Artificial Neural Networks with small Datasets. A practical Approach

Masterarbeit

zur Erlangung des akademischen Grades

Master of Science in Engineering (M.Sc.)

Eingereicht bei:

Fachhochschule Kufstein Tirol Bildungs GmbH

Data Science & Intelligent Analytics

Verfasser:

Paul Leitner, BA

1910837299

Erstgutachter	:	Dr. Johannes Luethi
Zweitgutachter	:	Lukas Demetz, PhD

Abgabedatum:

31. October 2021

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Masterarbeit selbständig und ohne fremde Hilfe verfasst und in der Bearbeitung und Abfassung keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe. Die vorliegende Masterarbeit wurde noch nicht anderweitig für Prüfungszwecke vorgelegt.

Kufstein, 31. October 2021

Paul Leitner, BA

Contents

List of Figures

List of Tables

List of Listings

List of Acronyms

CNN Convolutional Neural Network

GB Gradient Boosting

nn Neural Network

ml Machine Learning

SMOTE synthetic minority oversampling technique

FH Kufstein Tirol

Data Science & Intelligent Analytics

Abstract of the thesis: **Artificial Neural Networks with small Datasets. A practical Approach**

Author: Paul Leitner, BA

First reviewer: Dr. Johannes Luethi

Second reviewer: Lukas Demetz, PhD

After giving a summary on the literature and history of neural networks, I elucidate the trade-offs between deep learning and other machine learning approaches. I show that machine learning approaches such as Gradient Boosting (GB) mostly trade increased data requirements in favor of data scientist worktime in data preparation and feature engineering. I then investigate whether more complicated Neural Networks (nns) may be used by synthetically enlarging the training data present and thereby achieving comparable accuracy while saving data preparation time, effectively trading processing time (synthetic data enlargement being resource-intensive) for manual feature-engineering time by creating a nn model and benchmarking it against a GB reference model on a standard Machine Learning (ml) dataset with small data, the diabetic retinopathy dataset.

insert result - how much better does this perform? tradeoffs!

note - synthetic data Hittmeir et al. (2019)

31. October 2021

1. Introduction

In 2012 Krizhevsky and his colleagues entered and won the ImageNet classification contest with a deep convolutional neural network Convolutional Neural Network (CNN), outperforming other models by a significant margin Krizhevsky et al. (2012). This marked a turning point in machine learning in general, and in perceptual tasks specifically.

Currently, data scientists spend a significant amount (how much? sources!) of their time, when solving 'shallow' machine learning tasks (such as???) in feature engineering / preprocessing. Source! examples! This is due to the fact that shallow approaches such as decision trees, GBM and SVM models require features that 'directly' connect the prepared data to the searched-for outcome. (source)

Deep learning (neural networks) create intermediate representations via **stacked layers** at the cost of increased training data (source). Thereby

Shearer (2000)

1.1 Problem Situation

As can be seen in Figure 1 . . .

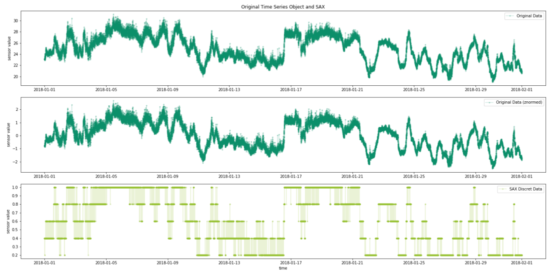

Figure 1: Sax approximation of a time series

1.2 Objectives

1.3 Methods

1.4 Structure

1.5 Tables

Table 1 shows an example table.

Table 1: This is a table

Column 1	Column 2	Column 3
A	B	C
D	E	F
G	H	I

1.6 Source Code

Listing 1: Hello World in Java

```java
public class Hello {
    public static void main(String[] args) {
        System.out.println("Hello World");
    }
}
```

Listing 1 shows the classic Hello World in Java.

Listing 2: Hello World in Python

```python
# This is a comment
print('Hello World')
```

Listing 2 shows the classic Hello World in Python.

Listing 3: Hello World in JavaScript

```javascript
function hello() {
    console.log('Hello World');
}

hello();
```

Listing 4: Hello World in JavaScript (ES6)

```
const hello = async () => {
    await console.log('Hello World');
}

hello();
```

2. Synthetic Data in Privacy

cite –> paper from source, different models on synthetic data!

2.1 Synthetic Data for model performance

When training nns for image classification, (source) a common practice is **data augmentation**, a range of random transformation applied to images in order to synthetically increase the breadth of data that the model is exposed to. Such operations include

- rotation
- shearing
- zoom
- height & width shift

effectively, these operations transform an Image while preserving the underlying signals in the data. However, with other types of data this might be possible. Attributes of another dataset may not be feasibly 'shifted' in one direction or another without fundamentally changing the signal and misleading the model.

note - the infeasibility of pretraining on non-image datasets - representations of the visual world

2.2 Deep Learning

3. Comparison with other solutions to the small data problem

- synthetic minority oversampling technique (SMOTE)
- crossvalidation (k-fold, single holdout)
- transfer learning (word embeddings, image filter layers)
- wholesale synthetic data approaches, Hittmeir et al. (2019) **more sources needed**

4. Results

5. Discussion

Bibliography

Hittmeir, M., Ekelhart, A., and Mayer, R. (2019). On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, ARES '19, New York, NY, USA. Association for Computing Machinery.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

A. List of Interview Partners

B. Code Table