARD Interpreter (FRWI) was created by the Algorithm 1.

Algorithm 1 Algorithm to generate local mental image

```
Require: $C$ (classifier), $image$ (input image), $fs$ (feature size),
  $fp$ (features proportion), $S$ (total samples)
  $totalFeatures \leftarrow fp * (width(image) - fs) * (height(image) - fs)$
  $tupleSize \leftarrow fs * fs$
  $rew \leftarrow RegressionWisard(tupleSize)$
  for $i$ from 0 to $S$ do
    $mask \leftarrow selectFeatures(totalFeatures, width(image), height(image))$
    $permutation \leftarrow applyMask(image, mask)$
    $c \leftarrow l2norm(C(permutation), C(image)$
    $d \leftarrow l2norm(image, sample)/getMaxL2Norm(image)$
    $rf \leftarrow applyFuzzy(cf, df)$
    $rew.train(mask, rf)$
  end for
  $lmi \leftarrow rew.getRegressionMentalImage()$
  return $normalize(lmi)$
```

3.2 Generating permutations from the input image

To generate permutations is needed to create the binary mask where the values ones define which position will keep the value of the input image and the values zeros will represent the erased positions. Erase value in this work was assumed as value zero but could be any others values. To tackle the binary mask, a group of random positions are selected. For each selected position the neighbour positions are defined as ones, the size of the neighbourhood is defined by feature size parameter in the Algorithm 1. The remain positions are defined as zeros.

3.3 Calculating factors

The distance factor is calculated by the l^2-norm between the input image and its permutation divided by the max value of the l^2-norm in those conditions. This factor gives information about how far the permutation image is from the input image and therefore the locality of the data. The classifier factor is calculated by the l^2-norm after getting the output of the classifier. This one shows how far is the classification of the permutation from the classification of the input image. It gives information about the locality of the classifier function. It is expected that the classifier's output be a probabilistic vector whose the sum of all values is equal to 1.

3.4 Fuzzy Rules

The fuzzy rules [8] in the present work determine how relevant are the selected features to the classifier. Those rules can be found below:

$$p = (\mu_L(c) \wedge \mu_H(d)) \vee (\mu_L(c) \wedge \mu_M(d)) \vee (\mu_L(c) \wedge \mu_L(d))$$

$$n = (\mu_H(c) \wedge \mu_H(d)) \vee (\mu_H(c) \wedge \mu_M(d)) \vee (\mu_H(c) \wedge \mu_L(d))$$

$$rf = \begin{cases} p - n, & \text{if } p > n \\ 0, & \text{otherwise} \end{cases}$$

Where c and d are the classifier factor and distance factor respectively, p is the positive rule and n is the not positive rule. The membership functions are μ_H to high, μ_M to middle and μ_L to low, where μ_H and μ_L are trapezoidal memberships functions and μ_M is a triangular membership function. The rf is the result, a value between 0 and 1, where 0 is no relevance and 1 full relevance. In the rules above the fuzzy operators($\wedge$; $\vee$) were defined in the following way:

$$a \wedge b \rightarrow a * b$$

$$a \vee b \rightarrow a + b - a * b$$

3.5 FRWI Local Mental Images

In this process, only the structure of Regression WiSARD [9] is used to aggregate all the information and generate the local mental image. The mental image [7] from WiSARD was needed to be adapted to be used with the structure of the Regression WiSARD due to the two dimensions of it. In that case, the same reverse process is made, reading from RAMs and writing in the image, but as there are two dimensions so the y value stored and the counter value are summed separated for each position of the tuple of the RAM and become one dividing $sum(y)$ by the $sum(counters)$. As each RAM address represent a position in the input structure thus it just need to do the reverse process using the mapping as a guide. And so the Regression Mental Image is built, but to achieve the local mental image the result from the above process is normalized by the maximum and minimum values.

4 Methodology, Experiments and Results

4.1 Methodology of evaluation of interpretable models

To evaluate the interpretable models a novel metric was defined, the Interpretation Capacity Score. With such metric, it is made three types of evaluation. The first one evaluates if the classifier keeps the same answer after apply the interpretation output as a mask over the image. The second one evaluates if classifier changes its answer after applying the opposite of the interpretation mask over the image. And the third was added to penalize the case where the whole image is defined as relevant, so the complexity was introduced by calculating how much the interpretation mask is filled, where 1 is full and 0 is empty. Apply those evaluations over several images and calculating the mean for each one, it can be applied the below equation to determine the final metric.

$$ics = (1 - c) * 2 * \frac{p + n}{p * n} \tag{1}$$

Where c is the complexity, p is when it keeps the same answer and n is when it changes the answer.

4.2 Experiments

To test the performance of this work was used the MNIST[10] database and the FASHION MNIST[11] and it was compared with the LIME. The FRWI and the LIME were tested several times over the above mentioned databases to obtain the mean. In each step, a WiSARD model, Linear model and Random Forest model were trained with the training set, and the testing set was used to evaluate the interpreter's models over each classifier, and in the final the interpretation capacity score is applied. The results in Fig. 2 show that the FRWI model performs better than LIME. The first reason for this to happen is the LIME select the most of the image as relevant. Thus, the final score reaches low values. This situation is observable in Fig. 1. Despite this, LIME has an interesting result at Fig. 1 image (c), where it selects until the border of the number seven in the image. One possible explanation for that is the WiSARD model attributes relevance to the frontier of number seven. That behaviour better is seen at the picture (d) from Fig. 1. Nevertheless, the FRWI model gives us a local mental image with different degrees of relevances. Fig. 1 (h) shows that most of the mental image has a high degree of relevance. This makes it difficult to point out which features are really important. However, by selecting the highest degrees of relevance, a smaller region can be delimited as relevant.

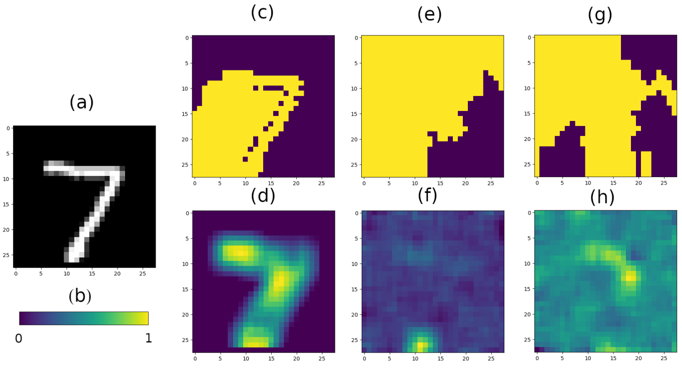

Figure 1: (a) Input image, (b) range of relevance from local mental images(LMI), (c,e,g) LMI from lime, (d,f,h) LMI from FRWI, (c,d) interpreter applied to WiSARD, (e,f) Linear model, (g,h) and Random Forest model

5 Conclusion

The Fuzzy Regression WiSARD Interpreter – FRWI – was developed as a tool to help to understand and explaining the responses of different classifiers over

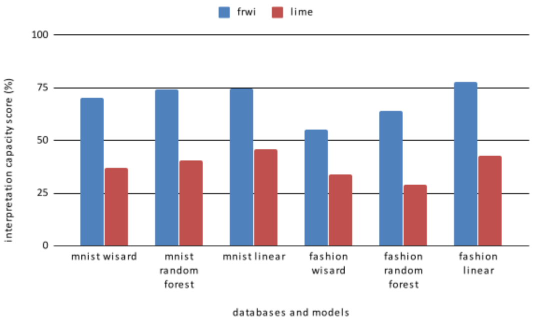

Figure 2: Performance of the interpreter with the interpretation capacity score

different scenarios. In order to do so, several permutation images are generated and fed to target classifiers. The main strategy is to use fuzzy rules to determine what is most relevant to the classifier over a given sample image, so a local mental image is generated to show this information. Also, a quantitative comparison was made concerning the interpretation efficiency of both FRWI and LIME models based on a new metric, the Interpretation Capacity Score. FRWI presented a much better performance than LIME in the evaluated scenarios. One thinks it also suggests a better understanding of humans.

References

[1] D. Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2, 2017.

[2] M. T. Ribeiro et al. 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*, pages 1135–1144. ACM, 2016.

[3] R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[4] I. Aleksander et al. Wisard, a radical step forward in image recognition. volume 4, pages 120–124. MCB UP Ltd, 1984.

[5] ANDERS BJ ORKSTR OM. Ridge regression and inverse problems. *Stockholm University, Department of Mathematics*, 2001.

[6] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[7] B. et al. Grieco. Extracting fuzzy rules from “mental” images generated by modified wisard perceptrons. In *Proc. E*, volume 26, pages 101–773, 2008.

[8] L. Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[9] L. Lusquino Filho et al. Prediction of palm oil production with an enhanced n-tuple regression network. *ESANN*, 2019.

[10] Y. LeCun et al. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[11] H. Xiao et al. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.