**CIDR report- interpretability framework**

Interpretability methods try to take machine learning models and understand their predictions in a more comprehensible way. We focused on interpretability methods that give local explanations sometimes called attributes, i.e., the method explains a predication of trained model on a specific data sample.

We offer a framework for evaluating an interpretability method. Our approach can be summarized in the following scheme [Figure1]:

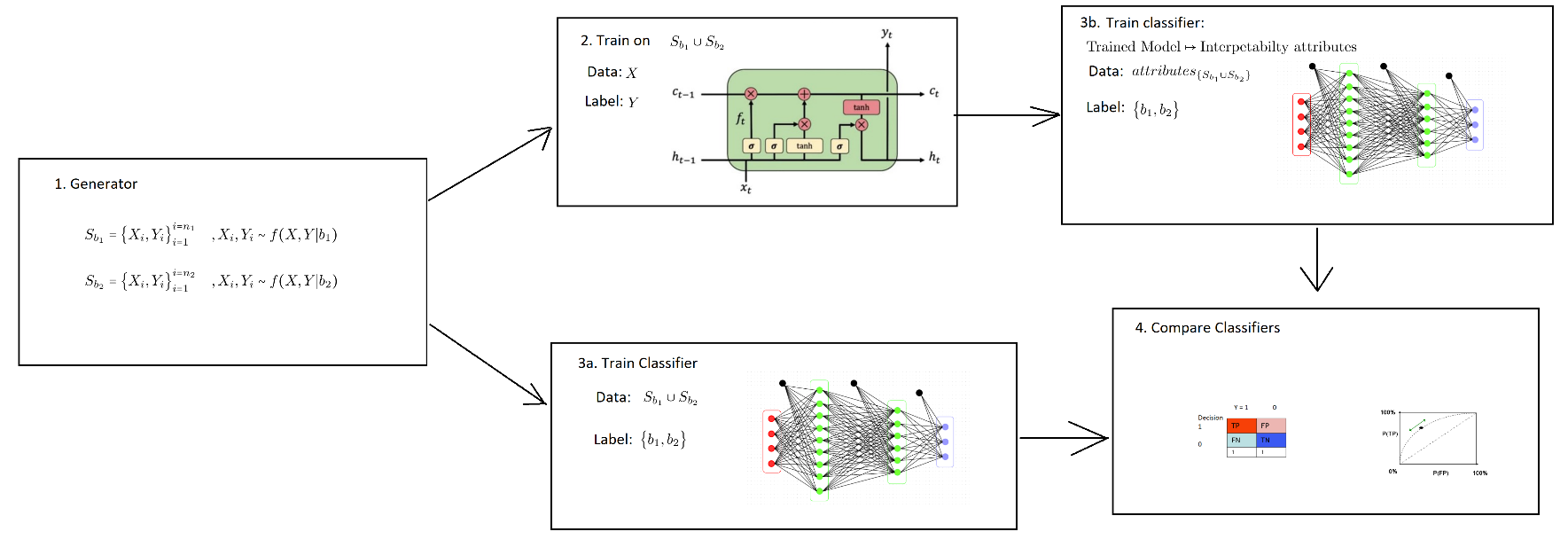
1. Generate 2-modal datasets.
2. Train a machine learning model on the data set.
3. Train two classifiers for the 2-modalities of the data set:
   1. A *clean* view – training only with the data set.
   2. An *interpreted* view – training only with attributes of the trained model.
4. **Evaluate interpretability method by comparison of the *clean* classifier to *interpreted* classifier with simple classification metrics.

Figure 1 Evauation Scheme

We used this framework to evaluate the *Integrated Gradients* [1] interpretability method. Our data sets were taken from the hydrology domain (DREAM model), we basically took two synthetic (but realistic) rainfall data time series, representing two weather scenarios (**step 1 data**). They were merged and fed a rainfall-runoff model that simulated streamflow (**step 1 labels**). We trained LSTM model (**step 2**) and run several interpretability methods and then trained the *clean* and *interpreted* classifiers (**steps 3a+3b**). Finally, we used classical ML classification the metric of test classification accuracy to evaluate how good was the interpretably method.

Our way of "challenging" the interpretably method was to see how performance of the classification metric diminishes while we train over smaller and smaller number of samples from the train set. We compared the results to the *clean view* classification. Our basic assumption was that if we treat an interpretably method as a feature importance rank method, a classifier trained with less and less data samples of the explanations will do better in this sense rather than using the original data samples that were used for the explanations.

Chart, line chart

Description automatically generated

Figure 2

In figure 2 we see the results of the evaluation framework for Integrated Gradient method for 3 types of classifiers: 1. DNN- deep neural network 2-3 SVC -support vector classifier with non-linear kernel and different regularization parameters. These classifiers were trained of different data sizes (x-values) and were evaluated on the test set accuracy (y-values). Our main conclusion was that the clean classification in most of the case prevails the interpreted classification.

Chart, line chart

Description automatically generated

Figure 3

In figure 3 we see the results of the evaluation framework for an Integrated Gradient variant called Local Integrated Gradients. In these cases, we saw that the interpretability method was able to extract explanation that helped the classifier reach better than the clean view.

**References:**

1. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution  
   for Deep Networks". In: CoRR abs/1703.01365 (2017). arXiv: 1703.01365.  
   url: <http://arxiv.org/abs/1703.01365>.