Master's Thesis

Superb English Titel

Toller Deutscher Titel

Bastien Grasnick
bastien.grasnick@student.hpi.uni-potsdam.de

Hasso Plattner Institute for IT Systems Engineering
Enterprise Platform and Integration Concepts Chair

August-Bebel-Str. 88
14482 Potsdam, Germany
http://epic.hpi.de/

Supervisors:

Prof. Dr. Hasso Plattner
Dr. Matthias Uflacker
M. Sc. Cindy Perscheid

Hasso Plattner Institute
Potsdam, Germany

August 28, 2017

Abstract

Awesome Abstract

Zusammenfassung

Deutsche Zusammenfassung

Contents

List of Figures

List of Tables

1

Introduction

According to the WHO, cancer was the second largest cause of death around the world in 2015 being the reason of almost 1 out of 6 deaths.[1] Statistics from 2012 estimate that there were 14.1 million new cancer patients while 8.2 million people died from it. [1] Moreover, prognosis for the future estimate around 23.6 million new cancer patients per year in 2030. [2] This shows how severe this disease impacts humankind.

However, cancer is not just one single disease. There are more than 100 different types and subtypes of cancer[3]. Moreover, treatments and their effectiveness may vary substantially from type to type. Therefore, an accurate diagnosis of cancer type is crucial in order to improve treatment success and reduce costs.

One approach to advance cancer diagnosis is the utilization of computational methods. In particular, sequencing the transcriptome of samples from patients with various cancer types enables the analysis of the disease for research through computational methods. One aspect of this analysis is to try and classify samples. In recent years, advances in sequencing technology enabled the development of RNA-Sequencing (RNA-Seq), a method that provides higher coverage and greater resolution than previous techniques. Most RNA-Seq experiments feature few samples (in the tens or hundreds) and measure gene expression levels in the order of ten thousand.

Therefore, classification using machine learning approaches in these scenarios suffers from the curse of dimensionality resulting in feasibility problems, mainly extremely long processing times. To prevent this, feature selection techniques have been utilized and even especially developed for the field of bioinformatics.

[1] http://www.who.int/cancer/en/

[2] http://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer

[3] https://www.cancer.gov/types

Their goal is to reduce the set of genes used in machine learning models to a small number that achieves good performance while providing reasonable computation time. Moreover, the selected genes can be seen as associated with the investigated disease.

On the other hand, existing knowledge bases like the Comparative Toxicogenomics Database (CTD)[4] or UniProt[5] already contain a lot of information about the connection between diseases and genes. The information about these associations could already be used in the gene selection stage. This would erase the need of employing feature selection.

Therefore, the goal of this master's thesis is to examine whether the utilization of external biological knowledge for gene selection in cancer classification achieves performance comparable to the results of popular feature selection techniques.

The rest of the thesis is organized as follows. Chapter 2 covers related work and background knowledge. Chapter 3 describes the experiment conducted and lists the results. Those results are evaluated in chapter 4. A discussion follows in chapter 5. Finally, chapter 6 will conclude and discuss possible future work.

[4] http://ctdbase.org/

[5] http://www.uniprot.org/

2

Motivation

Super Motivation.

3

Problem Statement

In this work we consider the problems. Therefore we want to answer the following research questions in our work.

1. Q1 ?
2. Q2 ?

For this work the following restrictions apply.

1. Restriction 1
2. Restriction 2

To answer our research questions, the remainder of this work is structured in the following way.

4

Related Work

Related Work.

5

Concept and Implementation

This Chapter provides a detailed understanding of our approach.

5.1 Mathematical Model

Mathmatical Model.

5.2 Architecture

Architecture.

6

Evaluation

We evaluate our architecture.

6.1 Benchmarking Setup

6.2 Experiments

7

Conclusion

We answered our research questions in the following way.

8

Future Work

Future Work

References

[1] Lindsey A. Torre, Freddie Bray, Rebecca L. Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 65(2):87–108, 2015.

A

Appendix

Appendix

Eigenständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die genannten Quellen und Hilfsmittel verwendet habe.

Potsdam, 28. August 2017

Bastien Grasnick