

LUDWIG-MAXIMILIANS-UNIVERSITY
MUNICH

INSTITUTE FOR STATISTICS

Master Thesis

A comparison study of prediction
approaches for multiple training
data sets and test data with
block-wise missing values

Author:

Frederik
LUDWIGS

Supervisor:



Dr. Roman
HORNUNG



February 8, 2020

Abstract

Contents

1	Introduction	3
2	Block-wise missingness in multi-omics data	6
	2.1 Block-wise missingness in train data	6
	2.2 Block-wise missingness in test data	6
3	Methods	7
	3.1 Overview of the approaches	7
	3.2 Random Forest for Classification	7
	3.3 Reference Approach 1 - clinical data only	7
	3.4 Reference Approach 2 - removing missing data	7
	3.5 Imputation Approach - missForest	7
	3.6 Adjustment Approach 1 - blockwise RF	7
	3.7 Adjustment Approach 2 - foldwise RF	7
4	Benchmark Experiments	8
	4.1 Datasets	8
	4.1.1 Own simulated data	8
	4.1.2 Simulated data from Jonas Hagenberg's thesis	8
	4.1.3 Real-Life data	8
	4.2 Accessing the Performance	8
	4.2.1 CV - Testingsituations!	8
	4.2.2 Metrics	8
5	Results	9
	5.1 Own simulated data	9
	5.1.1 Scenario 1	9
	5.1.2 Scenario 2	9
	5.1.3 Scenario 3	9
	5.1.4 Scenario 4	9
	5.2 Simulated data from Jonas Hagenberg's thesis	9
	5.3 Real-Life data	9
6	Discussion and Conclusion	10
7	Bibliography	11
8	Attachment	13

1 Introduction

On October 1, 1990 the international scientific research project named *Human Genome Project* was launched, with the aim to sequence the first complete human genome ever [3]. After investments of totally \$2.7 billion the sequencing was officially finished in 2003 [14]. Since then there was a biomedical progress on the one hand, where for example a “number of disease genes have [...] been identified, leading to improved diagnosis and novel approaches in therapy” [16]. On the other hand there was also an “extraordinary progress [...] in genome sequencing technologies” [[6], p. 333], which led to a sharp drop in sequencing prices. Nowadays whole genome sequencing is available and affordable for almost everyone - e.g. ‘Veritas Genomics’, offers whole genome sequencing for ~\$700 [18]. Besides the ‘genome’, that carries the whole genetic material of an organism, there are also many other types of ‘-omes’, such as ‘epigenomes’, ‘transcriptomes’, ‘proteomes’, ‘microbiomes’, etc., each carrying a different type of information. Similar to the ‘genome’ the time and cost needed to obtain data from ‘-omes’ in general, reduced drastically [[1], [2], [4], [5], [17], [19]]. The methods used, to obtain “fast, automated analyses of large numbers of substances including DNA, RNA, proteins, and other types of molecules” [15] are summarized under the term ‘High Throughput Technologies’. These technologies have made data from molecular processes available for many patients on a large scale. The data, collected from diverse molecular processes are generally referred to as ‘omics’ data.

In the clinical context it is highly interesting to incorporate omics data into statistical approaches. A common example in this context is the survival time prediction for cancer patients, where additionally to the regular clinical data (e.g. ‘BMI’, ‘age’, ‘blood type’, ...) gene expression data has been incorporated into the survival models. This additional omics data has “often been found to be useful for predicting survival response[s]” [[2], p. 1]. In “the beginning, only data from single omics was used to build such prediction models, together or without standard clinical data” [[8], p. 1]. The usage of multiple different types of omics in a single prediction approach was the logical next step and coined the term ‘multi-omics’. The theoretical aspects of integrating several omics types into a single prediction model and how to deal with the block-wise structures has been topic of several papers already - e.g. Hermann 2019 [8], Hornung et al. 2019 [11], Klau et al. 2018 [12], Hieke et al. 2015 [9], Zhao et al. 2015 [20],

In this paper we will deal with a special type of missing data, “that is common in practice, particular in the context of multi-omics data” [10], the so called *Block-wise Missingness*. This occurs, when there are multiple training

sets available that all have the same response target, but different covariates [10]. In the context of multi-omics data, this can for example arise, when different hospitals do research on the same disease, but collect different types of omics for this. Patients from the different hospitals will therefore have different observed blocks of omics data. When concatenating these diverse training sets, it results in a data frame where the patients from the different hospitals miss omics blocks, that were observed any other hospital but theirs. When for example the patients from hospital 'Y' only have 'Y-Omic' as observed omics block, they miss all the other omics blocks, that were collected in the remaining hospitals.

To fit a statistical model, most of the approaches require fully observed data. In the setting of block-wise missingness this is clearly not the case, so that we either need to adjust our prediction models or process the data in a way, that we can fit regular models on it. This emerges the following challenges: How can we fit a model on the data, without removing observations or whole feature blocks? How does a model, that uses fully observed features/ observations only perform in comparison? Does imputation work properly in these settings? How does a model that uses clinical data only perform in comparison to the alternative approaches?

Additional to the problem of block-wise missingness, omics data also have the challenge of high dimensionality. Data from a single omics type can easily exceed 10,000 features and usually result in dataframes with less observations than features [$n < p$] [8]. Klau et al. 2018 [12] states, that besides the predictive performance of an approach it is furthermore important for the approach to be sparse. "Sparsity is [...] an important aspect of the model which contributes to its practical utility" [[12], p. 3], as it makes the model much more interpretable than models including several thousands of variables.

A type of model that naturally handles high dimensional data, even if the number of observations is lower than the amount of features [$n < p$], is the random forest method [8]. Additionally to this, it can handle different input types, doesn't need a lot of tuning and yields comparable predictive performances [7]. The only drawback is that it is not as interpretable as "models yielding *in* coefficient estimates of few relevant features" [[8], p. 35], as for example penalised regression approaches. Nevertheless variable importance measures can be extracted with the random forest method, aswell as partial dependence plots. Furthermore the random forest method has been used successfully in various articles dealing with multi-omics data - e.g. Hornung et al. 2019 [11], Herman 2019 [8], ... Additionally, there have been two proposals by Roman Hornung 2019 [10] and Norbert Krautenbacher 2018 [13], that modify the random forest approach, so that it can deal with block-wise missing data.

An other promising approach in the context of multi-omics data are different adaptations of penalised regression, as for example the priority-lasso [12] or the IPF-Lasso [2]. These approaches can also be modified so that they handle block-wise missingness. Even though the theoretical aspects of these approaches are not part of this thesis, we compare the performances of these with the different random forest methods/ adaptations described in the 'Methods' chapter of this work. The theoretical aspects of these penalised regression adaptations are part of Hagenberg's thesis.

The thesis at hand aims to provide a large scale comparison of classifying methods capable to deal with block-wise missingness in multi-omics data. For this **we** compare the predictive performance of a naive approaches, a random forest based imputation method, two random forest adaptations and the adaptations of penalised regression on multiple data sets. At first the term 'block-wise missingness' is defined in more detail and it is explained how it can arise in multi-omics data - aswell in train, as in test data. Following, the theory of the random forest method for classification is explained, as well as how this method can be used for imputation. Then the two adjustments of the random forest approach for the block-wise missingness are illustrated. In the section 'Benchmark Experiment' **the three different data-sources** for the benchmark data are described and **the the** data itself is investigated. Also the metrics and approaches used for evaluation are defined. In the 'Results' chapter the results are analysed and all approaches are compared over the three different data sources. In last chapter all findings of the thesis are discussed, conclusions are drawn and an outlook is given.

2 Block-wise missingness in multi-omics data

2.1 Block-wise missingness in train data

2.2 Block-wise missingness in test data

3 Methods

3.1 Overview of the approaches

3.2 Random Forest for Classification

3.3 Reference Approach 1 - clinical data only

3.4 Reference Approach 2 - removing missing data

3.5 Imputation Approach - missForest

3.6 Adjustment Approach 1 - blockwise RF

3.7 Adjustment Approach 2 - foldwise RF

4 Benchmark Experiments

4.1 Datasets

4.1.1 Own simulated data

Subsetting the omics blocks

Inducing blockwise missingness

4.1.2 Simulated data from Jonas Hagenberg's thesis

4.1.3 Real-Life data

4.2 Accessing the Performance

4.2.1 CV - Testingsituations!

4.2.2 Metrics

5 Results

5.1 Own simulated data

5.1.1 Scenario 1

5.1.2 Scenario 2

5.1.3 Scenario 3

5.1.4 Scenario 4

5.2 Simulated data from Jonas Hagenberg's thesis

5.3 Real-Life data

6 Discussion and Conclusion

7 Bibliography

- [1] Ke Bi et al. “Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales”. In: *BMC genomics* 13.1 (2012), p. 403.
- [2] Anne-Laure Boulesteix et al. “IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data”. In: *Computational and mathematical methods in medicine* 2017 (2017).
- [3] Francis S Collins. “Medical and societal consequences of the Human Genome Project”. In: *New England Journal of Medicine* 341.1 (1999), pp. 28–37.
- [4] Valeria D’Argenio. “The high-throughput analyses era: are we ready for the data struggle?” In: *High-throughput* 7.1 (2018), p. 8.
- [5] Gregory B Gloor et al. “Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products”. In: *PloS one* 5.10 (2010).
- [6] Sara Goodwin, John D McPherson, and W Richard McCombie. “Coming of age: ten years of next-generation sequencing technologies”. In: *Nature Reviews Genetics* 17.6 (2016), p. 333.
- [7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- [8] Moritz Herrmann. “Large-scale benchmark study of prediction methods using multi-omics data”. PhD thesis. 2019.
- [9] Stefanie Hieke et al. “Integrating multiple molecular sources into a clinical risk prediction signature by extracting complementary information”. In: *BMC bioinformatics* 17.1 (2016), p. 327.
- [10] Roman Hornung. “Random forests for multiple training data sets with varying covariate sets”. manuscript - unpublished yet. 2019.
- [11] Roman Hornung and Marvin N Wright. “Block Forests: random forests for blocks of clinical and omics covariate data”. In: *BMC bioinformatics* 20.1 (2019), p. 358.
- [12] Simon Klau et al. “Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data”. In: *BMC bioinformatics* 19.1 (2018), p. 322.

- [13] Norbert Krautenbacher. “Learning on complex, biased, and big data: disease risk prediction in epidemiological studies and genomic medicine on the example of childhood asthma”. PhD thesis. Technische Universität München, 2018.
- [14] *National Human Genome Research Institute*. <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>. Accessed: 2020-01-07.
- [15] *National Institutes of Health*. <https://commonfund.nih.gov/arra/highthroughput>. Accessed: 2020-01-30.
- [16] Belinda JF Rossiter and C Thomas Caskey. “Impact of the Human Genome Project on medical practice”. In: *Annals of surgical oncology* 2.1 (1995), pp. 14–25.
- [17] Shrutii Sarda and Sridhar Hannenhalli. “Next-generation sequencing and epigenomics research: a hammer in search of nails”. In: *Genomics & informatics* 12.1 (2014), p. 2.
- [18] *Veritas - The Genome Company*. <https://www.veritasgenetics.com/myGenome>. Accessed: 2020-01-19.
- [19] Forest M White. “The potential cost of high-throughput proteomics”. In: *Sci. Signal.* 4.160 (2011), pp. 8.
- [20] Qing Zhao et al. “Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA”. In: *Briefings in bioinformatics* 16.2 (2015), pp. 291–303.

8 Attachment

Figures

Tables