Using Conceptual Modeling to Drive Machine Learning Solutions Development - A Case Report on applying GR4ML

Soroosh Nalchigar[1] and Eric Yu[1,2]

[1] Department of Computer Science, University of Toronto, Toronto, Canada
[2] Faculty of Information, University of Toronto, Toronto, Canada
{soroosh,eric}@cs.toronto.edu

Despite remarkable advances in machine learning and analytical technologies, there has been little attention paid to conceptual modeling for the development of such solutions. While modeling techniques have been proposed to assist in several areas related to the design of machine learning and advanced analytics solutions, we are not aware of any systematic framework that provides model-based support to connect all stages from goal-based requirements to business analytics design (including machine learning) to data preparation.

GR4ML is a conceptual modeling framework for requirements elicitation, design, and development of machine learning solution. The key objectives of this framework include:

- Enabling the modeling, elicitation and clarification of business analytics requirements
- Allowing analysts to derive and model design of machine learning solutions for addressing the analytical requirements
- Modeling data preparation pipelines and workflows for transforming the raw data into prepared datasets for execution of analytical algorithms
- Modeling the link from business strategies, stakeholders, and decisions towards analytics solutions design and algorithms, and thereafter to data preparation workflows and enterprise data assets
- Modeling business analytics knowledge in the form of design catalogues that are expressed in conceptual models
- Providing solution patterns as an explicit way of representing well-proven machine learning designs for commonly-known and recurring business analytics problems based on user, data, and model contexts
- Providing a model-driven methodology for analysis and design of business analytics and machine learning solutions
- Providing modeling guidelines for analysis and design of business analytics solutions (i.e., guiding the use of the framework)

GR4ML includes meta-models, methods, design catalogues and patterns, guidelines, and instantiations. It consists of three modeling views, representing different aspects of a solution and viewpoints of different roles involved in the development of such systems. The Business View supports the elicitation of business analytical requirements by capturing stakeholders, strategic goals,

decisions, questions and required insights. The Analytics Design View supports the design of the solution by capturing algorithms, metrics, and quality requirements, focusing primarily on machine learning solutions. The Data Preparation view supports the design of transformation workflows by capturing data tables, flows, and preparation tasks and operations. These views are linked together to represent a holistic view and bridge the gap from business strategies to machine learning algorithms to data preparation operations. The framework comes with a set of design catalogues and solution patterns that encode and represent generic and well-proven machine learning solutions for commonly-known recurring business problems. The framework is also augmented with methodological steps and modeling guidelines for supporting domain users in working with the framework. It also includes model-based support for linking analytics-driven insights to consequent enterprise actions and changes.

GR4ML has been applied to multiple real-world projects and in empirical studies. These studies have been focused on evaluating the adequacy and expressiveness of the framework, testing its usefulness towards the claimed benefits, revealing its drawbacks, and finding difficulties in using it.

In this talk we report on an empirical study, in the health care domain, that was conducted to validate the expressiveness and usefulness of the framework. Here, the case study research method was employed to evaluate expressiveness and usefulness of GR4ML. Regarding the expressiveness, the case study provides evidence that the framework includes an adequate set of concepts for expressing machine learning requirements and solution design. It reports that using GR4ML modelers could arrive at a characterization of an existing analytics product which was deliberately kept unknown to them during the study. Regarding usefulness, the study reports that GR4ML can be useful by discovering requirements that would be missed without the use of framework, as well as by enhancing the communication and mutual understanding among different roles in an analytics project. Overall, positive feedback on the framework was collected from participants, especially around the goal-oriented thinking approach on machine learning technologies; to start from business goals, refine them into decisions and questions, and thereafter reveal machine learning requirements. Several important areas of improvement were also identified as a result of this study, including lack of guidelines for scoping the goal models in the Business View, and lack of support for prioritizing analytical requirements and for modeling their impact.