FOLD-R++ AND FOLD-RM

SAMUEL SLOMOWITZ

INTRODUCTION

- Many deep learning and machine learning models have excellent metrics but lack explainability. Wang and Gupta have develop an efficient, fast, scalable, and easy-to-implement algorithm to classify machine learning datasets (Wang and Gupta, 2022). FOLD-R++ is made for binary decision making while FOLD-RM can handle three or more decision making criteria.

TENNIS

- The Kaggle dataset on Tennis (https://www.kaggle.com/sveneschlbeck/beginners-classification-dataset) has two numeric predictors (age, interest level in tennis) and a binary outcome variable success (in the tennis sport overtime). The FOLD-R++ has an accuracy of 1.0 and performed the task in 22 microseconds.

CHURN

- The Kaggle dataset on Churn (https://www.kaggle.com/shubh0799/churn-modelling) has a binary outcome variable of Exited (whether a bank customer left or remained with the bank). The FOLD-R++ had an accuracy of 0.86 and a F1 ratio of 0.92, taking 1.09 seconds to perform.

APPLE OR ORANGE

- The apple or orange dataset (https://www.kaggle.com/raykleptzo/classification-data-apples-oranges) has a binary outcome variable if the row is an apple or orange. The FOLD-R++ had an accuracy of 1 and a precision value of 1, taking 0.22 milliseconds to perform.

BANK

- The bank dataset (https://www.kaggle.com/itsmesunil/bank-loan-modelling) has a binary outcome if a credit card is or is not approved for a customer. The FOLD-R++ had an accuracy of 0.51 and a F1 ratio of 0.45, taking 0.64 seconds to perform.

SURVIVAL

- The survival dataset (https://www.kaggle.com/mitishaagarwal/patient) has 82 predictors and a binary outcome variable of whether someone is identified as a hospital death or not. The FOLD-R++ had an accuracy of 0.92 and a F1 ratio of 0.96, taking less than 2 minutes to perform.

EXERCISE

- The exercise dataset (https://www.kaggle.com/kukuroo3/body-performance-data) is a multi-classification outcome variable of A, B, C, or D where A is the most athletic group of individuals. The FOLD-RM had an accuracy of 0.23 and a F1 ratio of 0.34, taking 0.93 seconds to perform.

DATA SCIENCE

- The data science dataset (https://www.kaggle.com/scarecrow2020/tech-students-profile-prediction) has a multi-classification outcome variable of profile such as “advance backend,” “advanced data science,” “advanced front-end,” “beginner backend,” “beginner data science,” or “beginner front-end.” The FOLD-RM had an accuracy of 0.42 and a F1 ratio of 0.56, taking 2.7 seconds to perform.

AIR

- The air dataset (https://www.kaggle.com/sid321axn/beijing-multisite-airquality-data-set) has a multi-classification outcome variable of station such as “Aotizhongzin,” “Changping,” or “Dingling.” The FOLD-RM had an accuracy of 0.11 and a F1 ratio of 0.19, taking 18 seconds to perform.

ANALYSIS

- *hospital_death(X,'0') :- apache_4a_hospital_death_prob(X,N68), N68=<0.13, not ab3(X), not ab5(X), not ab8(X), not ab9(X), not ab11(X), not ab13(X), not ab14(X), not ab16(X), not ab17(X).*

- The Survival dataset perform well. In addition, the explainability is provided. For example, a patient will survive the value of apache_4a_hospital_death_probaliity is less than 0.13. One exception is if the patient’s peripheral oxygen saturation during the first 24 hours of their unit stay is less than 98%.

LIMITATIONS

- Some limitation included not being able to apply the algorithm to natural language processing and computer vision data. Also, some complex datasets output an overwhelming amount of rule to interpret and select from. Further, multi-classification outcome variables with 12 predictors, for example, has poor metrics as seen with the air dataset included 12 stations instead of 3.

REFERENCES

- Wang, Huaduo, and Gopal Gupta. “Fold-R++: A Scalable Toolset for Automated Inductive Learning of Default Theories from Mixed Data.” ArXiv.org, 14 Feb. 2022, https://arxiv.org/abs/2110.07843.