Learning to Aggregate on Structured Data

Master Thesis Proposal & Work Plan

Clemens Damke

Matriculation Number: 7011488

cdamke@mail.uni-paderborn.de

October 9, 2019

1 Motivation

Most of the commonly used supervised machine learning techniques assume that instances are represented by d-dimensional feature vectors $x \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ for which some target value $y \in \mathcal{Y}$ should be predicted. In the regression setting the target domain $\mathcal{Y}$ is continuous, typically $\mathcal{Y} = \mathbb{R}$, whereas $\mathcal{Y}$ is some discrete set of classes in the classification setting.

Since not all data is well-suited for a fixed-dimensional vector representation, approaches that directly consider the structure of the input data might be more appropriate in such cases. One such case is the class of so-called *learning to aggregate* (LTA) problems as described by Melnikov and Hüllermeier [1]. There the instances are represented by compositions $\boldsymbol{c}$ of constituents $c_i \in \boldsymbol{c}$, i.e. variable-size multisets. The assumption in LTA problems is that for all constituents c_i a local score $y_i \in \mathcal{Y}$ is either given or computable. The set of those local scores should be indicative of the overall score $y \in \mathcal{Y}$ of the entire composition $\boldsymbol{c}$. The goal of LTA is to learn a variadic aggregation function $A : \mathcal{Y}^* \to \mathcal{Y}$ that estimates such composite scores, i.e. $\hat{y} = A(y_1, \ldots, y_n)$ for a composition with n constituents. Additionally the aggregation function A should be associative and commutative to fit with the multiset-structure of compositions.

Current LTA approaches only work with multiset inputs. In practice there is however often some relational structure among the constituents of a composition. This effectively turns LTA into a graph classification or regression problem. The overall aim of this thesis is to look into the question of how aggregation function learning methods might be generalized to the graph setting.

2 Related Work

This thesis will be based on two currently mostly unrelated fields of research: 1. Learning to Aggregate 2. Graph classification. A short overview of the current state-of-the-art approaches in both fields will be given now.

2.1 Learning to Aggregate

Two main approaches to represent the aggregation function in LTA problems have been explored. The first approach uses *uninorms* [1] to do so. There the basic idea is to express composite scores as fuzzy truth assignments $y \in [0, 1]$. Such a composite assignment y is modeled as the result of a parameterized logical expression of constituent assignments $y_i \in [0, 1]$. As the logical expression that thus effectively aggregates the constituents, a uninorm U_λ is used. Depending on the parameter λ, U_λ interpolates between t-norms and t-conorms which are continuous generalizations of logical conjunction and disjunction respectively.

TODO: Details

Recently Melnikov and Hüllermeier [2] have also looked at an alternative class of aggregation function. Instead of using fuzzy logic to describe score aggregation, *ordered weighted average* (OWA) operators were used.

TODO: Details

2.2 Graph Classification

3 Goals

3.1 Required Goals

3.2 Optional Goals

4 Approach

5 Preliminary Document Structure

1. Introduction
2. ...

6 Time-Schedule

Figure 1: Sketch of the time schedule for the work on the thesis

References

[1] Vitalik Melnikov and Eyke Hüllermeier. "Learning to Aggregate Using Uninorms." In: *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2016, pp. 756–771 (cit. on pp. 1, 2).

[2] Vitalik Melnikov and Eyke Hüllermeier. "Learning to Aggregate: Tackling the Aggregation/Disaggregation Problem for OWA." In: *ACML* (2019) (cit. on p. 2).

Supervisor

Student