Analytics Meta Learning

Zi Yang

Abstract

Data analytics is the process to analyze massive data to discover useful knowledge and make conclusion about the information, to improve predictions and support decision making. Solving analytics tasks requires more than just merely applying analysis algorithms, instead it combines high-level decision making and low-level process execution, which makes it more difficult than performing individual analyses or analysis steps. As more and more analysis components have become available nowadays, it has been more challenging and time-consuming than ever to quickly design an extensible architecture, and effectively and efficiently compose an information system from these components in order to achieve a desired or optimal level of performance on a given analytics task. In this these, we study both theoretically and empirically the problem and the solution to assist, if not replace, humans the design, planning, and evaluation in the development of intelligent information systems for analytics tasks. We refer this problem as analytics meta learning.

In the thesis proposal, we formally define the problem of analytics meta learning and propose a solution framework that consists of three steps: analytics procedure definition, analysis component construction, and analytics space exploration. From the theoretical perspective, we focus on design and rigorous analysis of algorithms to extract procedural knowledge for analytics procedure definition and learn to optimize procedures from task benchmarks. We implement a software architecture framework that enables analytics meta learning, and we leverage the framework to solve real-world analytics tasks on three domains of problems, including general biomedical information seeking task, pharmaceutical decision support task, and product recommendation task, and empirically study the performance of the proposed algorithms.

To date, we have made initial achievements in one or some aspects of analytics meta learning. We have developed representation languages for analytics procedures (ECD and DPT), and proposed a greedy algorithm for optimizing a configurable information pipeline (CSE algorithm). The ECD framework has been successfully used in developing intelligent information systems across various domains since its first release in mid 2012, which includes TREC Genomics passage retrieval task, BioASQ QA task, target validation for drug discovery, etc. We have seen positive outcomes from a set of preliminary experiments, e.g., our BioASQ exact answer generation pipeline was ranked #1 in 5 out of 6 focused categories in the official evaluations. The framework has also been employed in the related courses domestically and abroad for educational purposes since 2012. The goal of the thesis is to formally define and systematically study the analytics meta learning problem through theoretical analyses and empirical case studies on a wide range of applications.