Unsupervised representation learning for time series anomaly detection and beyond

Ferreira Correia; L.

Summary

This thesis investigates the feasibility of data-driven and truly unsupervised representation learning to time series anomaly detection problems. Such problems are relevant to the real world, as problems from it often lack available labels, for example due to prohibitively high costs. Therefore, advances in this field would benefit a variety of applications, though unfortunately such progress is currently hindered by the lack of reliable evaluation metrics and high-quality public datasets. This is further complicated by the fact that comparatively little work has been dedicated to contributions in benchmarking. Without metrics and datasets, it is difficult to objectively locate how far the current research has come and how robust claims in the literature really are.

To address this, this thesis contributes to the research field by providing a novel and non-trivial time series dataset, which is generated with state-of-the-art simulation tools. While technically a synthetic dataset, the simulation model it is generated with is real-world inspired, meaning that much of the complexity of a real-world system is modelled. Harnessing simulation to generate data also allows for a greater level of granularity over the anomaly generation process than would otherwise be possible during collection of a real-world dataset.

Another aspect hindering scientific advances from being applied to real-world problems is also the rather loose definition of unsupervised methods in literature. Many publications revolving around time series anomaly detection may claim to be unsupervised, though further inspection reveals that at least one component within the proposed methodology, usually the threshold used to isolate anomalous data points, assumes the presence of labelled data. Therefore, this thesis also investigates how well *truly* unsupervised methods perform on real-world and real-world-like datasets, not only in terms of anomaly detection performance, but also in terms of generalisation capability. Results show that, when training subset is contaminated with anomalies,

truly unsupervised methods, i.e. using the unsupervised threshold, struggle, with the best approach achieving $F_1 = 0.06$. This can be largely attributed to the threshold choice, as setting it to a value that maximises the F_1 score, shows that some approaches achieve much more respectable results of up to $F_1 = 0.58$. When the training subset is cleansed of anomalous data, using the unsupervised threshold yields much better calibrated approaches compared to a contaminated training subset, achieving up to $F_1 = 0.67$. Here, the ideal threshold yields better results with an F_1 score of up to 0.86; a gap much smaller than previously with a contaminated training subset. Overall it becomes clear, that modelling alone does not dictate how well a model performs, since if it is paired with a poorly chosen threshold, the potential of the resulting approach is masked. Furthermore, the experiment investigating the generalisation capability of a model shows that, for the automotive powertrain data used, the model is able to generalise better over the data if the training subset features especially dynamic or high vehicle-speed time series.

Instead of assuming the availability of labelled data, this thesis investigates the feasibility of active learning to further enhance truly unsupervised anomaly detection methods. The proposed framework intelligently collects sequences most dissimilar to one another in an effort to maximise diversity. In real-world problems, these sequences can be presented to a domain expert who can label them, thereby forming a small labelled dataset which can be used for further threshold optimisation. The novel query strategy proposed shows that, especially in low query budget scenarios, it outperforms approaches calibrated with an unsupervised threshold almost across the board, almost matching the results obtained with the best possible threshold. Since even domain experts can make mistakes, this work also considers the effect of mislabelling on the anomaly detection performance, illustrating the robustness of active learning. Even in the most extreme case investigated, where the domain expert was simulated to mislabel almost a third of the queries, approaches calibrated with the resulting active learning-based threshold still outperformed their unsupervised counterpart, indicating a level of robustness to human error.

Moreover, this work also shows how the same data-driven modelling procedure can be applied to other time series problems like predictive maintenance. In industrial settings, predictive maintenance aims to inform about the right time for component changes due to deterioration and wear. This is in contrast to more traditional practices of preventive maintenance, which involves swapping components at an early stage, leading to unused remaining useful life within the component, and run to failure, which involves running a component until it is no longer usable, which can damage

other components in the process. For a case-study from a real-world automotive test bench, the data-driven model was able to detect significant deterioration of the testee 54 min earlier than the currently implemented rule-based monitoring system.

Lastly, this work concludes that several aspects of the research field deserve future work to be dedicated to them. Benchmarking especially requires further contributions, not only in terms of dataset diversity but also in terms of general and flexible evaluation metrics, in addition to the need for the quantification of approach complexity. Additionally, the experiments involving active learning show that robustness to mislabelling can be further improved, which is of large interest to industrial applications. For cases where a domain expert is not available, future work should also seek to provide improved ways of obtaining truly unsupervised thresholds.