JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Explainability Enhancement in Orange Tool: A Hybrid LIME–LRP Interpretation Framework for Neural Models

Dr. Monika Rathore
Associate Professor
Manipal University Jaipur
monika.rathore@jaipur.manipal.edu

Abstract: In recent years Artificial Intelligence and Machine Learning have shown an outstanding performance and have achieved notable attention both in the field of Research and various Industries. And Deep Learning has made a significant contribution in the field of AI. However Deep Learning is treated as a black box because it is unable to give the explanation about how a system has planned or has achieved a particular result. And if a human needs to accept a particular system one needs to understand why/why not the system works. And human users must be able to determine when to trust the system and when the system should not be trusted. In the real-world applications explainability has become essential for both the people and developers who are affected by AI decisions. The decision made by the system can sometime be critical to life, death, and personal wellness. So, there is a need for proper explanation about the decision made by the system. There is a need to approximate the black box in an interpretable way. Explainable Artificial Intelligence (XAI) is a method used in AI which tells how a particular system decides which can be understood by humans or gives an explanation how a system arrives at a particular solution. When a user gets an explanation, he knows when to trust and distrust a system.

Index Terms: Explainable AI(XAI), LRP, LIME.

1. INTRODUCTION:

In the recent years Deep Learning has taken artificial Intelligence to the next level and has shown an outstanding performance in every field be it medical science, e-commerce etc. But the main disadvantage of deep learning is that it is a black box. One does not know how the system or model has taken a particular decision. Previously People also doesn't care about the how the model come to the decision. If the people got the correct result from the model than it is treated as good. But in the few years people started thinking and trying to find out the reason behind how a model takes decisions. And the main reason behind this is trust. There is always an issue when to trust a system as the decision take by the model can sometimes be critical to life, death, and personal wellness, for example in the medical sector. The solution to this kind of problem is Explainable AI(XAI).

Explainable AI(XAI) is the approach which deals with the explainability of the system. Interpretability and explainability plays an important role in the evaluation of a model. Interpretability deals with the 'what' part and explainability deals with the 'why' part. And XAI reduces the gap between interpretability and explainability.

There has been several approaches for explainable AI(XAI):

- Composite framework of Decision Tree and Neural Network.

- Layer-Wise relevance propagation (LRP).
- Neural-Backend Decision Trees (NBDTs).
- Local-interpretability model-agnostic explanation (LIME).

This paper proposes an approach which deals with explainable AI and solve the problem of how a system make a decision and the problem of trust issue.

2. DATASET USED:

For experiment, iris dataset and MNIST handwritten digits data is used for creating two models, one with LRP technique and one with LIME technique.

3. SOFTWARE TOOL USED:

Orange tools are used for the construction of decision tree as well as artificial Neural Network. Orange is an open-source machine learning and data visualization tool.

Python language and Jupiter Notebook is used for construction of two models with LRP and LIME technique. Python is an interpreted, high-level and general-purpose programming language and Jupiter Notebook is open-source IDE used for python coding.

4. CONSTRUCTION OF DECISION TREE AND NEURAL NETWORK

For construction of decision tree and neural network, the iris dataset needs to be imported into orange tool. The dataset contains the following attributes:

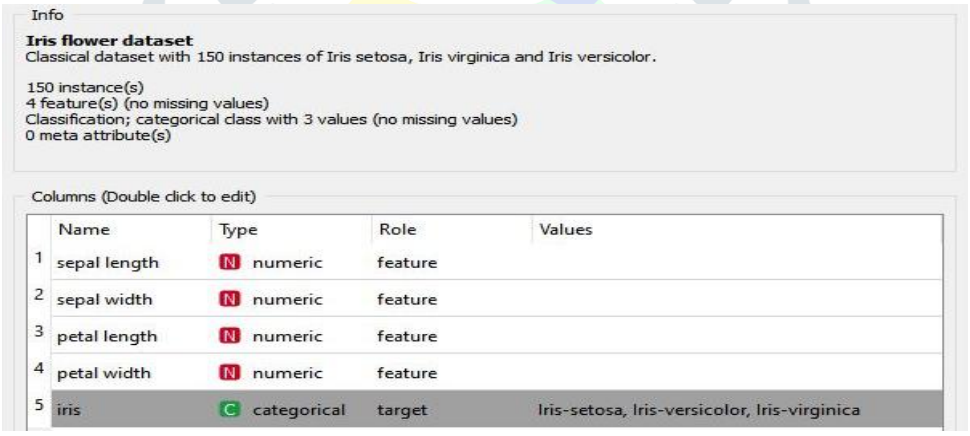

Fig . 1 Iris Dataset

The iris attribute is taken as an target with values Iris-serosa, Iris-versicolor, Iris-virginica. The decision is taken based on 4 Parameter, that is, sepal length, sepal width, petal length, petal width. The Decision tree and neural network is created in orange tool using drag and drop feature.

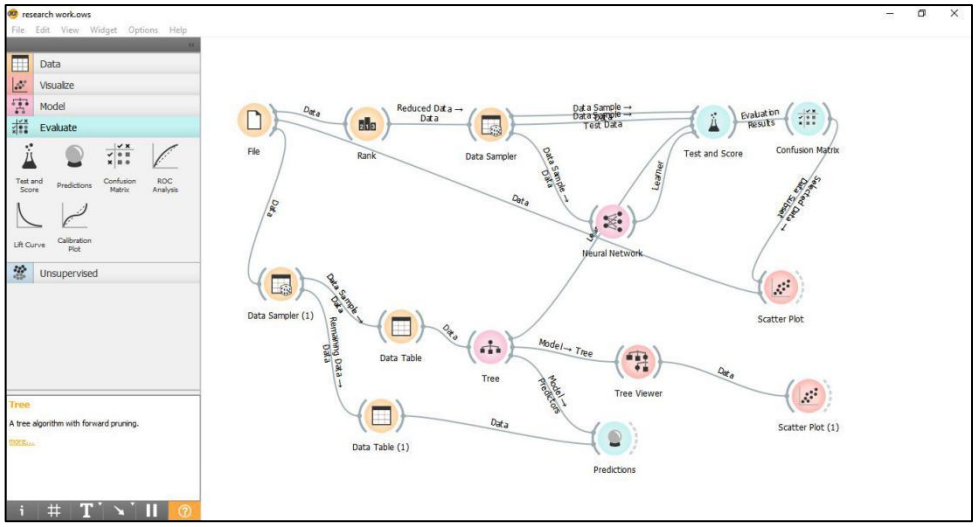

Fig 2. Decision Tree and Neural Network

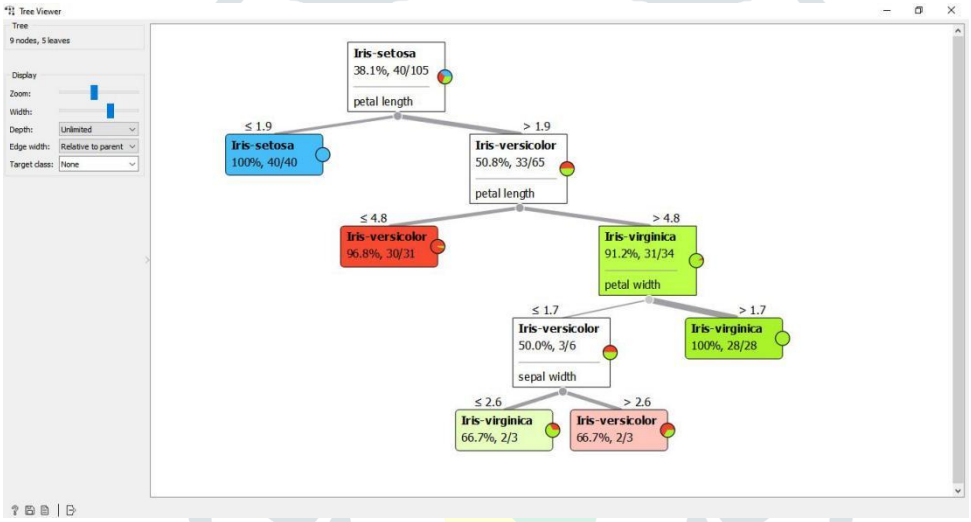

Fig 3. Decision tree with tree viewer

Here, the decision tree is giving some sort of explanation that how the model has taken the decision on the basis of split value. But the neural network has not shown any kind of explanation, it has directly comes to the conclusion. So to solve this problem we have built two more model with LIME and LRP technique respectively.

5. CONSTRUCTION OF LIME MODEL:

LIME(Local Interpretable Model-Agnostic Explanation) is a technique which help in understanding what is going inside a black box model(whiten a black box) which is our agenda for this paper. This method modifies a single data sample by changing the feature values and observing the resulting impact on the output. For each data sample, it attempts to play the role of the 'explainer' explaining prediction. LIME provides a list of explanation which reflect the participation of each feature in forecast of data sample as an output.

For the construction of lime models, iris dataset is used and a random forest is trained. After training the model, the accuracy score of the model was 0.96666666666666667.

After training and testing, the explainer is created and only the numerical features is taken into consideration.

The computed statistics is used for two things:

I. Scaling of the data to figure out the distances when the attributes are not on the matching scale.

II. Sampling unsettled instances - done by sampling from a Normal(0,1), multiplying by the std and adding back the mean.

Here a single instance is explained:

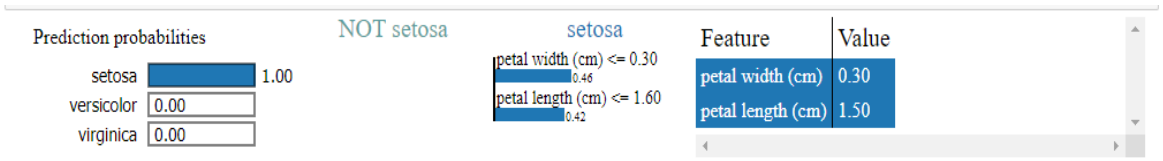

Fig 4. LIME technique showing explanation of two features.

On the right side, in table format it is the row which we are explained. It is to be noted that only the features used in the explanation are displayed because the show all parameters are set to false while explaining a single instance. On the right side, in the table, the value field is the original value for each feature.

Note that the discretize_continous property is set to true because of which the LIME has discretized the features in the explanation and discretized features make for more intuitive explanations.

After explaining single instance, check the local linear approximation for which we have increase the value of petal length and petal width and we got the below output.

```
Increasing petal width
P(setosa) before: 1.0
P(setosa) after: 0.48

Increasing petal length
P(setosa) before: 1.0
P(setosa) after: 0.55

Increasing both
P(setosa) before: 1.0
P(setosa) after: 0.03
```

Fig 5. Output of petal length and petal width after increasing their value

Note that both features (petal width and petal length) had shown some chance after increasing their value. The scale at which they need to be perturbed or unsettled of course depends on the scale of the feature in the training set.

Here all the features is shown, just for completeness:

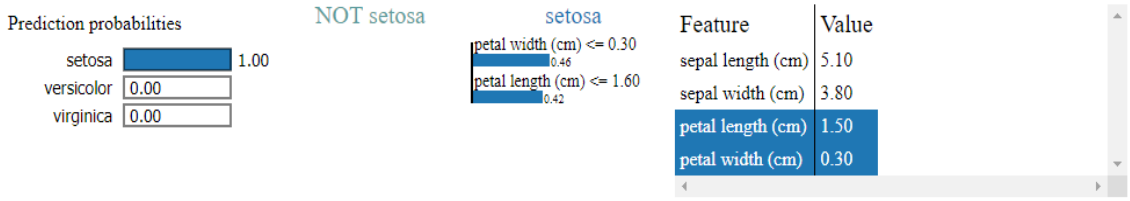

Fig 6. LIME technique showing explanation of all the four features.

From the above image, it has been seen that the LIME technique has giving some sort of explanation on the basis of table and it is easy to understand the explanation that if the petal width and petal length is less than or equal to their original value i.e. 0.30 and 1.60 respectively than the output should be setosa and our prediction probabilities is also showing the output setosa. So the model has shown its reason why the model has come to a particular conclusion on the basis of LIME tabular explainers.

6. CONSTRUCTION OF LRP MODEL:

In our study, For the creation of LRP model we have used the MNIST handwritten digits data. Firstly 12 exemplary MNIST test digits uploaded.

Fig 7. MNIST digits stored as 784-dimensional vector of pixel values

Each digit in the dataset is represented as a 784-dimensional vector, where each element corresponds to a pixel value and ranges from **–1.0** (black) to **+1.0** (white). These vectors are fed into a fully connected neural network configured with layer dimensions **784–300–100–10**, using ReLU activation functions in all hidden layers.

Then an additional top-layer ReLU activation has been added and compared to the original neural network. This however doesn't alter the calculation while focusing at positive output scores. The top layer activations are scores estimating the evidence the network has found for each class. In the following, we show the primary three digits and the scores produced for each class at the output:

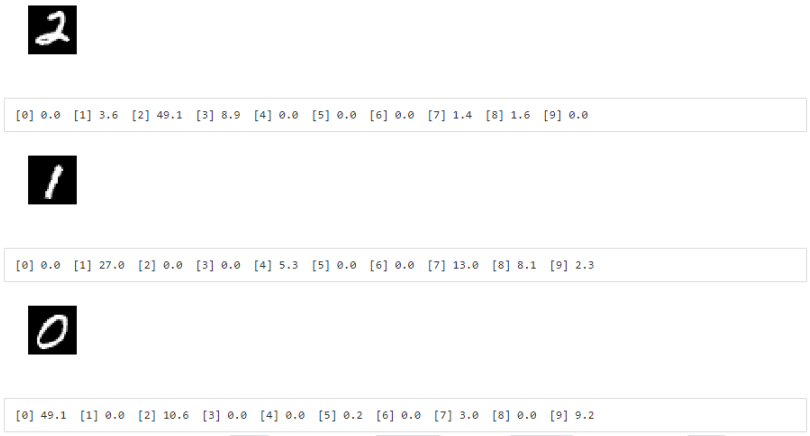

Fig 8. Highest score systematically corresponds to the correct digit.

Next, the Layer-Wise Relevance Propagation (LRP) algorithm is applied to the trained network from top to the bottom of the network using three propagation rules (LRP-0, LRP-ϵ, and LRP-γ) and the retrieved pixel-wise relevance scores from the bottom layer can be rendered as a heatmap.

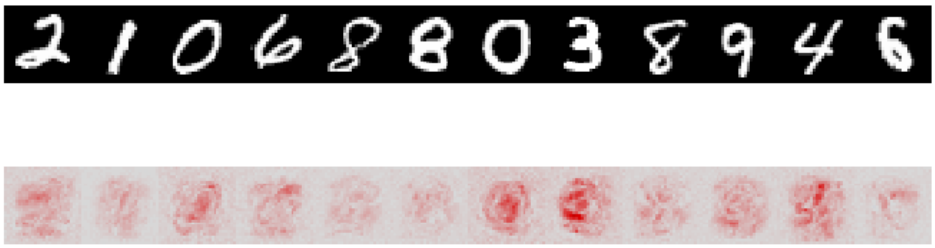

Fig 9. MNIST digit rendered as heatmap

In the LRP maps, red marks show important pixels and blue marks show less important ones. Most relevance appears on the digit itself. For example, extra red lines near **"3"** hint it could be mistaken for **"8"**, and a small mark above **"4"** supports it being **"4"** instead of **"9"**

LRP technique has giving explanation using heatmap about which feature has contributed most in taking the decision which can be helpful in explaining how a system has taken a particular decision.

7. CONCLUSION

A decision tree and neural network has been created in orange tool but neural network was unable to give any sort of explanation so to solve this problem we have created two more model with LIME and LRP technique. Both the technique are giving some sort of explanation by using explanation table and heatmap respectively.

But if we combine LIME and LRP technique and create a new Hybrid model, we can come up we a good model which is good at giving explanation about how a system has taken a particular decision and trust issue can be solved if one get a proper explanation about a system black-box part.

8. REFERENCES

1. Feiyu Xu, Hans Usxkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao and Jun Zhu, *Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges*, September, (2019).
2. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, Aug 9, (2016).
3. Gregoire Motavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Muller, *Layer-Wise Relevance Propagation: An Overview*, Jan, (2016).
4. Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Henry Jin, Suzanne Petryk, Sarah Adel Bargal, Joseph E. Gonzalez, *NBDT: Neural-Backed Decision Trees*, June 11, (2020).
5. Heatmapping, everything about LRP, Available: http://heatmapping.org/
6. Jaime Zornoza, *Explainable Artificial Intelligence*, April 15, 2020. Available: https://towardsdatascience.com/explainable-artificial-intelligence-14944563cc79
7. Patric Ferris, *An introduction to explainable AI, and why we need it*, April 2019. Available: https://www.kdnuggets.com/2019/04/introduction-explainable-ai.html
8. Alvin Wan, *Making Decision Trees Accurate Again: Explaining what Explainable AI did not*, April 18, (2020). Available: https://medium.com/riselab/making-decision-trees-accurate-again-explaining-what-explainable-ai-did-not-abb73e285f22
9. Dr. Matt. Turek, *Explainable Artificial Intelligence (XAI)*, Available: https://www.darpa.mil/program/explainable-artificial-intelligence