META-EMBEDDINGS FOR AUTOML MODEL SELECTION IN ANOMALY DETECTION

Milos Kotlar[1]

[1] School of Electrical Engineering, University of Belgrade
e-mail: kotlarmilos@gmail.com

Abstract:

This poster reviews meta features for choosing an optimal unsupervised model for anomaly detection by using domain-specific meta features. Also, it proposes improvement by using multi-dimensional dense vectors to limit the dimensions of meta features and to improve speed and performance.

Key words: anomaly detection, automl, meta-embeddings, meta features, meta-learning

1. Meta features for anomaly detection

Detecting anomalies in data is a challenging task where the goal is to observe patterns that differ from the data distribution. Increased development of sensors and edge devices has led to the generation of large amounts of data that are analyzed by systems for their analysis and processing [1], [2]. The performance of such systems solely depends on the quality of the data, the selected model and the model parameters.

Machine learning algorithms can be used for anomaly detection, where for different types of data, some methods give better results than others. According to the no free lunch theorem, there is no one model that works best for every problem. Choosing an optimal unsupervised model for anomaly detection by using domain-specific meta features with multi-dimensional vectors is proposed in this poster. This poster presents follow-up research of an article published in IEEE Access journal [3]. This paper was written according to the guidelines from [6]"

1.1 AutoML systems

Model selection can be automated by using automated machine learning systems (AutoML) that proposes a model for detecting anomalies based on data and meta features that are extracted from the data A growing number of research papers shed light on AutoML frameworks, which are becoming a promising solution for building complex machine learning models without human expertise and assistance [4]. The key challenge in enabling AutoML frameworks to build an efficient model for anomaly detection tasks is to determine the best underlying model for a given task and optimization metric.

1.2 Meta learning in AutoML systems

The meta-learning approaches based on a set of meta features that describes data properties can enable efficient model selection in AutoML frameworks. Meta features are calculated from the data using the appropriate meta functions. In certain cases, meta features are calculated from the created model and later used to create relations between model characteristics and algorithm performance. Meta-learning is present in supervised and unsupervised learning. In both cases, meta features are used to describe the main characteristics of the data and thus transfer knowledge to other domains where they are predictive to the model's performance.

The existing meta-learning approaches based on statistical and information-theoretic meta features require large amounts of data and computational resources to extract data properties. Paper [3] proposes a new set of meta features based on domain-specific knowledge only, where it is shown that the proposed meta features achieve accuracy of 87% and meet the critical requirements for application in AutoML systems, while the existing solutions achieve accuracy of 73%. In cases where there is no significant number of datasets available for evaluation, the proposed solution achieves 25% worse performance compared against the existing solutions.

2. Meta features embeddings

The existing and proposed solutions listed in the previous section uses one-hot encoding where each meta feature is encoded with a sparse vector. Set of meta features is represented as a sparse matrix that leads to the "curse of dimensionality" by creating a new dimension for each new meta feature. Measuring similarity between meta features with increased number of instances could affect speed and performance.

Potential improvements that would solve the "curse of dimensionality" and sparse matrix representation is meta features embeddings. Figure 1 depicts an idea of using embeddings for anomalies sentiments in data and correlating them with the optimal models.

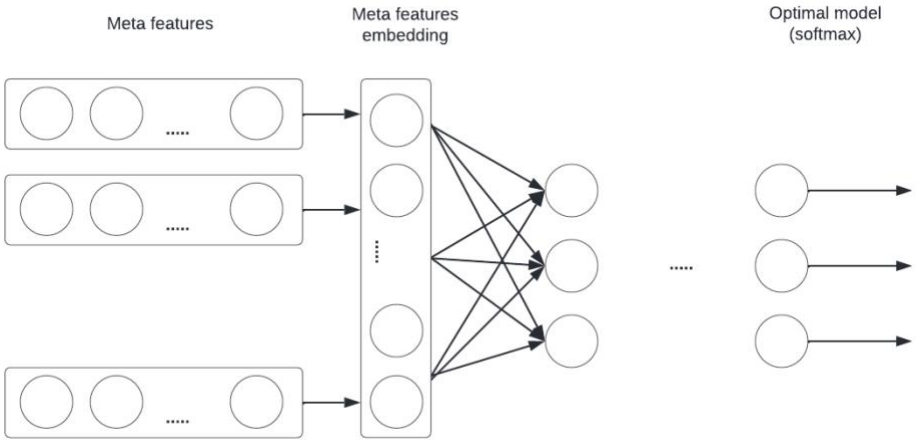

Fig. 1. Proposed meta features embeddings for anomalies sentiments in data and correlating them with the optimal models

3. Conclusion

Proposed changes in this poster could limit the dimensions of meta features and to improve speed and performance. The experiments will be designed and conducted in the future research.

References

[1] A. Oussous, F. Z. Benjelloun, A. A. Lahcen и S. Belfkih, „Big data technologies: A survey," Journal of King Saud University-Computer and Information Sciences, vol. 30, no. 4, pp. 431-448, 2018

[2] W. Günther, M. Mehrizi, M. Huysman и F. Feldberg, „Debating big data: A literature review on realizing value from big data," The Journal of Strategic Information Systems, vol. 26, no. 3, pp. 191-209, 2017

[3] Kotlar, Miloš, et al. "Novel Meta-Features for Automated Machine Learning Model Selection in Anomaly Detection." IEEE Access 9 (2021): 89675-89687

[4] X. He, K. Zhao и X. Chu, „AutoML: A Survey of the State-of-the-Art," Knowledge-Based Systems, т. 212, p. 106622, 2021

[5] A. Rivolli, L. Garcia, C. Soares, J. Vanschoren и A. de Carvalho, „Towards reproducible empirical research in meta-learning," arXiv preprint arXiv:1808.10406, p. 32–52, 2018

[6] Milutinovic V., „The Best Method for Presentation of Research Results," IEEE TCCA Newsletter, 1-6, 1996