# Learning to Aggregate on Structured Data

## Master Thesis Proposal & Work Plan

Clemens Damke

Matriculation Number: 7011488

cdamke@mail.uni-paderborn.de

October 7, 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 valuation $y_i \in \mathcal{Y}$ is either given or computable. The set of those local valuations should be indicative of the overall valuation $y \in \mathcal{Y}$ 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 valuations, 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. Aggregation function learning 2. Graph classification. A short overview of the current state-of-the-art approaches in both fields will be given now.

## 2.1 Aggregation Function Learning

## 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 p. 1).

Supervisor

Student