Explainable Self-Learning Self-Adaptive Systems

Verena Klös[1]

1 Self-Learning Self-Adaptive Systems

Self-adaptivity enables flexible solutions in dynamically changing environments. However, due to the increasing complexity, uncertainty, and topology changes in cyber-physical systems (CPS), static adaptation mechanisms are insufficient as they do not always achieve appropriate effects. Furthermore, CPS are used in safety-critical domains, which requires them and their autonomous adaptations to be safe and, especially in case of self-learning systems, to be explainable. To achieve this, we have developed an engineering and analysis approach for self-learning self-adaptive systems [KGG16, KGG18a, KGG18b] based on our notion of timed adaptation rules. These rules are extended condition-action rules, that additionally include a timed expected effect of the adaptation actions. In contrast to techniques like online planning, timed adaptation rules make adaptation decisions explicit and comprehensible due to their explicit application condition and timed expected effect.

At run-time, we employ history information of system and environment parameters together with history information of applied adaptation rules to retrace past adaptation decisions and record whether their expected effect was achieved and in which system contexts. We use this information to evaluate the accuracy of adaptation rules and to dynamically improve the adaptation logic through learning. Our rule accuracy evaluation detects inaccurate adaptation rules and classifies the observed effects of previous rule executions w.r.t. the observed deviation from the expected effect. The evaluation results are used to refine deviating adaptation rules into more specific context-dependent rules that capture the observed effects. This enables the accurate co-evolution of the adaptation rules. To learn adaptation rules for new situations that are not covered by the existing rules, we employ executable run-time models of the system and its environment. Those models provide an interface to update them with the dynamically gathered knowledge of the system and environment, to realise co-evolution of models and actual environment. As a result, we achieve an updated adaptation logic that consists of more accurate, yet comprehensible, timed adaptation rules. A separate verification phase enables us to provide offline and online guarantees of evolving adaptation logics based on human-comprehensible formal models.

[1] TU Berlin, TU Berlin, Softw. and Embedded Syst. Eng., Ernst-Reuter-Platz 7, 10587 Berlin, verena.kloes@tu-berlin.de

2 Traceability of Decisions for Comprehensibility

To achieve explainability of autonomous adaptation decisions, we propose precise retracing of previous adaptation decisions, containing their cause (i.e., violated adaptation goals), their contexts (i.e., the current values of system and environment parameters), the chosen rule, the expected, and the actual effect. For explanations of adaptation rule learning, we provide an explanation basis for the following questions: "Why was learning necessary?" "How were the results achieved?" and "Which assumptions were made?"

To this end, we build explanation objects that contain the original adaptation rule (if existent), the kind of learning (observation- or simulation-based learning), the learning result (refined context-specific rules, or newly learned rule), and, for observation-based learning, the underlying evaluation results. To explain the results of simulation-based learning, we propose to log the updated run-time models as these will further evolve over time. Furthermore, we add the fitness function of the learning algorithm to explain the evaluation of found solutions.

3 From Comprehensibility/Traceability to Explainability

Each object in our explanation basis can already be seen as explanation for a single adaptation/ learning decision. While this explanation format is sufficiently comprehensible for experts, i.e. self-adaptive software engineers, further processing to generate textual explanations may be beneficial for non-experts and is part of future work. Furthermore, we plan to consider customized explanations for different target groups. This can be achieved by generating customized textual explanations based on our explanation base. Currently, our explanation base contains a lot of information. To avoid information overload, but still provide sufficient information, we plan to cooperate with cognitive science to investigate human needs and expectations on explanations to achieve really helpful explanations. Additionally, we plan to investigate the use of machine learning on user feedback to learn the characteristics of helpful explanations.

References

[KGG16] Klös, Verena; Göthel, Thomas; Glesner, Sabine: Formal Models for Analysing Dynamic Adaptation Behaviour in Real-Time Systems. In: 3rd Workshop on Quality Assurance for Self-adaptive, Self-organising Systems (QA4SASO). IEEE, pp. 106–111, 2016.

[KGG18a] V. Klös, T. Göthel, and S. Glesner, "Comprehensible and dependable self-learning self-adaptive systems," *Journal of Systems Architecture*, vol. 85-86, pp. 28–42, 2018.

[KGG18b] V. Klös, T. Göthel, and S. Glesner, "Runtime Management and Quantitative Evaluation of Changing System Goals in Complex Autonomous Systems," *Journal of Systems and Software*, vol. 144, pp. 314–327, 2018.