Contents

Chapter 1

Introduction

In the last years, the amount of available data has been increasing permanently. Companies in most industries started to realise that their data contains a lot of useful information and that they can use it to optimise their processes. Also research benefits a lot from the increasing availability of data. Machine learning algorithms use data to learn various tasks, e.g. how to recognize a person by their face. These algorithms not only do well learning such tasks, but they even start performing better than humans. An example that shows the increasing amount of available data is the amount of websites as shown in figure 1.1. It took the world-wide web 23 years (1989-2012) to reach one billion websites where the next billion websites only needed six years (2012-2018). Another domain where similar changes can be observed is image data, that is accessible though different platforms such as Facebook and Instagram. By having a large amount of data, machine learning algorithms can perform very well.

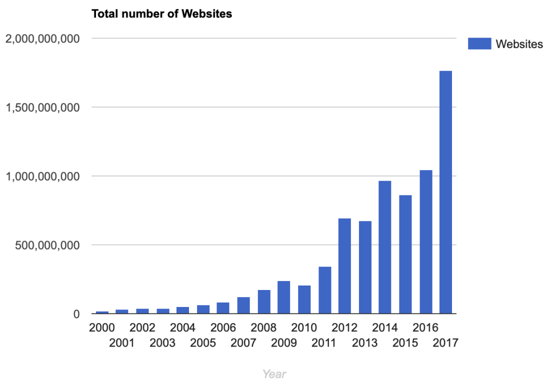

FIGURE 1.1: increasing amount of websites ...

However, one problem of many state-of-the-art machine learning algorithms is that they can solve one task well, but only this task. Imagine a classifier that can distinguish between different animals. It may have learned to perform very well and can differentiate different animals such as a leopard and a tiger. The classifier learned some certain representations that identify the animals occurately. Nevertheless, the classifier only learned to describe animals. An image such as the one shown in 1.2 may be able to trick the classifier by being evaluated as a leopard.

Figure 1.2: not a leopard

Where we can give general information about existing algorithms, their runtime and their accuracy, this thesis aims to design algorithms that are adapted to certain text and image data and thus are able to perform more efficiently on the data than existing general solutions. Also, these algorithms are designed to be adaptable to other clustering tasks within the same data domain.

The here proposed algorithms perform clustering tasks, i.e. they divide existing data into different subspaces. Imagine different animals, one potential subspace could be pets where on the other side there are wild animals. The proposed clustering algoriths belong to a specific family that will be introduced in chapter 2.

Chapter 2

Background Theory

2.1 Data-driven Algorithm Design

This increasing amount of data allows us improve the learning capabilities of machines. We know how well existing algorithms perform for any kind of data and which runtime guarantees they have. However, the algorithms' guarantees are general observations and can vary a lot between different data. In many real-world applications the data does not vary that much, e.g. the data for clustering websites into different types may vary quite much on a yearly base, but as this task gets exectued thousands of times each second for certain search algorithms, the data will not change much. By assuming a static context, it is then possible to leverage the context to improve the algorithmic results, e.g. say you want to cluster person data for different genders. By having this a-priori information, you can use a k-means clustering algorithm with $k = 3$ in order to differentiate between female, male and non-binary people.

However, such observations are mostly not that trivial and often require more effort in order to obtain useful a-priori information. In order to cluster financial standing, one could imagine seeing different clusters depending on the age or the education. But how many clusters would result here? The data has to be processed and evaluated for different values in this case.

2.2 Transfer Learning

Once our algorithm performs well for our data and our tasks, we then want to transfer the gained knowledge to different tasks. Say the algorithm already learned how to differentiate images of the handwritten digits zero, one and too, the same algorithm should then be able to apply the gained knowledge to distinguish between other handwritten digits too.

2.3 Linkage-based hierarchical clustering

2.3.1 Single Linkage

2.3.2 Complete Linkage

2.3.3 Average Linkage

Chapter 3

Related Work

Chapter 4

α-Linkage

We define α as the parameter with which the output of an algorithm is weighted. In this chapter we propose different distance measures depending on the weight parameter α.

4.1 Linear Interpolation between two different linkage strategies

In the first setting we are using the single linkage distance $d_{SL}(X, Y)$ and the complete linkage distance $d_{CL}(X, Y)$. By combining the two distances we can create a linear model that ranges from $\alpha = 0$ (single linkage) to $\alpha = 1$ (complete linkage) resulting in equation 4.1.

$$\begin{aligned} d_{SC}(X, Y, \alpha) &= (1 - \alpha) \cdot d_{SL}(X, Y) + \alpha \cdot d_{CL}(X, Y) \\ &= (1 - \alpha) \min_{x \in X, y \in Y} d(x, y) + \alpha \max_{x \in X, y \in Y} d(x, y) \end{aligned} \tag{4.1}$$

4.2 Linear Interpolation between three different linkage strategies

4.3 Proposed Algorithms

4.4 Performance Optimizations

Chapter 5

α-β-Linkage

5.1 Bilinear Interpolation between three different linkage strategies

5.2 Adapted Algorithm

Chapter 6

Experimental Setup

6.1 Datasets

6.1.1 NELL data

6.1.2 MNIST handwritten digits

6.1.3 CIFAR-10

6.1.4 CIFAR-100

6.2 Cost functions

6.2.1 Majority Cost

6.2.2 Hamming Distance

6.3 Experiments

Chapter 7

Results and Discussion

Chapter 8

Conclusion

Appendix A

An Appendix

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus at pulvinar nisi. Phasellus hendrerit, diam placerat interdum iaculis, mauris justo cursus risus, in viverra purus eros at ligula. Ut metus justo, consequat a tristique posuere, laoreet nec nibh. Etiam et scelerisque mauris. Phasellus vel massa magna. Ut non neque id tortor pharetra bibendum vitae sit amet nisi. Duis nec quam quam, sed euismod justo. Pellentesque eu tellus vitae ante tempus malesuada. Nunc accumsan, quam in congue consequat, lectus lectus dapibus erat, id aliquet urna neque at massa. Nulla facilisi. Morbi ullamcorper eleifend posuere. Donec libero leo, faucibus nec bibendum at, mattis et urna. Proin consectetur, nunc ut imperdiet lobortis, magna neque tincidunt lectus, id iaculis nisi justo id nibh. Pellentesque vel sem in erat vulputate faucibus molestie ut lorem.

Quisque tristique urna in lorem laoreet at laoreet quam congue. Donec dolor turpis, blandit non imperdiet aliquet, blandit et felis. In lorem nisi, pretium sit amet vestibulum sed, tempus et sem. Proin non ante turpis. Nulla imperdiet fringilla convallis. Vivamus vel bibendum nisl. Pellentesque justo lectus, molestie vel luctus sed, lobortis in libero. Nulla facilisi. Aliquam erat volutpat. Suspendisse vitae nunc nunc. Sed aliquet est suscipit sapien rhoncus non adipiscing nibh consequat. Aliquam metus urna, faucibus eu vulputate non, luctus eu justo.

Donec urna leo, vulputate vitae porta eu, vehicula blandit libero. Phasellus eget massa et leo condimentum mollis. Nullam molestie, justo at pellentesque vulputate, sapien velit ornare diam, nec gravida lacus augue non diam. Integer mattis lacus id libero ultrices sit amet mollis neque molestie. Integer ut leo eget mi volutpat congue. Vivamus sodales, turpis id venenatis placerat, tellus purus adipiscing magna, eu aliquam nibh dolor id nibh. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Sed cursus convallis quam nec vehicula. Sed vulputate neque eget odio fringilla ac sodales urna feugiat.

Phasellus nisi quam, volutpat non ullamcorper eget, congue fringilla leo. Cras et erat et nibh placerat commodo id ornare est. Nulla facilisi. Aenean pulvinar scelerisque eros eget interdum. Nunc pulvinar magna ut felis varius in hendrerit dolor accumsan. Nunc pellentesque magna quis magna bibendum non laoreet erat tincidunt. Nulla facilisi.

Duis eget massa sem, gravida interdum ipsum. Nulla nunc nisl, hendrerit sit amet commodo vel, varius id tellus. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc ac dolor est. Suspendisse ultrices tincidunt metus eget accumsan. Nullam facilisis, justo vitae convallis sollicitudin, eros augue malesuada metus, nec sagittis diam nibh ut sapien. Duis blandit lectus vitae lorem aliquam nec euismod nisi volutpat. Vestibulum ornare dictum tortor, at faucibus justo tempor non. Nulla facilisi. Cras non massa nunc, eget euismod purus. Nunc metus ipsum, euismod a consectetur vel, hendrerit nec nunc.

Bibliography