Analyzing predictive models with the help of visualizations

Alessandra Legretto and Paolo Buono
University of Bari Aldo Moro, Italy
{alesandra.legretto,paolo.buono}@uniba.it

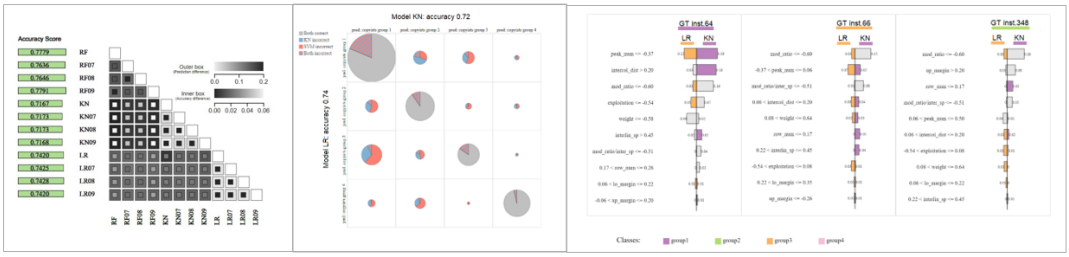

Figure 1: Visualization workflow: (a) Comparison matrix, (b) Pie chart matrix, (c) Instance level explanation

ABSTRACT

Technology may help people in taking everyday decisions exploiting the power of predictive models. Different models applied on the same dataset may have the same performance but produce different outcomes. Making correct predictions is difficult and *"predictive modelers often only explore relatively few models"* [2]. The identification of the most accurate model is also challenging [3, 4]. This research aims supporting the analyst to explore, compare models and select the best one for a given dataset, through the use of visualizations. The approach is taking into account Keim's Visual Analytics Mantra: *"Analyze first, show the important, zoom, filter and analyze further, details on demand"* [1].

KEYWORDS

Data Mining, Visualizations, Predictive Visual Analytics

1 A VISUALIZATION WORKFLOW

As described in the following, the analyst starts by comparing the performance of models in pairs, using a comparison matrix. Then, using a pie chart matrix, she focuses on the behaviour of selected models, according to their predicted classes. Finally, she explores the importance of the features using an instance level explanation visualization technique.

Comparison matrix. It compares the performance of pairs of models prediction. Models are reported on rows and columns of the triangular matrix (Figure 1.a). Each cell contains two nested boxes. One represents the difference in prediction given by the ratio between the number of predicted instances that differ in the two models and the total number of instances in the test set. The other box shows the performance similarity given by the accuracy score difference between the two models. The values of boxes are displayed using a grayscale color that fills the box. The analyst typically seeks cells that have dark outer box and dark inner box, because the task searches for cells with maximum difference in prediction and maximum similarity in accuracy score.

Pie-chart matrix. the pie-chart matrix visualization shows the comparison between the two selected models. (Figure 1.b). Rows and columns represent the classes available in the dataset. The radius of each pie chart is proportional to the number of instances in the cell. The analyst seeks for models that make a more wrong or correct classification in a specific class or in specific patterns.

Instance level explanation. It shows more details about the selected pie chart such as the features used by the two models to classify the selected instances. For instance, each box in Figure 1.c reports an instance and its ground truth, the classifications of the two models and the rules applied to the features to classify instances.

2 CONCLUSION

This proposed visualization workflow helps the analyst to select the models that best fit the dataset and the task the user is most interested in. We are currently studying the reported approach in different dataset with different users.

ACKNOWLEDGMENT

This work is funded by Italian MIUR through PON Ricerca e Innovazione 2014-2020 - Asse I "Investimenti in capitale umano" - Azione I.1 "Dottorati Innovativi con caratterizzazione industriale" (CUP H96D17000040006 approved with D.R.n.986 on 30/03/2017 of University of Bari Aldo Moro).

REFERENCES

[1] Daniel A Keim, Florian Mansmann, and Jim Thomas. 2010. Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter* 11, 2 (2010), 5–8.
[2] Max Kuhn and Kjell Johnson. 2013. *Applied predictive modeling*. Vol. 26. Springer.
[3] Kenney Ng, Amol Ghoting, Steven R Steinhubl, Walter F Stewart, Bradley Malin, and Jimeng Sun. 2014. PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. *Journal of biomedical informatics* 48 (2014), 160–170.
[4] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S Ebert. 2019. Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models. *IEEE transactions on visualization and computer graphics* 25, 1 (2019), 364–373.

WomENcourage 2019, September 16–18, 2019, Rome, Italy
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn