Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Bachelor Thesis

Bachelor Thesis

Meta-learning for encoder selection

Encoding is the process of embedding categorical data in a numerical space. An **encoder** is a set of translation rules; examples commonly used with tree-based models are *OneHotEncoder* and *TargetEncoder*[a]. Encoding is often considered a necessary evil in machine learning, as most algorithms can only learn from numerical data. On the other hand, properly encoding a dataset can save tuning and training time as well as improve model performance. The literature largely overlooks this opportunity and, to our knowledge, no strategy for encoder selection has ever been proposed.

Meta-learning [1] is best summarized as "learning how to learn". It is about generalizing results obtained from a set of learning tasks to an unseen learning task. To do so, one must first of all populate a *meta-dataset* M:

1. fix a learning algorithm (e.g., `RandomForest`) and a performance metric (e.g., `balanced accuracy`)
2. fix a set of datasets and a set of encoders
3. fix a set of meta-features (e.g., `cross-attribute correlations`) [2]
4. for each dataset: extract meta-features
5. for each dataset-encoder pair: compute the learning algorithm's performance on the encoded dataset, store it in M.

The *meta-task* associated to M is:

*Predict **performance** from **meta-features** and **encoder***

The primary goal of this thesis is to provide human experimenters with a tool to assist in selecting the best encoder. This tool, best in the form of a set of rules, is learned from the meta-task and is both **interpretable** - the experimenter knows what meta-features are most relevant when choosing an encoder - and **generalizable** to new datasets. Learning the ruleset is by no means a trivial task and can be achieved in a multitude of ways; see [3]. It is also of particular interest to check whether such a ruleset generalizes "well" to learning algorithms not considered while building the meta-dataset. If that is the case, it hints that the optimal encoder for a learning task depends only on the dataset's properties.

To summarize what above, the applying student should:

- Review
  - **Encoders** for tree-based models
  - **Meta-features** extraction from categorical and mixed datasets
  - **Meta-learning** for performance prediction
- Populate the meta-dataset M with an appropriate choice of meta-features
- Train some meta-learners and learn an interpretable and generalizable ruleset
  - Rule example: *If the attribute is strongly correlated with the target, encode it using TargetEncoder*
- Explain the learned ruleset
  - *Why are the rules what they are? Why do they work?*
  - **If** the ruleset generalizes well to new learning algorithms, *why is that the case?*

The applicant should possess working knowledge of English and basic knowledge of Python for machine learning.

[1] J. Vanschoren. "Meta-learning: A survey". In: *arXiv preprint arXiv:1810.03548* (2018).
[2] Z. Abedjan et al. "Profiling relational data: a survey". In: *The VLDB Journal* 24.4 (2015), pp. 557–581.
[3] T. Hailesilassie. "Rule extraction algorithm for deep neural networks: A review". In: *arXiv preprint arXiv:1610.05267* (2016).

[a] https://towardsdatascience.com/benchmarking-categorical-encoders-9c322bd77ee8

Ansprechpartner

Federico Matteucci, M.Sc. federico.matteucci@kit.edu Raum: 340

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34