KIT
Karlsruhe Institute of Technology

Carnegie
Mellon
University

Data-Driven Learning of Clustering Algorithms for Image and Text Data

Master’s Thesis of

Manuel Lang

at the Department of Informatics
Humanoids and Intelligence Systems Lab

Reviewer:	Prof. Dr. Rüdiger Dillmann
Second reviewer:	
Advisor:	

1. January 2019 – June 4, 2019

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

Baden-Württemberg
STIPENDIUM®

This thesis was written during an exchange at Carnegie Mellon University in Pittsburgh (Pennsylvania) and was kindly supported by the Baden-Württemberg Stipendium.

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Karlsruhe, June 4, 2019

. . . .

(Manuel Lang)

Acknowledgments

First, I would like to thank Prof. Dr. Rüdiger Dillmann for recommending me for the InterACT program and supporting me throughout my work.

In addition, I would like to thank Prof. Dr. Marina-Florina Balcan for supervising me during my stay at Carnegie Mellon University where she gave me the opportunity to work on very interesting research topics. Also, Nina supported me with insightful discussions, very helpful guidance and the option to bring in my own ideas and wishes. After my stay at CMU, Nina helped me to wrap up the project to have everything necessary to write this thesis and also submit this work as a contribution to NeurIPS 2019.

My thanks also go to the Automated Algorithm Reading Group and Nina's Learning Theory Group, where we had a lot of interesting discussions about state-of-the-art research that allowed me to learn a lot during my stay at CMU. Especially, I would like to thank Travis Dick, who supported me a lot during the implementation of the framework, the theoretical part of this work and also by finding interesting ideas to apply the introduced algorithms.

Last but not least, I would like to thank my family who supported me not only during my studies.

Contents

List of Figures

List of Tables

1. Introduction

Unsupervised grouping is used in various applications to categorize data observations into similar regions. As an example, similar documents can be combined into clusters so that for a new document or a search query, a list of corresponding documents can be shown [1]. The same procedure can also be applied for different tasks such as grouping products [2], searching images [3] or detecting anomalies [4]. In comparison to supervised learning, the data does not have to be (completely) annotated, i.e. potentially expensive labeling work can be avoided by using clustering algorithms.

As the amount of available data has been increasing in the past [5], data analysis is more often required for some specific use-case that includes a very specific dataset. State-of-the-art algorithms mostly provide general complexity- and runtime-guarantees, thus worst-case guarantees have to be assumed for the given dataset. However, as large datasets often don't adapt much over time, it is very likely that also runtime and complexity of certain algorithms applied on the given data will not change much. On the other hand, it is not trivial which algorithm can then be used to obtain the optimal results, i.e. the optimal clusters of the given data.

In addition, data is often split into different natural representations. For instance, images on websites can be seen as a matrix of pixels, but visually impaired people would rather use the image's alternative text description. For machine learning experiments it can be difficult to create a model based on various representation as it does not seem to be natural how to stack different data sources such as pixels and alternative texts.

This thesis proposes several algorithms to efficiently use a linear combination of clustering algorithms to overcome the hurdle of selecting the proper algorithm for the given data. In addition, the framework this algorithm is built in[1] will be also be applied on learning a weighted linear combination of feature representations. The proposed clustering algorithms belong to a specific family that will be introduced in chapter 2.

[1] The implementation is published open-source, see `https://github.com/manu183/TODO`.

2. Background Theory

2.1. Data-driven Algorithm Design

Explanation of best algorithm selection

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 executed 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.

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 two, the same algorithm should then be able to apply the gained knowledge to distinguish between other handwritten digits too. The gained knowledge is some kind of learned data, that can for example be the feature representation of a Convolutional Neural Network, where a potential goal can be to transfer the representation knowledge to another classification task.

For clustering tasks learned knowledge could be a number of clusters, a good feature representation for the input data or other useful information that allows performing similar clustering tasks better by transferring the knowledge.

2.2. Linkage-based hierarchical clustering.

This thesis focuses on agglomerative hierarchical clustering, i.e. clustering algorithms that merge clusters starting from each cluster as its own point until all points belong to the same cluster. At each iteration the clusters with the closest distance are merged together. As there are various clustering algorithms, there also are various distance measurements. One way of describe the distance between two clusters, say X and Y is by defining a linkage between them. There are three different methods to do so.

Single Linkage. Single linkage defines a distance between two clusters X and Y as the distance between the two nearest points of these clusters (see equation 2.1).

$$d_{SL}(X,Y) = \min_{x \in X, y \in Y} d(x,y) \tag{2.1}$$

Complete Linkage. Complete linkage defines a distance between two clusters X and Y as the distance between the two farthest points of these clusters (see equation 2.2).

$$d_{CL}(X,Y) = \max_{x \in X, y \in Y} d(x,y) \tag{2.2}$$

Average Linkage. Average linkage defines a distance between two clusters X and Y as the average distance between all points $x \in X$ and all points $y \in Y$ (see equation 2.3).

$$d_{AL}(X,Y) = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d(x,y) \tag{2.3}$$

Effects of different linkage strategies. Depending on the linkage strategy, the pairwise distances between all N clusters $C_1, ..., C_N$ will be different. As the clustering algorithm merges the closest pair of clusters in each iteration, the merging clusters C_i and C_j with $i, j \in 1, ..., N$ might vary as shown in figure 2.1, where ten clusters $C_0, ..., C_9$ get clustered with bottom-up hierarchical clustering using the Euclidean distance as distance $d(x,y)$ to calculate the pairwise distance according to the three mentioned linkage strategies.

As different points are merged together, this also means that the clustering may have a different quality. This thesis compares the clusterings' quality for different data by introducing algorithms to efficiently determine the quality not only for these linkage strategies but also for their linear combinations.

2.3. Generating Feature Representations

In order to improve overall the clustering performance, this work uses several techniques to obtain better feature representations of text and image data.

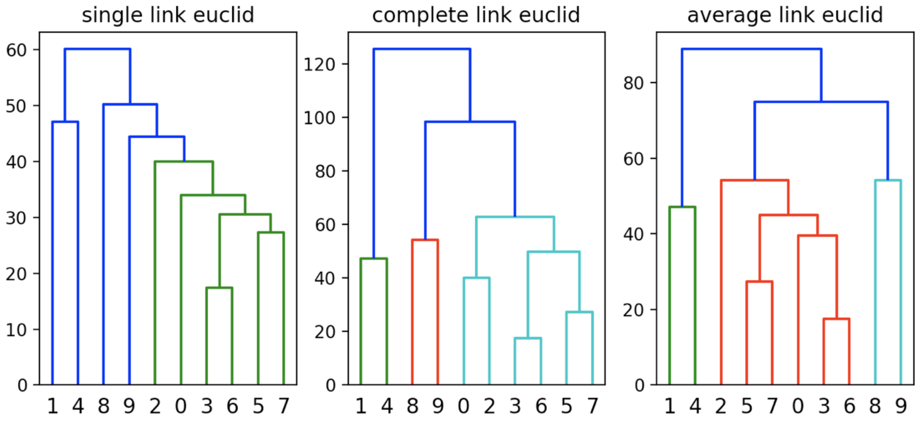

Figure 2.1.: Different distance measurements often result in different merges for bottom-up hierarchical clustering algorithms. The three discussed linkage strategies result in three different clusterings.

2.3.1. Text Features

To cluster different words, there are several ways to create a difference measurement of these words. A rather simple approach would be to calculate the edit distance that describes the difference of the characters in the words [6]. However, this approach does contain any semantic information, i.e. synonyms will have a larger distance than wanted. This motivates to leverage contextual information, where we use several pre-trained models on different datasets. Stanford's GloVe provides such models that incorporate knowledge from Wikipedia and social networks [7].

In addition, CMU's Machine Learning Department provides another way to compare words. Their Never-Ending Language Learner provides information in which contexts certain words are used online [8]. We can create a corpus containing all different contexts. Similar to the bag-of-words approach, we can then count the occurrences of the words in the corpus' contexts. However, the resulting data is very sparse and Euclidean distance will not work well to compare the "bag-of-context" representations, i.e. other measurements such as the cosine distance are preferred.

2.3.2. Image Features

3. Related Work

Explain [9].

Explain [10].

Explain [11].

Balcan et al. proposed the two infinite families to interpolate between different linkage strategies [12], such as shown in equations 3.1 and 3.2.

$$\mathcal{A}_1 = \left\{ \left(\min_{u \in A, v \in B} (d(u,v))^\alpha + \max_{u \in A, v \in B} (d(u,v))^\alpha \right)^{1/\alpha} \middle| \alpha \in \mathbb{R} \cup \{\infty, -\infty\} \right\} \tag{3.1}$$

Equation 3.1 shows a distances in the range between single linkage ($\alpha = -\infty$) and complete linkage ($\alpha = \infty$). Balcan et al. also show that $\mathbb{R} \cup \{\infty, -\infty\}$ contains a maximum of $O(n^8)$ different intervals, where each interval $[\alpha_{lo}, \alpha_{hi}]$ represents a different merging behavior.

$$\mathcal{A}_2 = \left\{ \left(\frac{1}{||A||\,||B||} \sum_{u \in A, v \in B} (d(u,v))^\alpha \right)^{1/\alpha} \middle| \alpha \in \mathbb{R} \cup \{\infty, -\infty\} \right\} \tag{3.2}$$

Equation 3.2 will also result in single linkage for $\alpha = -\infty$ and complete linkage for $\alpha = \infty$. In addition to that, the family $\mathcal{A}_2$ also contains the definition of average linkage ($\alpha = 0$). However, the guarantee for maximum $O(n^8)$ intervals does not apply to this family. A formal guarantee will be $O(n^4 2^n)$, but this thesis will show that the experimental results are much better than the actual formal guarantee.

Balcan et. al also provided a solution to calculate all different merges, however this approach solved the mathematical equations and led to the same clusters being used for a merge quite a often. As our solution only evaluates cases where different clusters get merged, the algorithm described in the following section has a lower runtime as well as a lower lower complexity.

4. Efficient Algorithm Selection

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 that allows us interpolating between different linkage strategies in a way similar to the proposed by Balcan et al. [12], where a infinite interval was proposed.

To have a real application, we need to have a finite set of intervals. So we interpolate between one algorithm with $\alpha = 0$ and another algorithm with $\alpha = 1$, where for $\alpha = 0$ the result will be the result of algorithm 1 and for $\alpha = 1$ the result of algorithm 2.

4.1. Linear Interpolation between two different linkage strategies

Interpolating between two of the three mentioned linkage strategies results in three different algorithmic settings. 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.

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

Equivalently we can interpolate between the single linkage distance $d_{SL}(X, Y)$ and the average linkage distance $d_{AL}(X, Y)$ instead of the complete linkage distance $d_{CL}(X, Y)$ for $\alpha = 1$ resulting in equation 4.2.

$$\begin{aligned} d_{SA}(X, Y, \alpha) &= (1-\alpha) \cdot d_{SL}(X, Y) + \alpha \cdot d_{AL}(X, Y) \\ &= (1-\alpha) \min_{x \in X, y \in Y} d(x, y) + \alpha \frac{1}{\|X\| \|Y\|} \sum_{x \in X, y \in Y} d(x, y) \end{aligned} \tag{4.2}$$

The last of the three settings describes the interpolation between the average linkage distance $d_{AL}(X, Y)$ and the complete linkage distance $d_{CL}(X, Y)$ resulting in equation 4.3.

$$\begin{aligned} d_{AC}(X, Y, \alpha) &= (1-\alpha) \cdot d_{AL}(X, Y) + \alpha \cdot d_{CL}(X, Y) \\ &= (1-\alpha) \frac{1}{\|X\| \|Y\|} \sum_{x \in X, y \in Y} d(x, y) + \alpha \max_{x \in X, y \in Y} d(x, y) \end{aligned} \tag{4.3}$$

4.2. Proposed Algorithms

Our goal is to find an algorithm that determines all different behavior depending on different values of α. To do so, we propose different algorithms. The goal of the first algorithm is to divide the interval of $\alpha \in [\alpha_{lo}, \alpha_{hi}]$ to subintervals where the behavior is consistent within each interval.

```
Data: input data $p_1, ..., p_N$, initial states $st$
Result: $k$ intervals $[\alpha_0, \alpha_1], ..., [\alpha_{k-1}, \alpha_k]$
for $iteration \leftarrow 1$ to $N - 1$ do
    foreach $state\ s \in st$ do
        remove state $s$;
        $cand_1, cand_2 \leftarrow$find merge candidates for $s.\alpha_{lo}$ and $s.\alpha_{hi}$;
        if $cand_1 == cand_2$ then
            $ms \leftarrow$ merge $cand_1$;
            add state $ms$ with interval $[\alpha_{lo}, \alpha_{hi}]$ to the end of $st$;
        else
            $\alpha_{split} \leftarrow$ calculate split;
            $s_1 \leftarrow$ merge $cand_1$;
            $s_2 \leftarrow$ merge $cand_2$;
            add state $s_1$ with interval $[\alpha_{lo}, \alpha_{split}]$ to the end of $st$;
            add state $s_2$ with interval $[\alpha_{split}, \alpha_{hi}]$ to the end of $st$;
        end
    end
end
```

Algorithm 1: We calculate all from splits resulting different intervals between α_{lo} and α_{hi}, merge the resulting clusters and do so until each state contains only one cluster with all points.

Starting from an interval $\alpha \in [\alpha_{lo}, \alpha_{hi}]$, we calculate the merging clusters by $\min_{X,Y} d(X, Y, \alpha)$ for both the minimum α_{li} and the maximum α_{hi} of the interval. In case both values of α return the same pair of merging clusters X and Y, we merge X and Y. In case the values of α_{lo} and α_{hi} lead to different merges, we can calculate a value α_{split} where we know that for values of $\alpha \in [\alpha_{lo}, alpha_{split})$ we merge the clusters found for $\min_{X,Y} d(X, Y, \alpha_{lo})$ and for values of $\alpha \in [\alpha_{split}, alpha_{hi}]$ we merge the clusters found for $\min_{X,Y} d(X, Y, \alpha_{hi})$. In order to calculate the value of α_{split} we can equalize the distance functions of the merges of clusters X, Y and the clusters A, B (say α_{lo} leads to merging X and Y and α_{hi} leads to merging A and B) as seen in equation 4.4.

$$d(X, Y, \alpha_{split}) = d(A, B, \alpha_{split}) \tag{4.4}$$

Applying 4.4 to a concrete example of distance functions leads to a concrete calculation for the value of α_{split}. Equation 4.5 shows the calculation for the in equation 4.1 introduced d_{SC}.

$$
\begin{gathered}
d_{SC}(X, Y, \alpha_{split}) = d_{SC}(A, B, \alpha_{split}) \\
(1 - \alpha_{split}) \min_{x \in X, y \in Y} d(x, y) + \alpha_{split} \max_{x \in X, y \in Y} d(x, y) = \\
= (1 - \alpha_{split}) \min_{a \in A, b \in B} d(a, b) + \alpha_{split} \max_{a \in A, b \in B} d(a, b) \\
(-\alpha_{split}) \min_{x \in X, y \in Y} d(x, y) + \alpha_{split} \max_{x \in X, y \in Y} d(x, y) + \alpha_{split} \min_{a \in A, b \in B} d(a, b) - \alpha_{split} \max_{a \in A, b \in B} d(a, b) = \\
= - \min_{x \in X, y \in Y} d(x, y) + \min_{a \in A, b \in B} d(a, b) \\
\alpha_{split}(- \min_{x \in X, y \in Y} d(x, y) + \max_{x \in X, y \in Y} d(x, y) + \min_{a \in A, b \in B} d(a, b) - \max_{a \in A, b \in B} d(a, b)) = \\
= - \min_{x \in X, y \in Y} d(x, y) + \min_{a \in A, b \in B} d(a, b) \\
\alpha_{split} = \frac{- \min_{x \in X, y \in Y} d(x, y) + \min_{a \in A, b \in B} d(a, b)}{- \min_{x \in X, y \in Y} d(x, y) + \max_{x \in X, y \in Y} d(x, y) + \min_{a \in A, b \in B} d(a, b) - \max_{a \in A, b \in B} d(a, b)}
\end{gathered}
\tag{4.5}
$$

After knowing the exact consistent range, we split the clusters for the different states and then calculate the merge candidates for the start and the end of the new intervals again. We can show the different possible merges in a tree of executions, where each node represents one interval $[\alpha_{lo}, \alpha_{hi}]$ where the same clusters get merged (see figure 4.1).

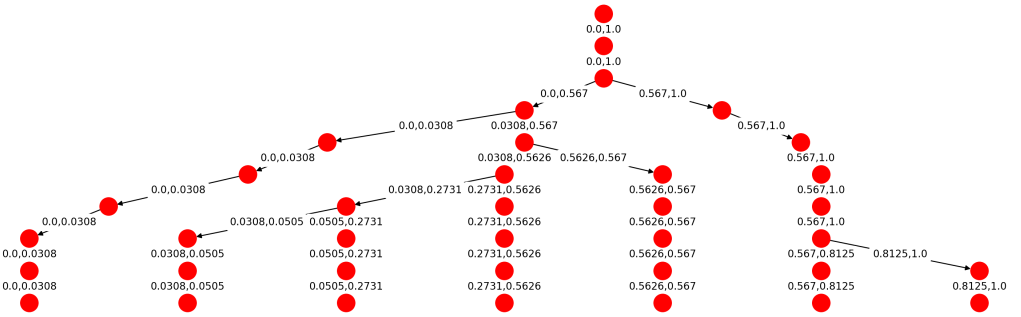

Figure 4.1.: Different values of α lead to different merges. We calculate a tree where we start with the entire range $[\alpha_{lo}, \alpha_{hi}]$ and split the interval into all different subintervals with consistent merges.

We perform the described procedure for iterations $i = count(points) - 1$ times until only one cluster containing all points is left. All the leaf nodes in the resulting tree of executions then represent one interval $[\alpha_{lo}, \alpha_{hi}]$ where the clustering is consistent within the interval and each interval contains a different clustering.

However just calculating the split values in a range $[\alpha_{lo}, \alpha_{hi}]$ does not necessarily yield to the best possible solution. One of these examples is demonstrated in figure 4.2, where the blue line for the constant value of d will not be considered, only the lines $\alpha \in [\alpha_{lo}, alpha_{split})$ (red) and $\alpha \in [\alpha_{split}, alpha_{high}]$ (black) will be.

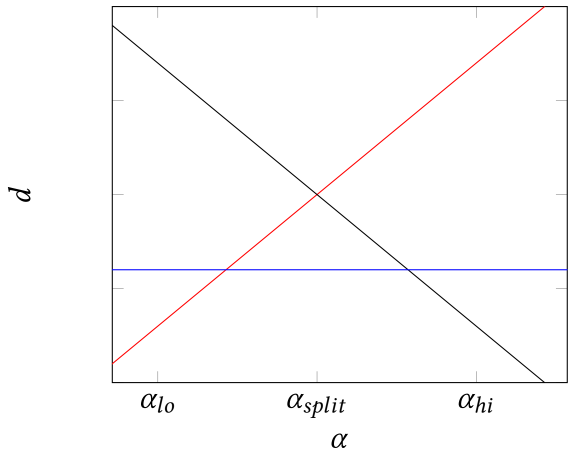

Figure 4.2.: Simply calculating the split values between the start and the end value of the range $[\alpha_{lo}, \alpha_{hi}]$ will not necessarily lead to the optimal values. By doing so, the blue line (constant d value) will not be considered.

In order to solve this, we can recursively check each resulting interval again if it contains different merging behaviors.

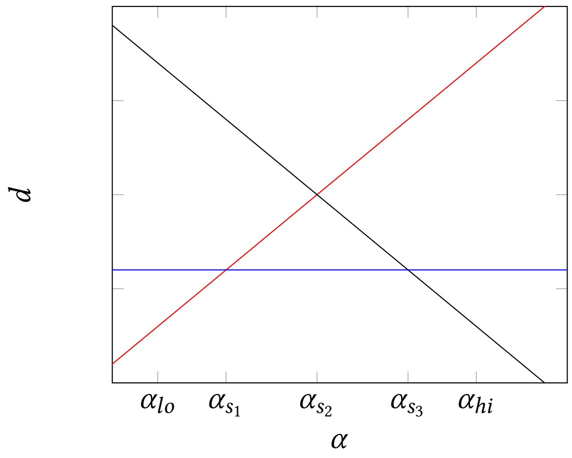

Figure 4.3.: Simply calculating the split values between the start and the end value of the range $[\alpha_{lo}, \alpha_{hi}]$ will not necessarily lead to the optimal values. By doing so, the blue line (constant d value) will not be considered.

By calculating the split points recursively, the example in figure 4.3 will result in the intervals $[\alpha_{lo}, \alpha_{s_1}]$, $[\alpha_{s_1}, \alpha_{s_2}]$, $[\alpha_{s_2}, \alpha_{s_3}]$ and $[\alpha_{s_3}, \alpha_{s_{hi}}]$. The optimal distance between α_{s_1} and α_{s_3} is covered now, but the results contain one unncessary interval as α_{s_2} still splits two intervals. The algorithm can check if older splits are still relevant, however the runtime cost to do so will be more expensive than carrying one additional interval with the same distance. We can use this knowledge and adapt algorithm 1.

```
Data: input data $p_1, ..., p_N$, initial states $st$
Result: $k$ intervals $[\alpha_0, \alpha_1], ..., [\alpha_{k-1}, \alpha_k]$
for $iteration \leftarrow 1$ to $N - 1$ do
    foreach state $s \in st$ do
        remove state $s$;
        ranges $\leftarrow$ find ranges between $s.\alpha_{lo}$ and $s.\alpha_{hi}$;
        foreach range $r \in ranges$ do
            $cand \leftarrow$ candidate for range;
            $ms \leftarrow$ merge $cand$;
            add state $ms$ with range $r$ to the end of $st$;
        end
    end
end
```

Algorithm 2: By calculating the split points between α_{lo} and α_{hi} recursively, we ensure that no optimal interval is left out.

As experimental results turn out to need a lot of memory (up to ≈ 20 GB for 300 points and 20,000 states), we want to adapt algorithm 2 so that it uses less memory. The memory usage scales relative to the amount of currently in-memory stored states, so the goal is to reduce these. As the amount of states is much larger than the amound of iterations, we calculate and evaluate the leave nodes of the tree and keep the alternative merges stored. This results in algorithm 3.

```
Data: input data $p_1, ..., p_N$, initial states $st$
Result: $k$ intervals $[\alpha_0, \alpha_1], ..., [\alpha_{k-1}, \alpha_k]$
while $||st|| > 0$ do
    foreach state $s \in st$ do
        remove state $s$;
        if $s$ is final then
            evaluate $s$;
        else
            ranges $\leftarrow$ find ranges between $s.\alpha_{lo}$ and $s.\alpha_{hi}$;
            foreach range $r \in ranges$ do
                $cand \leftarrow$ candidate for range;
                $ms \leftarrow$ merge $cand$;
                add state $ms$ with range $r$ to the beginning of $st$;
            end
        end
    end
end
```

Algorithm 3: Instead of calculating the nodes layerwise, this algorithm works pathwise, i.e. it goes down one path of a tree to a leaf node and evaluates it before continuing with the next split. This approach needs much less memory than the previous algorithms and has about the same runtime as shown in figure 4.4.

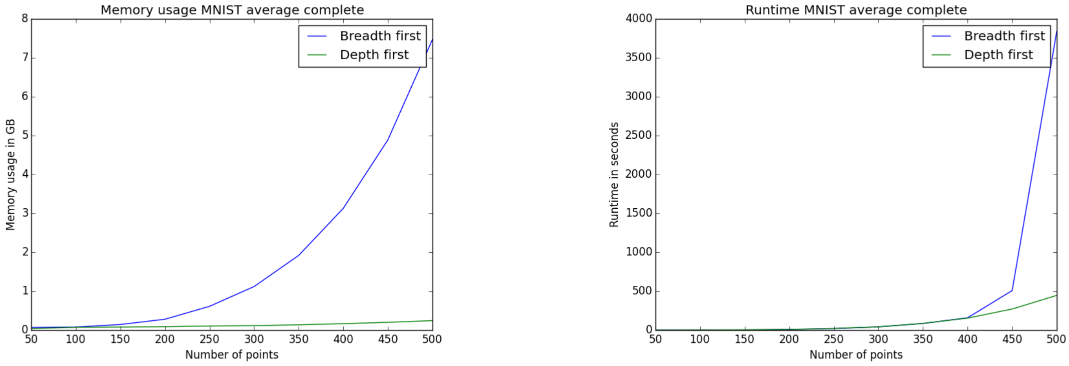

Figure 4.4.: The depth first implementation needs less memory and also has a better runtime compared to the breadth first implementation.

Instead of merging iteratively and steadily shrinking the intervals, we propose an algorithm with a geometric motivation. We are again evaluating an interval $[\alpha_{lo}, \alpha_{hi}]$, but we interprete the different merges as linear functions depending on α. We can start by calculating the merge candidate for the start value α_{lo} and calculate the next intersection that will yield to the next merge. By calculating all the intersections of linear functions, we can also determine all the different intervals for the range $[\alpha_{lo}, \alpha_{hi}]$, where different merging behaviors occure. Algorithm 4 describes this procedure.

```
Data: input data $p_1, ..., p_N$, start value $\alpha_{lo}$, end value $\alpha_{hi}$
Result: $k$ intervals $[\alpha_0, \alpha_1], ..., [\alpha_{k-1}, \alpha_k]$
$\alpha \leftarrow \alpha_{lo}$;
linear function $lf \leftarrow$ get lf for alpha;
while $\alpha < \alpha_{hi}$ do
    $\alpha_{new} \leftarrow$ calculate next split for $\alpha$;
    $lf \leftarrow$ get lf for $\alpha_{new}$;
    $\alpha \leftarrow \alpha_{new}$
end
```

Algorithm 4:

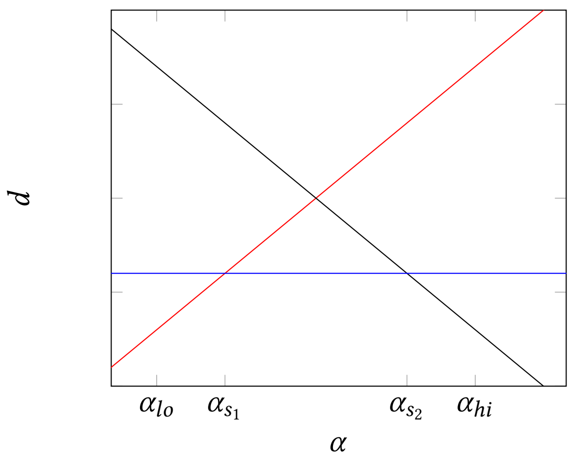

Figure 4.5.: Simply calculating the split values between the start and the end value of the range $[\alpha_{lo}, \alpha_{hi}]$ will not necessarily lead to the optimal values. By doing so, the blue line (constant d value) will not be considered.

4.3. Performance Optimizations

In order to have real-world applications, the proposed algorithms should run in an efficient way, i.e. it should not take the α-linkage algorithms too much time to run. A first python implementation took days to run, but switching to C++ and using its advantages took down the runtime to hours. However, there are more optimization methods that we used in order to improve the runtime.

4.3.1. Dynamic Programming

One of the most time-consuming parts was the calculating of the distances. For each pair of clusters C_i, C, j the distance had to be calculated for each clustering state. We optimized this by using dynamic programming and stored the distance matrices D_{lower} and D_{upper} for each state. The naming results from the different interpolation settings where we interpolate from one linkage distance (lower) to another linkage distance (upper), e.g. the setting in equation 4.1 describes the interpolation from single linkage (lower) to complete linkage (upper). In this example we then store the pairwise distances for both single linkage and complete linkage and in order to find the merge candidates we have to do iterate over the distance matrices instead of calculating the distances over and over again. When we merge two clusters, we then update the distance matrices for the given state. Table 4.1 shows an example for the pairwise distances of clusters i and j.

j\i	0	1	2	3	4
0	0	1.243	1.512	2.468	5.1243
1	1.243	0	2.443	3.1412	4.443
2	1.512	2.443	0	3.8988	6.827
3	2.468	3.1412	3.8988	0	5.72
4	5.1243	4.443	6.827	5.72	0

Table 4.1.: Storing the pairwise distances of all clusters avoids calculating the distances over and over again.

One observation that we can make is that the matrix has a lot of redundant values, because $D(i, j) = D(j, i)$. Removing these rendundant values will result in a trade-off between copying and indexing costs and will be discussed in the following section. Another optimization we can do is storing the indices of the active clusters, i.e. the clusters that can get merged. Once two clusters got merged, they cannot be merged any further, only the resulting cluster can. So we then do not have to consider the old clusters anymore and can remove them from the set of active indicies. This allows us to find the merge candidates faster as the pool of candidates gets smaller.

4.3.2. Trade-Off between Copying and Indexing Costs

Currently we can access the costs for a pair of clusters C_i and C_j through $D[i, j]$ or $D[i + j * width]$ for flattened matrices. These indices are very easy to calculate. In order to remove the redundant values from the distance matrix we remove all values below the diagonal as shown in table 4.2.

j\i	0	1	2	3	4
0	0	1.243	1.512	2.468	5.1243
1		0	2.443	3.1412	4.443
2			0	3.8988	6.827
3				0	5.72
4					0

Table 4.2.: Storing the pairwise distances of all clusters avoids calculating the distances over and over again.

In addition to that we can also remove the diagonal values as they represent the distances between the same clusters and are thus always zero. This results in table 4.3.

j\i	0	1	2	3	4
0		1.243	1.512	2.468	5.1243
1			2.443	3.1412	4.443
2				3.8988	6.827
3					5.72
4					

Table 4.3.: Storing the pairwise distances of all clusters avoids calculating the distances over and over again.

The matrices are now smaller, so they need less memory. In the example, we changed a matrix of the size 25 to a matrix of the size 10. In general a matrix of the size nxn will be compressed to a matrix of the size $\frac{n^2-n}{2}$. The lower amount of needed memory also results in less copying costs that will lead to a better runtime. However, the indexing is not as easy anymore. For easier storage, we again work with flattened matrices, the indexing for the resulting list is shown in equation 4.6.

$$index(i,j) = \frac{width * (width-1)}{2} - \frac{(width-j) * (width-j-1)}{2} + i - j - 1 \tag{4.6}$$

Calculating this index in a nested loop is very expensive, however we calculate the part that does not depend on i in the outer loop and thus only need to add i in the inner loop. This does not only yield to a lower memory usage of $\approx 30\%$, but also increases the runtime by TODO.

4.3.3. Implementation-specific Optimizations

In order to optimize the implementation even further, we will have a look into the implementation. One optimization that already was briefly described is the flatterning of the matrices, so the resulting list will be one-dimensional and can be iterated easier and faster.

Another observation is that copy operations are computationally expensive, so we avoid them as much as possible. In the described algorithms (1, 2 and 3) we removed a state from the list of states and added other states. In an optimized way, we do not remove the state and just overwrite the state with the resulting state. Once there are splits in the current interval, the state gets overwritten and additional states get added to the list.

We can also optimize the way of updating the distance matrices. Instead of adding new clusters there for a merge of clusters i and j we update the distances of i to all active clusters with the distances of the resulting cluster. The distances of the cluster j will not be considered for merges anymore as the index j gets removed from the active indices. This has the advantage that the size of the distance matrices will not increase after merges.

Also, the data types make an important contribution to the memory usage. Instead of using double precision floating point values, single precision is enough to clearly identify and separate all the resulting intervals. Same goes for the distances as we only need the minimum and maximum distances, that are not effected by loss of precision. To store the indices of the clusters, we know that they will not exceed 2^{16}, so they can be store as half precision values.

5. Optimizing the Metric

In a similar fashion as described in section 4 this sections aims to optimize a metric that is a linear combination of several metrics. For instance, images can have a 2D pixel representation and a text describing the each image. Combining these features for clustering tasks can be problematic as it is not how the optimal weight between these features should be. Does a word describe more than a subset of the image, are the features equally important or does the pixel image lead to better clusterings? With β-linkage we provide a framework based on α-linkage that calculates different merges based on linear combinations of representations and leads to optimized clusterings.

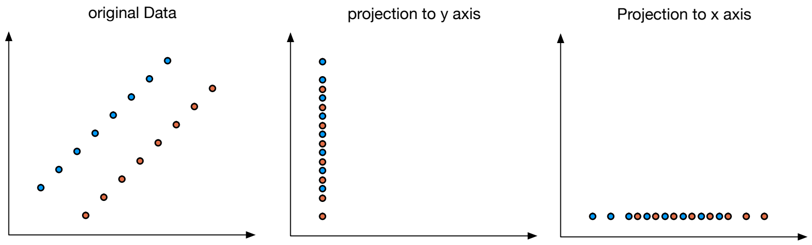

Figure 5.1.: Combining several metrics seems often natural and can lead to improved results as in this example where we project a dataset on both axes.

For instance, figure 5.1 shows a set of points that might be put in clusters easily. However, if you only look at the distance regarding the X_1-axis or the X_2-axis clustering will be very difficult, because each of the axis does not describe the spatial correlation anymore. This example is selected on purpose to motivate the following experiments where we learn optimal combinations of different metrics.

Explain difference to previous section.

6. Experimental Setup

This work evaluates the proposed algorithms for image and text data. This chapter describes the used datasets, the evaluation methods and the experimental setups.

6.1. Datasets

6.1.1. Synthetic Data

To motivate our approach, we manually created a dataset containing disks and rings as shown in figure 6.1. In this case, we know that single linkage performs well clustering the two rings, however it might be problematic to cluster the disks. On the other hand, complete linkage is expected to cluster the disks very well, but it might contact the two rings earlier than wanted. This data motivates our approach of interpolating between different linkage strategies, however the data is not natural and real-world datasets are very likely to have a different structure.

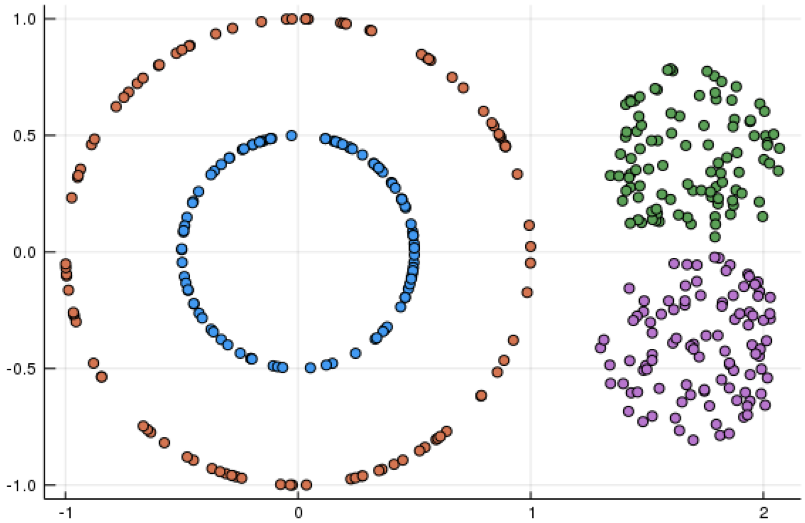

Figure 6.1.: We use disks and rings as a sample dataset to motivate our α-linkage approach. The dataset contains four clusters, two disks and two rings.

6.1.2. Never Ending Language Learner data

The Never Ending Language Learner (NELL) is a learning agent that reads the web, extracts data and verifies beliefs [13][14]. NELL for example knows that "Pittsburgh" is located in

"Pennsylvania". These beliefs represent different noun-phrases such as "Pittsburgh" and "Pennsylvania". The noun-phrases belong to certain categories. "Pittsburgh" is a "City" and "Pennsylvania" is a "State". These subcategories both belong to the main category "Geopolitical Location". While there are already different subcategories, the goal for a hierarchical clustering algorithm here is to extract new useful subcategories.

The used dataset, extracted web-information by NELL, contains 32 different main categories, such as "Animal", "Location" or "Person". Each of these consists of up to 250 different entities that belong to different subcategories. Exemplary entites for the category "Animal" are "Otter", "Squirrel" or "Wolf".

This thesis shows in chapter 7 the learned subcategories.

6.1.3. MNIST handwritten digits

The MNIST handwritten digit database contains images of the handwritten digits from zero to nine [15]. Samples of these images are shown in figure 6.2 Its training set contains a total of 60,000 images, where each image is represented as a 784-dimensional vector corresponding to a greyscale image with 28x28 pixels.

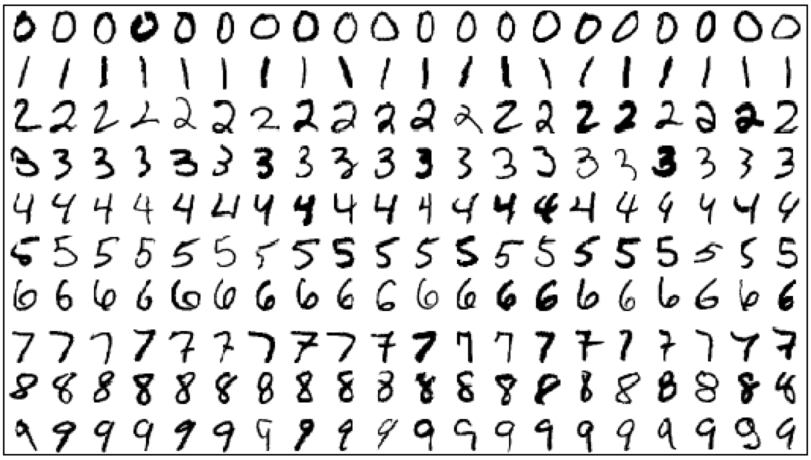

Figure 6.2.: The MNIST handwritten digits database contains 60,000 greyscale images of handwritten digits ranging from zero to nine. These samples show ten randomly drawn samples for each label represented as a 28x28 pixel image [15].

The goal of clustering MNIST images is to find an unsupervised learning method that can distinguish between greyscale images. In addition, we can define various clustering tasks where we pick a subsample of the ten labels and then try to transfer the results to other subsamples. For example, we first cluster images labeled as zero, one, two, three or four and later apply the knowledge the learned gained for clustering images labeled as five, six, seven, eight or nine. Theses types of experiments allow high-level transfer learning if we define several different clustering tasks, e.g. for five different labels there are $\binom{10}{5} = 252$ different combinations of labels.

Another obsevation that results from hierarchical clustering is the similarity of different labels, i.e. which labels are likely to get clustered together.

6.1.4. CIFAR-10

Another image dataset this thesis uses for evaluation is the CIFAR-10 dataset that contains 60,000 RGB images of ten different categories [16]. Each image consists of 32x32 pixels and is thus represented as a 3072-dimensional vector (32x32x3). The categories and ten random images from each are shown in figure 6.3.

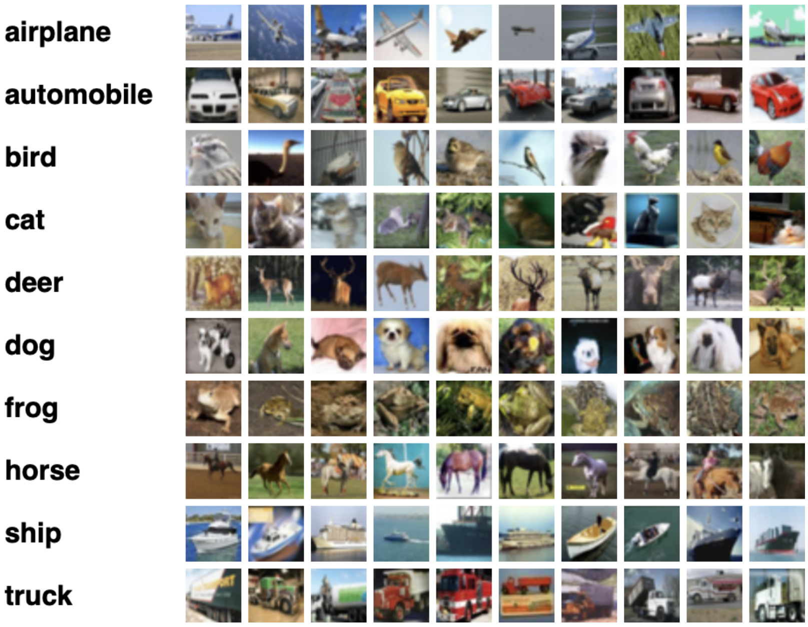

Figure 6.3.: The CIFAR-10 database contains 60,000 RGB images of the ten shown different classes. These samples show ten randomly drawn samples for each label represented as a 32x32 pixel image [16].

As the amount of images and the amount of classes is equal to the ones in the MNIST database, we can also try similar experiments. The main difference is that the images consist of RGB pixels instead of greyscale values.

6.1.5. CIFAR-100

The CIFAR-100 dataset contains similar images, but instead of 6,000 images each for 10 classes, it consists of 600 images each for 100 classes. The classes are divided into 20 superclasses each containing five subclasses. Examples of superclasses and corresponding subclasses are shown in table 6.1.

Having superclasses and subclasses allows clustering between different subclasses within a superclass and also between different superclasses. This allows more experiments than for the CIFAR10 data.

6.1.6. Omniglot

The omniglot dataset contains 1623 handwritten characters from 50 different alphabets, where each character is represented by 20 different images. Each image is grayscale and

superclass	subclasses
aquatic mammals	beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle

Table 6.1.: The CIFAR-100 dataset contains 20 different superclasses, each with five different subclasses leading to 100 classes overall. The images are represented in the same way as in the CIFAR-10 dataset, i.e. by a 3072-dimensional vector [16].

represented by 105x105 pixels [17]. Figure 6.4 shows characters of the more well-known Latin, Greek and Hebrew alphabets that are part of the dataset.

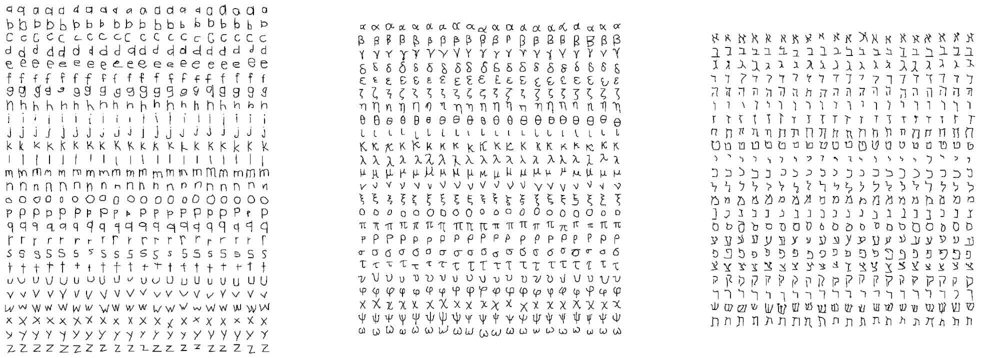

Figure 6.4.: The omniglot dataset contains handwritten characters of different alphabets, such as Latin, Greek and Hebrew [17].

The omniglot dataset is similar to the MNIST dataset as it also contains handwritten characters, however it has more different characters and less images of each of them. This allows us to run more learning tasks.

6.2. Cost functions

In order to evaluate the quality of a clustering, we need some kind of cost function that compares the generated clustering $C_1, ..., C_k$ with the target clustering $C_1^*, ..., C_k^*$. One method to compare them is the majority distance as shown in equation 6.1 where n ist the number of sampled points.

$$cost_{majority}(C_{1:k}, C_{1:k}^*) = \frac{1}{n} \sum_{i=1}^{k} (|C_i| - \max_j |C_i \cap C_j^*|) \tag{6.1}$$

This cost function is motivated by finding corresponding clusters with the lowest distance, i.e. each generated cluster gets matched with the optimal target cluster. However two generated clusters can be matched with the same target cluster. This motivates the hamming distance as shown in figure 6.2.

$$cost_{hamming}(C_{1:k}, C^*_{1:k}) = \max_{\sigma} \frac{1}{n} \sum_{i=1}^{k} (|C_i| - \max_{\sigma,j} |C_i \cap C^*_j|) \tag{6.2}$$

However, the hamming distance consists of an assignment problem to find the optimal matching σ between the generated clusters and the target clusters. Table 6.2 shows how such a matching can look like.

j\i	1	2	3	4	5
1	20	15	30	50	40
2	80	10	15	20	30
3	20	30	50	80	60
4	30	50	40	20	10
5	20	30	40	50	25

Table 6.2.: In order to calculate the hamming distance between two clusterings, we have to calculate the optimal mapping that results in lowest distance for these two clusterings. For random distances between clusterings $C^i_1, \ldots, C^i_k$ and $C^j_1, \ldots, C^j_k$ we can calculate the optimal mapping (blue highlighted cells) in a brute force way or more efficiently with the hungarian method [18][19].

While solving the assignment with a brute force strategy would result in $O(n!)$ complexity, Harold Kuhn introduced the hungarian method to solve the problem in $O(n^4)$ complexity [18]. Later on, James Munkred modified the algorithm to $O(n^3)$ complexity [19]. A detailed explanation of the hungarian method is included in appendix A.

6.3. Experiments

In order to find new subclusters for the NELL data, we cluster each of the 32 main categories seperately. This results in 32 different clustering tasks, where we compare the results of each clustering task with the target labels using the majority distance function. We will receive a cost function $cost(\alpha)$, that shows us for which value of α the resulting clusterings are good, for each category. By averaging all cost functions, we know for which values of α the α-linkage performs well in general. Beside having a value of α that can be used for other clustering tasks, the experiments also give different representation levels of clusters that are discussed in section 7.

To cluster the image data, we set up $\binom{10}{5} = 252$ different experiments by selecting all combinations of five out of the ten labels. In order to do so in efficient time, we subsample the dataset to 60 points for each label, so one experiment will cluster 300 points. By having a fixed set of point, we can show that a certain value of α will lead to good results for the subsampled data. We will use this kind of experiments for all in section **??** mentioned image datasets where all RGB-channels are treated equally for colored images.

In addition to these experiments, we will try to cluster as diverse as possible superclasses of the CIFAR100 dataset by manually picking the five superclasses fish, flowers, household furniture, people and vehicles 1. For each superclass we pick one subclass and evaluate the results for all $5 * \binom{5}{1} = 25$ different combinations of subclasses. In addition to the experiments with $k = 5$ clusters, we compare these results to the results for picking two different subclasses of each superclass ($5 * \binom{5}{2} = 50$ different experiments) resulting in $k = 10$ clusters and also for picking three different subclasses ($5 * \binom{5}{3} = 50$ different experiments) resulting in $k = 15$ clusters.

In comparison to picking as diverse as possible superclasses, we also evaluate the performance for as similar as possible subclasses. Similar subclasses are already given in the dataset through the subclasses within one superclass. We then evaluate the majority and the hamming cost for each superclass and again average the cost over all 20 superclasses to evaluate an optimal value for the parameter α.

6.4. Parameter Advising

The previously described settings average over multiple experiments and show one parameter α that represents the best clustering over all experiments, i.e. the algorithm automatically outputs the best result. For parameter advising, we select the top k values of α for each experiment and calculate the clustering's cost with the best of the k values of α [20]. We select the pool of α-values through the local optima for each experiment. The best k values of α, where k is much smaller than the number of experiments, can then be calculated with an integer optimization problem. A scenario where this setup can be used is by having a domain expert, who can select the best clustering from the k suggested ones.

$$\alpha_1^*, \ldots, \alpha_k^* = \underset{\alpha_1, \ldots, \alpha_k}{\arg\max} \sum_{i=1}^{N} \max_{j \in [k]} u\big(S, T(S, \alpha_j)\big). \tag{6.3}$$

Facility Location Advising

$$\begin{aligned}
\underset{x_{ij}, y_j}{\arg\max} \quad & \sum_{i=1}^{N} \sum_{j=1}^{m} x_{ij} u(S_i, T(S_i, \alpha_j)) \\
\text{subject to} \quad & \sum_{j=1}^{m} y_j = k \\
& \text{for each } i \in [N], \sum_{j=1}^{m} x_{ij} = 1 \\
& \text{for each } i \in [N], j \in [M], x_{ij} \leq y_j.
\end{aligned}$$

Greedy Parameter Advising

$$\sum_{i=1}^{N} \max\{u(S_i, T(S_i, \alpha_1)), u(S_i, T(S_i, \alpha_2)\} \tag{6.4}$$

The results of these experiments are discussed in the following section 7.

7. Results and Discussion

Update plots.

More results.

We evaluated the in chapter 4 proposed algorithms with the in chapter 6.1 discussed datasets aiming to find new subcategories for the text data and to generate better clusterings overall. The quality of the clusterings was calculated with the in chapter 6.2 explained cost functions.

7.1. Clustering Text Data

We subsampled the NELL data to a maximum of 250 points in each class and then evaluated each of the 32 classes separately. Figure 7.1 shows the resulting majority distances for the three different types of linear interpolation.

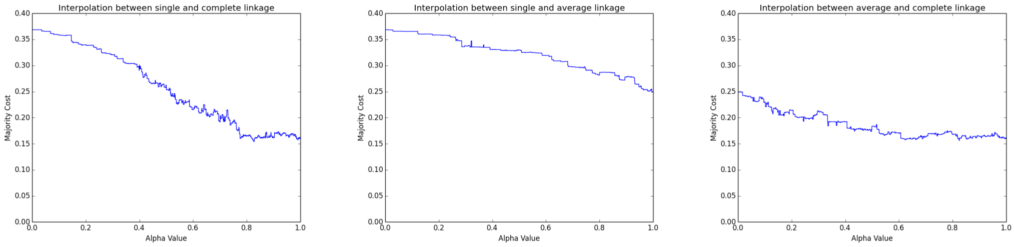

Figure 7.1.: The omniglot dataset contains handwritten characters of different alphabets, such as Latin, Greek and Hebrew [17].

We observe that single linkage performs very poorly and complete linkage performs very well for the NELL data. Table 7.1 shows the improvements we got over the other clustering strategies.

Strategy	Majority Cost
Single Linkage	0.36871
Average Linkage	0.248913
Complete Linkage	0.15935
$\alpha_{SC}(0.825)$	0.15442
$\alpha_{AC}(0.825)$	0.15569

Table 7.1.: Our proposed algorithm reduces the cost by $\Delta cost = 0.493\%$.

The total improvement over the common linkage methods is 0.493% and the best clustering we generated had an error of 15.442%. With this clustering we also managed to extract new subcategories as listed in table **??**.

In addition to averaged costs, we also evaluated the clusterings for multiple values of α. This is helpful in situations where a domain expert can select from multiple suggestions. For example, if we consider the best three values of α, the domain expert can choose from three different clusterings. In order to calculate the N best values of α, we selected the 32 optimal values resulting from the experiments that led to an integer optimization problem. This resulted in ?????.

As our only formal guarantee was that there will be a maximum of $O(n^8)$ intervals in the range between single and complete linkage, we also had a look at the actual results. Since the proof for single and complete linkage in Balcan et. al [12] relies on the fact that the distance $d_{SC}(X, Y, \alpha)$ is based on four points and a split between two merges thus is based on eight points, we would expect experiments containing average linkage to have more intervals, because the average linkage distance is based on all points of the clusters. Finding formal guarantees for the average distance is not a part of this thesis and will briefly be discussed in section **??**.

In addition to looking at the resulting loss, we also evaluated how well our algorithm is able to discover new subcategories. In the given data, only very specific class labels are given such as luxurious suites, broad road or denver international airport that all belong to a certain location, i.e. these noun phrases have the specific label and the only more generalized group is the location class. In B, we listed potential subcategories that our algorithm found and annotated them with a label manually. For instance, the noun phrase luxurious suites could then be grouped together with similar ones into the cluster suite that is a subset of the cluster office building room.

7.2. Clustering Image Data

We first clustered the MNIST data. First experiments were run with 250 points in each run.

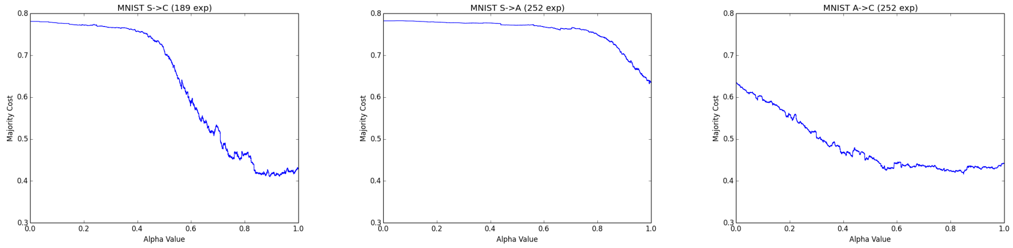

Figure 7.2.: The omniglot dataset contains handwritten characters of different alphabets, such as Latin, Greek and Hebrew [17].

Figure 7.2 shows the experimental results averaged over all 252 $\binom{10}{5}$ experiments, i.e. all different combinations of five unique labels. Again, we observe that interpolating between single and average linkage does not give us good results. Table 7.2 summarizes the results we obtain for these experiments.

Strategy	Hamming Cost
Single Linkage	0.782354
Average Linkage	0.634206
Complete Linkage	0.441931
$\alpha_{SC}(0.861624,)$	0.420714
$\alpha_{AC}(0.849407)$	0.416627

Table 7.2.: Our proposed algorithm reduces the cost by $\Delta cost = 2.5304\%$.

Reducing the cost by $\Delta cost = 2.5304\%$ seems to be a good result already. However the goal was to learn a parameter α that represents the entire dataset well. Because the procedure is very ressource-expensive, the results only looked at the first 500 points of the dataset as we clustered points from five labels in experiments of 250 points, i.e. we were using the first 50 points for each of the ten labels. This led us to running the same setting with other batches of the dataset to see if the subset of 500 points gives a good representation of the entire dataset. Figure ??? shows the results for the second batch and unfortunately the curves look quite differently, i.e. the batch did not give a good representation for the dataset.

To overcome this problem, we decided to scale up the experiments, so each of them used 1,000 instead of 250 points. As we used cloud computing to run the experiments in a reasonable time, we only evaluated the single to complete linkage and the average to complete linkage interpolation. We started with the single to complete linkage interpolation, where we show the results for the first six batches in figure 7.6. Each batch contains 2,000 points, i.e. the six batches cover the first 12,000 points of the dataset.

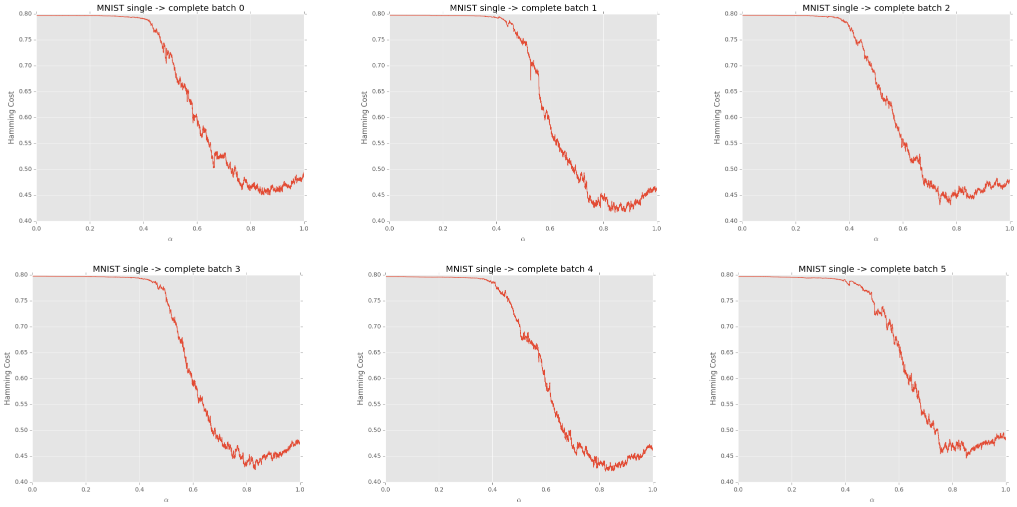

Figure 7.3.: The first six batches of the MNIST dataset result in similar curves when being evaluated between single and complete linkage.

As the curves look quite similar, we also want to analyse if the optimal values are similar. Thus we calculate the hamming cost for the optimal value of α in table 7.3.

Strategy	Batch 0	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
Single Linkage	0.796901	0.797345	0.797171	0.797405	0.796766	0.797024
Complete Linkage	0.490468	0.461063	0.479825	0.475329	0.463321	0.487111
α_{opt}	0.87228	0.84419	0.778498	0.83199	0.82338	0.852251
$cost_{opt}$	0.450012	0.416433	0.431143	0.423786	0.421103	0.446032
$\Delta cost$	4.0456%	4.463%	4.8682%	5.1543%	4.2218%	4.1079%

Table 7.3.: Our proposed algorithm reduces the cost by up to $\Delta_{max}cost = 5.1543\%$.

We were running the same experiments for the interpolation between average and complete linkage.

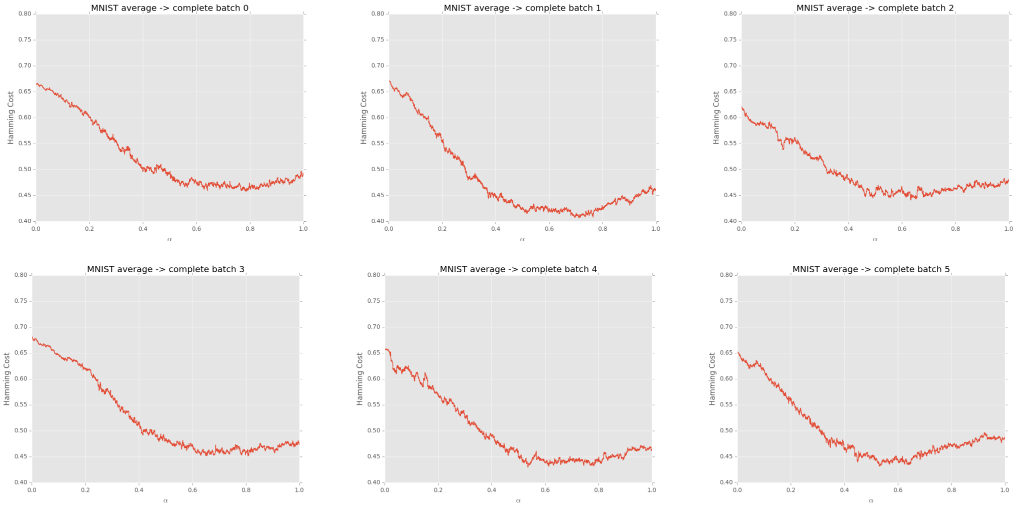

Figure 7.4.: The omniglot dataset contains handwritten characters of different alphabets, such as Latin, Greek and Hebrew [17].

The curves for interpolating between average and complete linkage in figure **??** vary more than the curves for interpolating between single and complete linkage 7.6. This results in a higher variation in the optimal values of α and the performance increase as shown in table 7.4.

In order to compare the both interpolation strategies, we averaged over the six batches to to see how well the different batches fit to each other. The results are shown in figure 7.5.

Figure 7.5 shows that both strategies give a valuable improvement over single, average and complete linkage on the first 12,000 points of the MNIST dataset. The exact improvement is shown in table 7.5.

Table 7.5 shows that the improvement for average to complete linkage interpolation is 3.3% and for single to complete linkage it is 3.7%. The optimal cost corresponds to $\alpha = 0.856557$ in the *SC* setting and $\alpha = 0.63275$ in the AC setting.

Strategy	Batch 0	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
Average Linkage	0.664952	0.672583	0.623325	0.679929	0.657857	0.652774
Complete Linkage	0.490468	0.461063	0.479825	0.475329	0.463321	0.487111
α_{opt}	0.7869	0.7124	0.634	0.807697	0.536073	0.5305
$cost_{opt}$	0.458167	0.406563	0.440964	0.451063	0.429849	0.431631
$\Delta cost$	3.2301%	5.45%	3.8861%	2.4266%	3.3472%	5.548%

Table 7.4.: Our proposed algorithm reduces the cost by up to $\Delta_{max} cost = 5.548\%$.

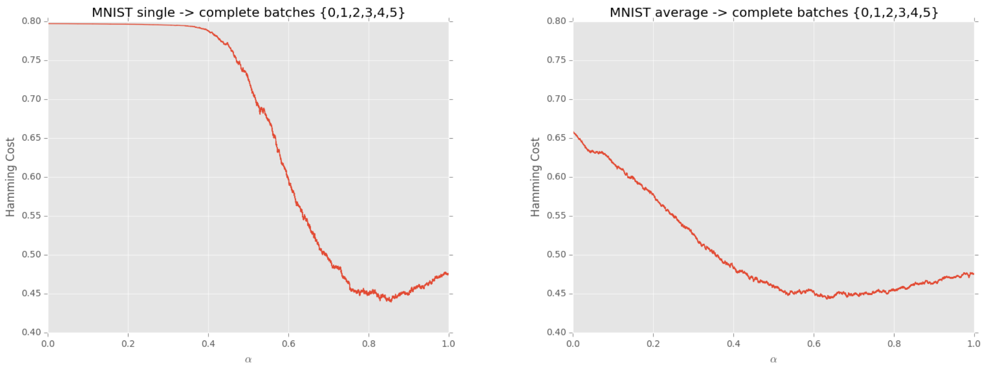

Figure 7.5.: Averaging the first six batches over both linkage strategies allows us to see how good they perform on the first 12,000 points of the MNIST dataset.

Strategy	Hamming Cost
Single Linkage	0.797102
Average Linkage	0.65857
Complete Linkage	0.476186
$cost_{opt_{SC}}$	0.439207
$\Delta cost_{SC}$	3.6979%
$cost_{opt_{AC}}$	0.443139
$\Delta cost_{AC}$	3.3047%

Table 7.5.: Over the first 12,000 points of the MNIST dataset our algorithm improvese the averaged hamming cost for 3.3%

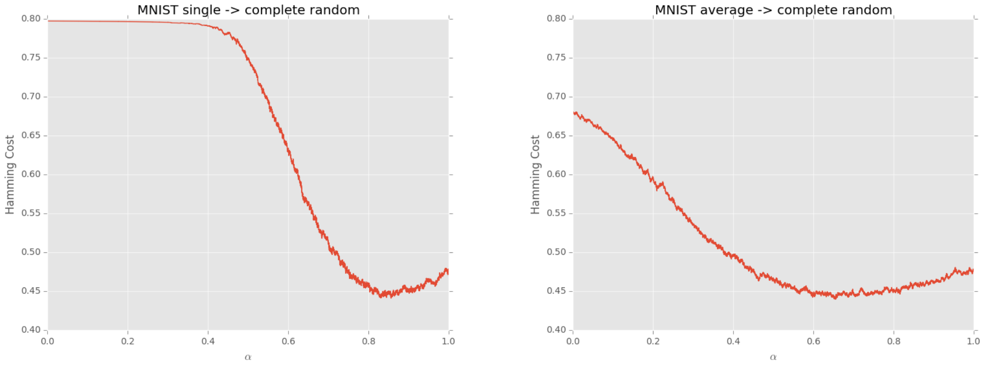

Figure 7.6.: The first six batches of the MNIST dataset result in similar curves when being evaluated between single and complete linkage.

8. Conclusion

Write section.

Bibliography

[1] Oren Zamir and Oren Etzioni. "Web document clustering: A feasibility demonstration". In: *SIGIR*. Vol. 98. Citeseer. 1998, pp. 46–54.

[2] Jaydeep Balakrishnan et al. "Product recommendation algorithms in the age of omnichannel retailing–An intuitive clustering approach". In: *Computers & Industrial Engineering* 115 (2018), pp. 459–470.

[3] Wen-Yan Lin et al. "Dimensionality's Blessing: Clustering Images by Underlying Distribution". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5784–5793.

[4] Zengyou He, Xiaofei Xu, and Shengchun Deng. "Discovering cluster-based local outliers". In: *Pattern Recognition Letters* 24.9-10 (2003), pp. 1641–1650.

[5] Samuel Fosso Wamba et al. "How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study". In: *International Journal of Production Economics* 165 (2015), pp. 234–246.

[6] Eric Sven Ristad and Peter N Yianilos. "Learning string-edit distance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.5 (1998), pp. 522–532.

[7] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: `http://www.aclweb.org/anthology/D14-1162`.

[8] Carnegie Mellon University Machine Learning Department. *CMU NELL all-pairs data, version 02-Feb-2012*. `http://rtw.ml.cmu.edu/rtw/allpairs`. Accessed: 2019-06-04.

[9] Rishi Gupta and Tim Roughgarden. "A PAC Approach to Application-Specific Algorithm Selection". In: *CoRR* abs/1511.07147 (2015). arXiv: `1511.07147`. URL: `http://arxiv.org/abs/1511.07147`.

[10] Maria-Florina Balcan, Travis Dick, and Ellen Vitercik. "Dispersion for Data-Driven Algorithm Design, Online Learning, and Private Optimization". In: *CoRR* abs/1711.03091 (2017). arXiv: `1711.03091`. URL: `http://arxiv.org/abs/1711.03091`.

[11] Eleni Triantafillou et al. "Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples". In: *CoRR* abs/1903.03096 (2019). arXiv: `1903.03096`. URL: `http://arxiv.org/abs/1903.03096`.

[12] Maria-Florina Balcan et al. "Learning-Theoretic Foundations of Algorithm Configuration for Combinatorial Partitioning Problems". In: *CoRR* abs/1611.04535 (2016). arXiv: `1611.04535`. URL: `http://arxiv.org/abs/1611.04535`.

[13] T. Mitchell et al. “Never-ending Learning”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.* AAAI’15. Austin, Texas: AAAI Press, 2015, pp. 2302–2310. ISBN: 0-262-51129-0. URL: `http://dl.acm.org/citation.cfm?id=2886521.2886641`.

[14] T. Mitchell et al. “Never-ending Learning”. In: *Commun. ACM* 61.5 (Apr. 2018), pp. 103–115. ISSN: 0001-0782. DOI: `10.1145/3191513`. URL: `http://doi.acm.org/10.1145/3191513`.

[15] Yann LeCun and Corinna Cortes. “MNIST handwritten digit database”. In: (2010). URL: `http://yann.lecun.com/exdb/mnist/`.

[16] Alex Krizhevsky. “Learning Multiple Layers of Features from Tiny Images”. In: 2009.

[17] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. “Human-level concept learning through probabilistic program induction”. In: *Science* 350.6266 (2015), pp. 1332–1338. ISSN: 0036-8075. DOI: `10.1126/science.aab3050`. eprint: `http://science.sciencemag.org/content/350/6266/1332.full.pdf`. URL: `http://science.sciencemag.org/content/350/6266/1332`.

[18] Harold W Kuhn. “The Hungarian method for the assignment problem”. In: *Naval research logistics quarterly* 2.1-2 (1955), pp. 83–97.

[19] James Munkres. “Algorithms for the assignment and transportation problems”. In: *Journal of the society for industrial and applied mathematics* 5.1 (1957), pp. 32–38.

[20] Dan DeBlasio and John Kececioglu. “Parameter advising for multiple sequence alignment”. In: *BMC bioinformatics.* Vol. 16. 2. BioMed Central. 2015, A3.

A. The Hungarian Method

Our goal is to find the best possible matching between two clusterings $C_1^i, ..., C_k^i$ and $C_1^j, ..., C_k^j$. In order to do so, we calculate the cost of matching each possible pair of clusters within the two clusterings.

To find the optimal matching in a brute force way, we have to look at each possible matching. Say we want to match each i to one j. For $i = 1$ we can pick from 5 different values of j, for $i = 2$ there are 4 potential values of j. This will overall result in $k! = 5! = 120$ different combinations, thus the complexity of the brute force approach is $O(k!)$. A more efficient algorithm (especially for higher values of k) was introduced by Kuhn and Munkres [18][19]. It consists of three major steps. In the first one, we subtract the row minima from each row. This step is performed in table A.1.

j\i	1	2	3	4	5	
1	5	0	15	35	25	(-15)
2	70	0	5	10	20	(-10)
3	0	10	30	60	40	(-20)
4	20	40	30	10	0	(-10)
5	0	10	20	30	5	(-20)

Table A.1.: Hungarian method step 1: Subtract the row minima from each row.

After subtracting the row minima, we now also subtract the column minima from each column as shown in table A.2.

j\i	1	2	3	4	5
1	5	0	10	25	25
2	70	0	0	0	20
3	0	10	25	50	40
4	20	40	25	0	0
5	0	10	15	20	5
	-	-	(-5)	(-10)	-

Table A.2.: Hungarian method step 2: Subtract the column minima from each column.

Now we try to find the optimal matching. To do so, we cover all zeros with lines and count the minumum needed lines to do so. Table A.3 shows that we need four lines.

After covering the zeros and counting the lines, we found the optimal matching in case the number of lines equals the number of rows (or columns) in the matrix. As we need four lines and the matrix has five rows in this example, we have to add more zeros. To do

j\i	1	2	3	4	5
1	5	0	10	25	25
2	70	0	0	0	20
3	0	10	25	50	40
4	20	40	25	0	0
5	0	10	15	20	5

Table A.3.: Hungarian method step 3: Cover all zeros with as few lines as possible.

that, we subtract the minimum value of the matrix (which is 5 here) from all uncovered values that are not zero and add it to all values that are not zero and covered twice. Now we can again check the needed lines as in table A.4.

j\i	1	2	3	4	5
1	5	0	5	20	20
2	75	0	0	0	20
3	0	10	20	45	35
4	25	45	25	0	0
5	0	10	10	15	0

j\i	1	2	3	4	5
1	5	0	5	20	20
2	75	0	0	0	20
3	0	10	20	45	35
4	25	45	25	0	0
5	0	10	10	15	0

Table A.4.: Hungarian method additional step: Create more zeroes until the number of minimal needed lines to cover all zeros matches the number of rows.

This will then result in the assignment seen in table A.5. Applying the matching to the input matrix then gives the optimal cost by summing the optimal values. For this example the optimal cost is then 95.

j\i	1	2	3	4	5
1	5	0	5	20	20
2	75	0	0	0	20
3	0	10	20	45	35
4	25	45	25	0	0
5	0	10	10	15	0

j\i	1	2	3	4	5
1	20	15	30	50	40
2	80	10	15	20	30
3	20	30	50	80	60
4	30	50	40	20	10
5	20	30	40	50	25

Table A.5.: Result of the hungarian method: The optimal matching between two clusterings.

B. Proposed NELL Subcategories

Luxury Room	Bathroom	Guest Room	Suite
spacious living room	large ensuite bathroom	elegant rooms	luxurious suites
comfortable living room	spacious marble bathroom	three guest rooms	one bedroom suites
guest room	one bathroom	large guest rooms	spacious suites
lounge room	full bathroom	deluxe guest rooms	deluxe suites
living room	upstairs bathroom	guests rooms	guest suites
superior room	large bathroom	spacious air conditioned rooms	bedroom suites
sleeping room	ensuite bathroom	furnished guest rooms	whirlpool suites
main bedroom	elegant bathroom	comfortable guest rooms	three suites

Table B.1.: Proposed Subcategories for “Office Building Room”.

Shoes	Uniform/Costume	Pants	Casual	Specialized
shoes	costume	kneepants	stocking cap	long stockings
high heel shoes	work uniforms	baggy pants	workout clothes	wide brimmed hat
sensible shoes	outfits	loose fitting pants	casual clothes	casual wear
old shoes	period costume	slacks	baseball caps	black stockings
pointe shoes	folk costumes	black shorts	skull caps	wear socks
dark shoes	halter top	special clothing	ball caps	high heels
spira shoes	period costumes	white shorts	evening clothes	surf wear
mens shoes	costumes	underpants	ball cap	wear gloves

Table B.2.: Proposed Subcategories for “Clothing”.

Stove/Oven	Machines	Bowls	Baking Sheets
full size stove	cookie cutters	large mixing bowl	oiled baking sheet
full size cooker	automatic washing machine	large serving bowl	rimmed baking sheet
red hot stove	washing machine	small bowl	large baking sheet
plastic jug	bread machine	single bowl	small baking sheet
toaster	cookie cutter	separate bowl	prepared baking sheet
greased baking dish	coffee machine	shallow bowl	ungreased baking sheet
wood burning pizza oven	cooking spray	separate mixing bowl	hot plate
ceramic top stove	coffee grinder	large bowl	greased baking sheet

Table B.3.: Proposed Subcategories for “Kitchen Item”.

C. Convolutional Neural Network Architecture for Feature Extraction

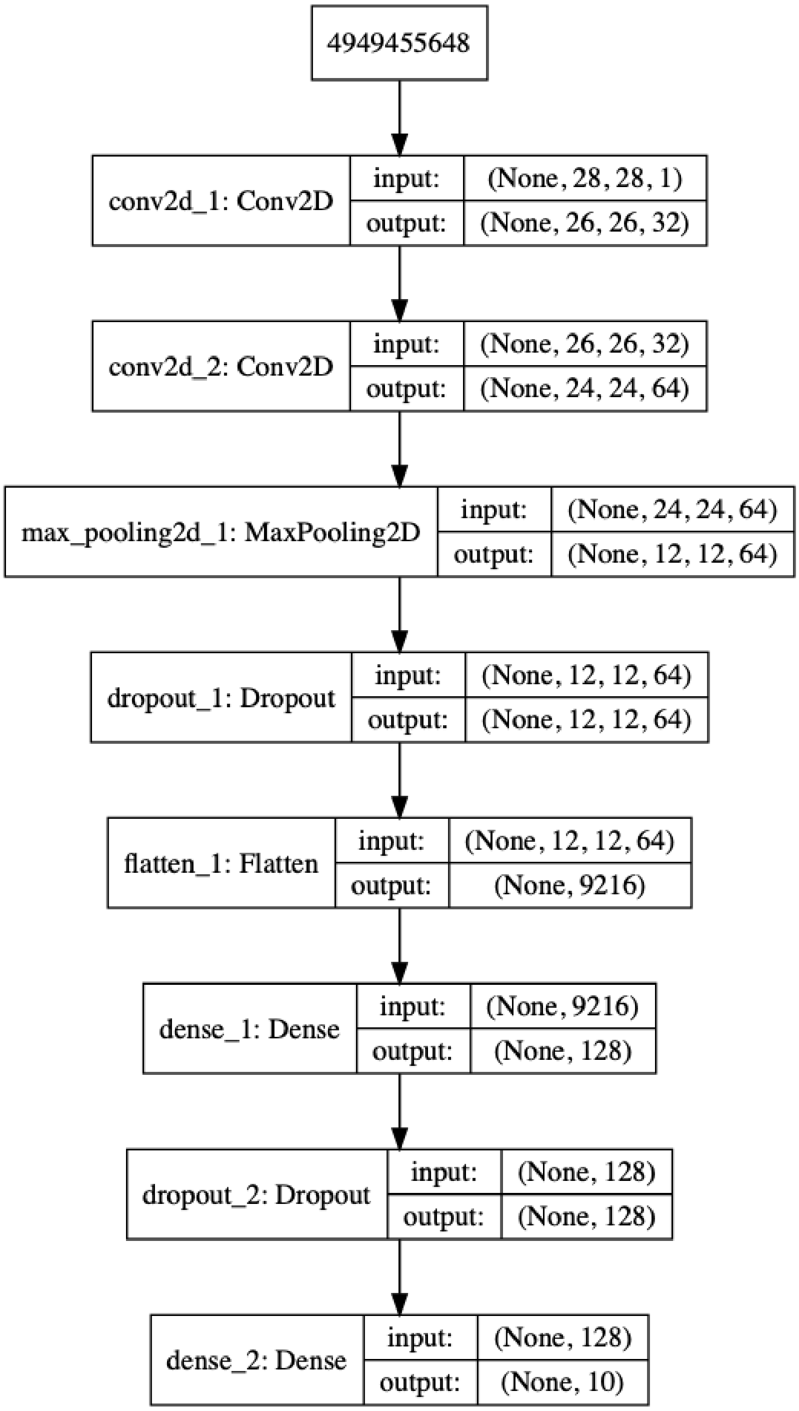

Figure C.1.: Combining several metrics seems often natural and can lead to improved results as in this example where we project a dataset on both axes.