Meta Feature Classification on a Synthetic Dataset Generator

João Carlos Fonseca Pina de Lemos { up201000660@edu.fe.up.pt }

Abstract. Machine learning (ML) models can perform a vast amount of tasks. Recognition of images and speech, statistical arbitrage (useful in finance), performing medical diagnostics through the analysis of medical exams, and performing predictive analysis are skills in their repertoire. The ML models must be trained using extensive data to accomplish these tasks.

Valid data is usually either scarce, expensive, or not readily available. Scrapping tools are commonly used to circumvent the scarceness of data, but their quality is questionable. A small dataset is also insufficient to produce an accurate prediction model. Even in cases where a large amount of data is available, its use may infringe privacy policies (a common problem when entering the medical field). These problems represent an accumulation of problematic data that leads to incorrect results. Synthetic datasets have a proven track record of improving ML effectiveness by providing a better dataset with fewer data imbalances than its ”organic” bredren.

The work developed in this study focus on the study of meta feature extraction methods and meta feature insertion analysis. Using an already existing Ticket-based Synthetic DataSet Generator as a basis, we also developed a meta extractor feature.

Each dataset has its group of meta features - its fingerprint. Our meta feature extractor feature can dissect any kind of CSV tabular dataset, even if those datasets were not synthetically generated by our generator.

Despite being referenced in many studies, meta features are not uniformly described. Nevertheless, some studies agree on lists of meta features, and those arranged lists are explored. Besides extracting parameters from existing datasets' meta features, it is also essential to analyse the parameter-parameter relationship.

We also explored the concept of meta feature inclusion in the generation process, where we try to force values of certain meta features into the final synthetic dataset.

Supervisor: Daniel Augusto Gama de Castro Silva, PhD; **Second Supervisor:** Leonardo Silva Ferreira.

Keywords: Synthetic Dataset. Meta Features. Machine Learning.

ACM Computing Classification System:

- CCS → Theory of computation → Theory and algorithms for application domains → Database theory → Data modeling
- CCS → Computing methodologies → Machine learning