# Learning to Aggregate on Structured Data

Clemens Damke

*November 8, 2019*

PADERBORN UNIVERSITY
*The University for the Information Society*
Department of Electrical Engineering,
Computer Science and Mathematics
Warburger Straße 100
33098 Paderborn

INTELLIGENT SYSTEMS
Intelligent Systems Group (ISG)

Master Thesis

# Learning to Aggregate on Structured Data

Clemens Damke

*1. Reviewer* Prof. Dr. Eyke Hüllermeier
Intelligent Systems and Machine Learning Group (ISG)
Paderborn University

*2. Reviewer* Prof. Dr. Axel-Cyrille Ngonga Ngomo
Data Science Group (DICE)
Paderborn University

November 8, 2019

**Clemens Damke**

*Learning to Aggregate on Structured Data*

Master Thesis, November 8, 2019

Reviewers: Prof. Dr. Eyke Hüllermeier and Prof. Dr. Axel-Cyrille Ngonga Ngomo

Supervisor: Vitalik Melnikov

**Paderborn University**

*Intelligent Systems and Machine Learning Group (ISG)*

Heinz Nixdorf Institute

Department of Electrical Engineering, Computer Science and Mathematics

Warburger Straße 100

33098 Paderborn

# Abstract

# Contents

# 1 Introduction

## 1.1 Motivation

The field of *machine learning* (ML) on graph-structured data has applications in many domains due to the general expressive power of graphs. Three common types of graph ML problems are

1. **Link prediction:** A graph with an incomplete edge set is given and the missing edges have to be predicted. The generation of friendship suggestions in a social network is a typical example for this.

2. **Vertex classification & regression:** Here a class or a score has to be predicted for each vertex of a graph. In social graphs this corresponds to the prediction of properties of individuals. Another example is the prediction of the amount of traffic at the intersections of a street network.

3. **Graph classification & regression:** In this final problem type a single global class or continuous value has to be predicted for an input graph. The canonical example for this is the prediction of properties of molecule graphs, e.g. the toxicity or solubility of a chemical.

In this thesis we will focus on the last problem type, *graph classification and regression* (GC/GR). A ML method for this problem has to accept graphs of varying size and should be permutation invariant wrt. the vertices. Those requirements are not met by the commonly used learners that only accept fixed-size feature vectors as their input, e.g. logistic models, support vector machines or multilayer perceptrons.

A GC/GR method has to account for two central aspects of the problem: 1. Local structural analysis and 2. global aggregation. The first aspect is about the extraction of relevant features of substructures of the input graph. The latter is about the way in which the local features are combined into a final class or regression value. The existing GC/GR methods are mostly motivated by local structural graph analysis. The aspect of global aggregation on the other hand is less emphasized by those methods.

There is however a separate branch of research that specifically looks at the problem of learning aggregation functions, called *learning to aggregate* (LTA). Current LTA approaches explicitly learn an aggregation functions for sets which can be interpreted as graphs without edges. The motivation for this thesis is to generalize LTA from sets to arbitrary graphs. The overall goal is to combine the aggregation learning perspective with existing GC/GR methods.

## 1.2 Goals

To extend LTA to graphs, three goals have to be achieved:

1. **Formalization of LTA:** Before LTA can be extended, its essential characteristics have to be defined. Those characteristics should provide the terminology to formally capture the differences and similarities between LTA and existing GC/GR methods.

2. **Give an LTA interpretation of GC/GR methods:** Using the LTA formalization, representative GC/GR approaches should be stated in terms of an LTA method. Currently there is no comprehensive formulation of the relation between both fields of research; this is addressed by the the second goal.

3. **Define an LTA method for graphs:**

## 1.3 Structure

**Chapter 2: Related Work**

**Chapter 3: Learning to Aggregate on Graphs**

**Chapter 4: Evaluation**

**Chapter 5: Conclusion**

# 2 Related Work

# Learning to Aggregate on Graphs

# 3

# 4 Evaluation

# 5 Conclusion

## 5.1 Review

## 5.2 Future Directions

# Appendix A

# List of Figures

# Erklärung zur Masterarbeit

Ich, Clemens Damke (Matrikel-Nr. 7011488), versichere, dass ich die Masterarbeit mit dem Thema *Learning to Aggregate on Structured Data* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Stellen der Arbeit, die ich anderen Werken dem Wortlaut oder dem Sinn nach entnommen habe, wurden in jedem Fall unter Angabe der Quellen der Entlehnung kenntlich gemacht. Das Gleiche gilt auch für Tabellen, Skizzen, Zeichnungen, bildliche Darstellungen usw. Die Masterarbeit habe ich nicht, auch nicht auszugsweise, für eine andere abgeschlossene Prüfung angefertigt. Auf § 63 Abs. 5 HZG wird hingewiesen.

*Paderborn, 8. November 2019*

Clemens Damke