Optimizing ML Model-based Query Plans: Navigating the Accuracy vs. Cost Trade-off

Ziyu Li, Asterios Katsifodimos, Marios Fragkoulis, and Alessandro Bozzon

Delft University of Technology
{z.li-14, a.katsifodimos, m.fragkoulis, a.bozzon}@tudelft.nl

Relational databases no longer store simple, structured tables, but are increasingly used for unstructured data, such as text, images or videos. Machine learning (ML) has done huge leaps forward in analyzing such unstructured data. Thus querying inside a database system when combining structured and unstructured data is becoming more and more important. Our work aims at incorporating ML models inside database systems and treating them as relational operators. For instance, binary classifiers can be used within filters.

Like in probabilistic databases, the data items that come as a result of applying ML models to them are probabilistic. For instance, a model classifies an object in an image as a car, with 80% accuracy. Although probabilistic databases have dealt with uncertainty, they have not dealt with *model selection*: given a set of models that can satisfy certain predicates, how do we choose the model with the highest accuracy such that the execution cost of that model is minimized? Moreover, oftentimes the goal is not only to select the most accurate model or a fast model, but to select a set of models that can meet the constraints and optimize for certain objectives. For example, with a constraint on accuracy of the query answer, we need to process the query with the minimum cost. In this case, when querying data with ML models, challenges raise in terms of several aspects: the uncertainty of model selection, the uncertainty of the results produced by the models and finally query evaluation.

In this work we aim to answer queries including predicates on operators covering conjunction, disjunction and negation. We propose an optimization method that can navigate the trade-off between accuracy and cost. To do that, we first introduce a model repository that records model “metadata”, i.e. the predicates those models can answer, and their performance (accuracy vs. cost) on those predicates. The accuracy of models is presented as the probability of an attribute, and applied when computing the probability of each item in the answer of a complex query. In addition, we discuss and elaborate the challenges raised by the optimization mechanism, and the potential for future work.