The Advanced-Maximum-Linkage Clustering-Algorithm

Lars Tschiersch; Matthias Zerbst

Department of Statistics, University of Dortmund, Germany

Abstract: The advanced maximum-linkage-algorithm (AMLA) is a derivative of the maximum-linkage-algorithm (MLA) given by Zerbst (2001). AMLA produces clusterings which have a good separation between the built classes. To reach the separation, centroids were calculated, which are the basis for classification. The separation between the classes is considered large, if this is also true for the so far calculated centroids. The selection of centroids having a large distance to each other is also guaranteed using the MLA. AMLA is improved in this way that the underlying frequency structure of the data can be adapted individually to the problem under consideration. Therefore an additional parameter $\rho \in [0, 1]$ is introduced. The parameter gives the degree to which the frequency structure is regarded.

Introduction

This paper introduces a clustering approach. The approach is called *Advanced-Maximum-Linkage-Algorithm* (AMLA). When deriving a clustering elements will be compendious within homogeneous classes. In general the number of classes will be given. Only in rare cases this number is derived by the clustering algorithm itself.

Classical clustering algorithms are based upon the k-means-approach (e. g. Anderberg, 1973), as well as Linkage-methods (e. g. Johnson & Wichern, 1992). Even neural networks, for instance self-organising maps (Kohonen, 2001), are used. All this approaches have one idea in common. All of them pool similar elements into classes.

A different basic approach, which guides to a clustering as well, is based on decollating dissimilar elements. This approach has advantages when beside homogeneity within classes a strictly separation of the classes is postulated. Such an approach is the maximum-linkage-algorithm (MLA) given by Zerbst et al (2000) and Zerbst (2001). MLA produces perspicuous decollated classes. Moreover classes for rare elements were inspected, provided that such elements built an own class due to the exposed site within the feature space.

As Zerbst (2001) illustrates the distinctiveness of MLA consists in considering the frequency distribution of the elements to cluster when using a fast clustering algorithm. MLA is especially designed for clustering pixels if aerial photographs or satellite images using RGB-colour format. The consideration of the frequency distribution for the elements to cluster when using MLA depends on a bounded feature space with a known maximum for the variance. When using other than RGB-data this does not hold always.

This diversification leads to the advanced-maximum-linkage-algorithm (AMLA). AMLA considers the frequency distribution for the elements to cluster essentially better than MLA. We will come back to this aspect in chapter 3.

In chapter 1 we will have a short introduction into theory of clustering. Chapter 2 contains the mathematical notation of the advanced-maximum-linkage-algorithm and the presentation of the advantages of AMLA in comparison to MLA. Chapter 3 gives some examples which show the change in clustering when different weight parameters ρ are used. At the end we will give an outlook on the further research.

1 Clustering in general

The aim of a clustering is to divide a set of elements in single classes. Thereby elements of the same class should be as similar as possible. Elements of different classes should be as dissimilar as possible.

The formal description of a clustering is as follows:

- Let Ω be the feature space of elements to cluster, whereby $\Omega \subseteq I\!R^{\,m}$.
- Let O be the set of elements to cluster $x_i \in \Omega$, $i = 1,...,n$, with $O = \{x_1,...,x_n\}$. The cardinality of the set O is given by n.
- Let q be the a priori given number of classes.

The clustering $C(O)$ with elements of O is defined as

$$C(O) = \left\{ c_1,...,c_q \right\} \quad \text{with} \quad c_i = \left\{ c_{i1},...,c_{in_i} \right\} \;, i = 1,...,q \,.$$

c_i, $i = 1,...,q$ represents a single class of the clustering.

The question of how to derive a clustering depends on the users' requirements and the initial situation. One has to bear in mind, that there exists no optimal clustering. Optimality of clustering is regarded with respect to various criterions. There exist different algorithms which constitute clusterings in dependency on different criterions.

The usually used criterion is the intra class variance clustering criterion (ICVC = Intra-Class-Variance-Criterion). According to this criterion, the optimal clustering is given, where the induced intra class variance is minimal. Bock (1998) shows, that the optimal clustering C out of all possible clusterings Γ is given by:

$$C(O) := \arg\min_{C(O) \in \Gamma} \frac{1}{n} \sum_{i=1}^{q} \sum_{k \in c_i} \left\| x_k - \frac{1}{n_i} \sum_{j=1}^{n_i} x_j \right\|_2 .$$

This minimum can only be achieved by the comparison of all possible clusterings, which takes long computational time, even with fast computers. For this reason algorithms are used which give solutions in proximity to this criterion. Such an algorithm is the k-means-approach (e.g. Hartigan, 1975). Several innovations are based upon this algorithm, such like the ISODATA-Program (Tou und Gonzales, 1974) or the PHASE-approach (Myers et al, 1997). The mentioned approaches tried to simplify the enormous computational effort.

To derive a clustering with the k-means approach, we have to detect a set Z of q centroids such that

$$Z = \{z_1, \ldots, z_q\}, \text{ with } z_i \in \Omega, \ i = 1, \ldots, q$$

holds.

The values $x_i \in \Omega, i = 1, \ldots, n$, will classified according to the centroids $z_i \in \Omega$, $i = 1, \ldots, q$. After the classification we have a clustering with q classes. To get new centroids we have to calculate the mean within each class. This means will be handled as new centroids in a new iteration step of the algorithm. Due to the fact that these values represent the classes, we are going to call them class representatives.

A totally different cluster criterion than the ICVC is the maximum linkage-criterion (MLC). The criterion requires from the class representative that the minimal distance between these class representatives should be maximal.

First choose q centroids from the set of values to cluster, which fulfil the above condition. A classification with respect to calculated centroids leads us to an optimal clustering, with regard to this criterion. The centroids are calculated with respect to the criterion such that

$$x_{1*},...,x_{q*} \in Z \quad \Leftrightarrow \quad \{1^*,...,q^*\} = \operatorname*{arg\,max}_{\{1^*,...,q^*\} \subset \{1,...,r\}} \operatorname*{arg\,min}_{i,j \in \{1^*,...,q^*\},\ i \neq j} \left\| x_i - x_j \right\|_2 \ ,$$

with r equal to the number of different elements of $x_i \in \Omega$, $i = 1,...,n$, holds (Zerbst, 2001). This condition is just of formal interest. When using the MLA the values out of the set of values to cluster can be distinguished. The elements calculated with MLA fulfil approximately the condition above. Further consideration according to the MLC and the centroids calculated with MLA can be read by Zerbst (2001). Beside the difference between MLC and MLA considers MLA the frequency of the underlying elements.

2 Advanced Maximum Linkage

The idea of advanced maximum linkage is to derive a clustering which is able to separate strictly between the classes. Therefore the algorithm first calculates centroids from the set of the values to cluster. The centroids are used to classify the remaining values. The selection of the centroids ensures two assumptions. On the one hand the distance between the centroids should be as large as possible. On the other hand the frequency structure should also be taken into account. In highly densed areas we would like to pick more centroids then in sparsed densed areas. But the sparse densed area should not be totally neglected. Especially the elements of those areas should be collected in separate classes because of their exposed site (large distances to other elements) in the feature space.

2.1 The Advanced Maximum Linkage Algorithm (AMLA)

The following definitions are necessary for the description of AMLA:

- Let $\Omega \subseteq \mathbb{R}^m$ be the feature space, which contains the elements to cluster.
- Let $O = \{x_1,...,x_n\}$, with $x_i \in \Omega$, $i = 1,...,n$, be the set of the values to cluster. This means we will cluster n elements.
- The set O can be split up in r different values $x_1,...,x_r$. Let $h_1,...,h_r$ the corresponding absolute frequencies.
- Let q be the a priori given number of classes.

- Let $\rho \in [0,1]$ a user specified parameter. This parameter gives the percentage of the frequency ($\rho \cdot 100\%$) of the data which will be considered by the clustering. For $\rho = 0$ we will not consider the frequency at all and for $\rho = 1$ the frequency is considered maximal.

- AMLA is an iterative algorithm. Therefore describes s the iteration step.

With respect to the above definitions we can make further definitions.

- Let the joint, relative empirical frequency λ_{ij}, $i, j \in \{1,\ldots,r\}$ be defined as

$$\lambda_{ij} = \frac{h_i \cdot h_j}{\left(\sum_{k=1}^{r} h_k\right)^2} = \frac{h_i \cdot h_j}{n^2}, \quad i, j \in \{1,\ldots,r\}, \text{ such that } \lambda_{ij} \in \left(\frac{\left(\min_{k=1,\ldots,r} h_k\right)^2}{n^2} ; 1 \right] \subset (0,1]$$

Let λ_i, $i \in \{1,\ldots,r\}$ be the simple empirical frequency:

$$\lambda_i = \frac{h_i}{\sum_{k=1}^{r} h_k} = \frac{h_i}{n}, \quad i \in \{1,\ldots,r\}, \text{ so that } \lambda_i \in \left[\frac{\min_{k=1,\ldots,r} h_k}{n} ; 1 \right] \subset (0,1]$$

- The consideration of the frequency distribution is realized with a weight function f. This function has two parameters. The first parameter is $\rho \in [0,1]$, which gives the part of the frequency taken into account for calculation. The second parameter is λ, which is the joint or simple relative empirical frequency, respectively, according to the iteration setup. The function is given by:

$$f(\rho, \lambda) = \rho \cdot \lambda + (1 - \rho), \quad \text{with } \rho \in [0,1] \text{ and } \lambda = \begin{cases} \lambda_{ij}, & s = 1 \\ \lambda_i, & s > 1 \end{cases}.$$

To illustrate how the function works, consider to following example.

Example

ρ	$f(\rho,\lambda)$	interval for f
0	1	$[1\,,\,1]$
½	½ λ + ½	$[\text{½}\,,\,1]$
¾	¾ λ + ¼	$[\text{¼}\,,\,1]$
1	λ	$\left[\min\lambda\,,1\right]$

Tab. 1: Counter-domain of $f(\lambda,\,\rho)$ for different ρ's

The function f can be used for weighting with relative frequency in dependence on the parameter ρ. As ρ gets big the weight of different λ_{ij} , $i,\,j \in \{1,\ldots,r\}$ increases. As smaller ρ gets the weight difference will decrease.

The algorithm works as follows:

Initialization

The user has to specify the parameters q and ρ.

Step 1

Calculate the matrix Δ, which contains the weighted distances of all r different values $x_1,\ldots,x_r$. For weighting use function f. That is

$$\Delta = \left(\delta_{ij}\right)_{ij=1,\ldots,r}, \text{ with } \delta_{ij} = f\left(\rho,\lambda_{ij}\right)\cdot\left\| x_i - x_j \right\|_2 . \tag{2.1}$$

Choose the maximum of the elements from Δ. The corresponding two values will be chosen as the first two centroids. The set of centroids will be marked with $Z^{(s)}$, where the superscript s stands for the iteration step. The set $Z^{(1)}$ is given by:

$$Z^{(1)} = \left\{ x_{i*}, x_{j*} \right\} \;\Leftrightarrow\; \left\{ i*, j* \right\} = \underset{i,j\in\{1,\ldots,r\}}{\arg\max}\,\delta_{ij} . \tag{2.2}$$

Step 2

Let $R^{(s)}$ and $Q^{(s)}$ be index sets which depend on the iteration step s. The indices of all marked values x_{i*} , $i* \in \{1^*,\ldots,(s+1)^*\}$ within the iteration step s be included in $Q^{(s)}$: $i \in Q^{(s)} \;\Leftrightarrow\; x_i \in Z^{(s)}$. The indices of the values x_j, $j \in \{1,\ldots,r\}\backslash \{1^*,\ldots,(s+1)^*\}$, which are not marked as centroids are contained in $R^{(s)}$: $R^{(s)} = \{1,\ldots,r\}\backslash Q^{(s)}$.

6

Now, calculate the distances between all elements in $R^{(s)}$ to all elements in $Q^{(s)}$. These distances are also weighted by f. The corresponding minimum will be entered in the distance vector $\delta_{iQ}, i \in R^{(s)}$:

$$\delta_{iQ} = f(\rho, \lambda_i) \cdot \min_{k \in Q} \| x_i - x_k \|_2, \quad i \in R^{(s)}. \tag{2.3}$$

The maximum of all entries from $\delta_{iQ}, i \in R^{(s)}$ is chosen and the corresponding value will be added to the sets of centroids:

$$Z^{(s+1)} = Z^{(s)} \cup \{x_{i*}\} \quad \Leftrightarrow \quad i* = \arg\max_{i \in R^{(s)}} \delta_{iQ}. \tag{2.4}$$

Step 2 will be repeated until $|Z| = q$. Therefore the step will be run through $(q - 2)$ times.

Step 3

Classify the different values $x_1, ..., x_r$ according to the calculated centroids $x_{i*}, i* = 1, ..., q$. After that the mean within each class is derived. When deriving the mean the empirical frequencies of the values are taken into account, by weighting which these frequencies. The mean values within each class are called class representatives. These values will be needed for further analysis and visualization.

Step 4

To judge the quality of a clustering we need some criterions. On the one hand there is the intra- and inter class variance. Is the intra class variance small, the class can be considered homogeneous. On the other hand is the average minimal distance between the class representatives important. The separation between the classes is better when the average minimal distance is large. Further information on judging the quality is given by Zerbst (2001).

<u>Type of weighting</u>

Why do we consider the joint relative frequency in the weight function (2.1) and the simple relative frequency in (2.3)? In principle, the frequency of centroid candidates is considered. Due to the fact that we judge two candidates in (2.1) we use the joint relative frequency. With (2.3) we judge just one candidate. Therefore we consider only the simple relative frequency. Furthermore, there will be an additional reason not to use the joint relative frequency in (2.3). See the following example. Let two centroids be already calculated. Now we are interested in calculating a third one. Two potential candidates are possible.
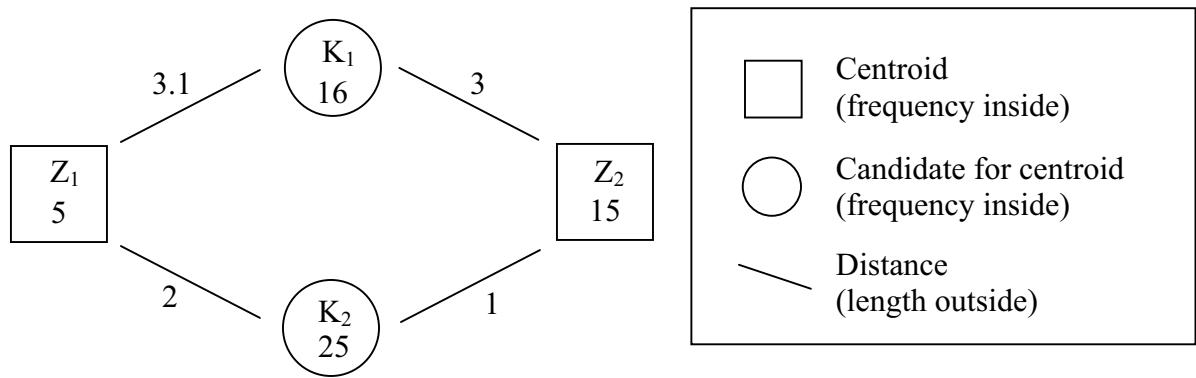
The above graphic shows that the candidate K_2 is no alternative to K_1. This is due to the enormous gap between the minimum distances which are 1 for K_2 and 3 for K_1.

Suppose the total frequency is 100. If we consider the joint frequency we get a distance matrix such as in Tab. 2. If we consider the simple frequency we get a distance matrix such as in Tab. 3.

	joint frequency consideration	
	Z_1	Z_2
K_1	$\frac{16}{100} \cdot \frac{5}{100} \cdot 3.1 = 0.0248$	$\frac{16}{100} \cdot \frac{15}{100} \cdot 3 = 0.072$
K_2	$\frac{5}{100} \cdot \frac{25}{100} \cdot 2 = 0.025$	$\frac{25}{100} \cdot \frac{15}{100} \cdot 1 = 0.0375$

Tab. 2: Matrix to calculate δ_{iQ} with respect to the joint frequency.

The grey coloured cells shows the results of the Minimum build over rows. After building the maximum we choose the candidate K_2. But candidate K_2 isn't good enough, because of the

above mentioned reasons. Even when we consider the frequency of the values, the maximal minimal distance is essential for clustering. This is not reached when using the joint frequency. Then K_2 is chosen, which has a minimal distance of 1. If we choose K_1 instead we have a minimal distance of 3. Tab. 3 shows that this doesn't happen when considering only the simple frequency, due to not using the frequency at all in the step where the minimum is derived.

	simple frequency	
	Z_1	Z_2
K_1	$\frac{16}{100} \cdot 3.1 = 0.496$	$\frac{25}{100} \cdot 2 = 0.5$
K_2	$\frac{16}{100} \cdot 3 = 0.48$	$\frac{25}{100} \cdot 1 = 0.25$

Tab. 3:   Matrix to calculate δ_{iQ} with simple frequency.

When building the maximum now, the "right" candidate K_1 will be chosen. We calculate the minimum independently of the frequency. After this calculation the frequency we'll be taken into account as well. The idea behind this strategy is that the absolute maximum wouldn't be reached by this procedure, but this is acceptable if a greater set of values (larger frequency) is represented. The frequency of the centroids calculated so far will not be considered.

2.2 Differences between AMLA und MLA

The substantial difference between AMLA and MLA is based in whether considering the frequency or not. When using AMLA, the frequency is taken into account in dependence of the parameter ρ when we going to calculate the weighted differences via (2.1) and (2.3). The parameter allows to fit the weights individually to the underlying problem. Where as when using MLA the desired set of centroids consists of two different sub sets. One sub set is derived without considering any frequency. The other sub set is derived with the distances from (2.1) or (2.3) weighted with the absolute frequencies of the data values. The cardinality of both sub sets depends on the total variance of the dataset (Zerbst, 2001). The number of weighted centroids, in the case of 24-bit-RGB-colorvalues, is given by

$$g_W(\kappa, q) = \left\lfloor \left(\sin\left(\frac{\pi\kappa}{48768.75} - \frac{\pi}{2} \right) + 1 \right) \frac{q}{2} \right\rfloor \text{ with } \kappa = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{3} \left(x_{ik} - \frac{1}{n} \sum_{j=1}^{n} x_{jk} \right)^2 .$$

9

The amount of unweighted centroids is given by $g_{UW}(\kappa,q) = q - g_W(\kappa,q)$.

The selection of centroids with respect to the weighting is given by

$$\Delta_h = (h_{ij} \cdot \delta_{ij}), \ \ \delta_{ij} = \left\| x_i - x_j \right\|_2, \ \ i, j = 1,...,r,$$

with $h_{ij} = h_i + h_j$, $i, j = 1,...,r$, instead of (2.1). Analogous we get for (2.3)

$$\delta_{iQ_h} = h_i \cdot \min_{k \in Q} \left\| P_i - P_k \right\|_2, i \in R.$$

When we choose unweighted centroids (2.1) changes to

$$\Delta = (\delta_{ij}), \ \ \delta_{ij} = \left\| x_i - x_j \right\|_2, \ \ i, j = 1,...,r,$$

and (2.3) to

$$\delta_{iQ} = \min_{k \in Q} \left\| x_i - x_k \right\|_2, \ \ i \in R.$$

The advantages of AMLA against MLA can be summarized as follows:

1. When using AMLA, the centroids are calculated with one formula.
2. The derived cardinalities $g_W(\kappa,q)$ and $g_{UW}(\kappa,q)$ by MLA correspond to the parameter ρ of AMLA. Hence, the adjustability of the weights is only given when using AMLA.
3. Both centroid calculating methods of MLA can be found in AMLA as well. A choice of centroids without weighting corresponds to the case where $\rho = 0$. A choice of centroids with weighting corresponds to the case where $\rho = 1$. Hence, in MLA two sets of centroids are calculated with extreme weighting. But better results can be achieved by using AMLA for calculating centroids with no extreme parameter for ρ (see chapter 3).

3 Example on the capacity of AMLA

To demonstrate how AMLA works and showing the competitiveness we are going to cluster the pixel of an image. Exemplary we choose an image from a farming study in West Africa. The study was supported by the Collaborative Research Centre „Adapt farming in West Africa" (SFB 308) of the German research foundation (DFG). The image was made with a captive balloon (dragon) close to the township of Chikal (3°26'E, 14°14'N) in the western part of the Sahelstate Niger (Gérard et al, 1997). The following image contains 393216 Pixel and 8730 different colours.
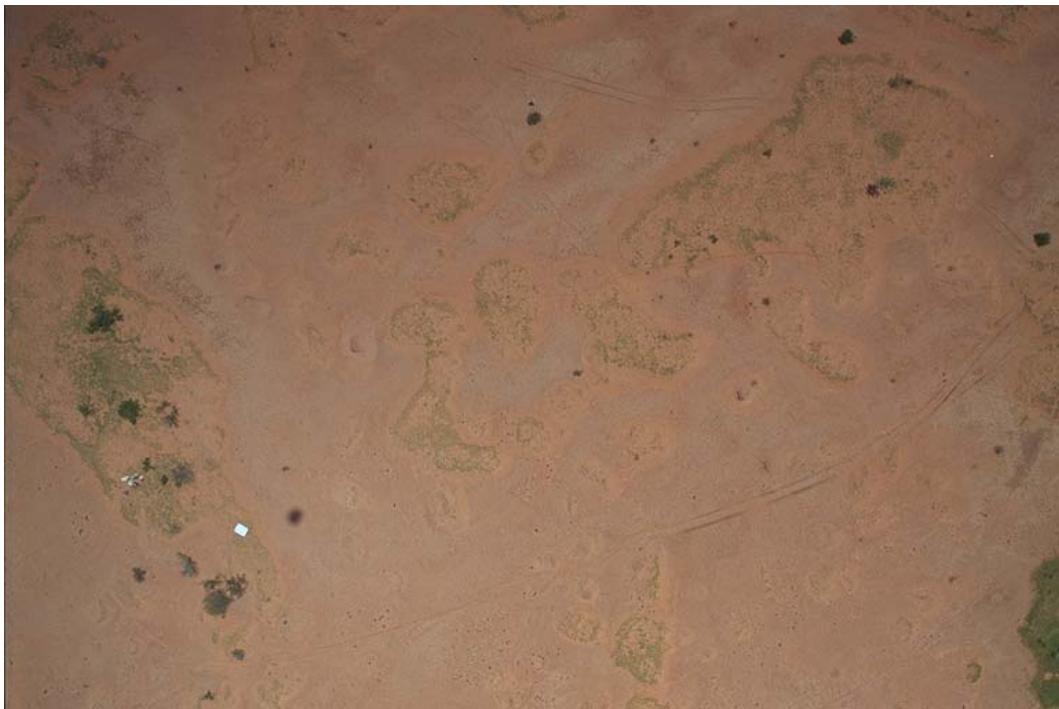
Fig. 1: Aerial photograph close to the township of Chikal (3°26'E, 14°14'N), western Niger.

We are going to cluster the different colours values of such images. In each class the mean is calculated. These mean values will be marked as class representatives, which will be used further for represent such an image. Thus, images similar to thematic maps arise. Those images give valuable information to ecological experts.

In this example we are going to reduce the number of classes with AMLA from 7830 to 25.
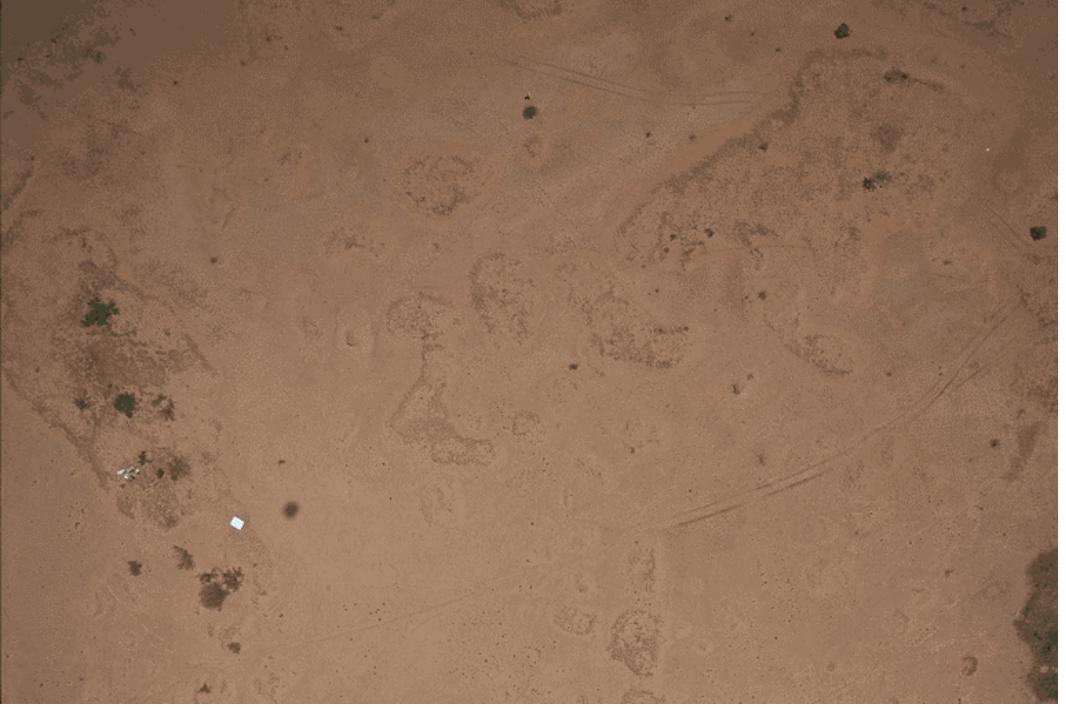
Fig. 2: Image based on class representatives derived from Fig. 1 with 25 classes.

To judge the quality of a clustering we need some criteria and the visualization of the class representatives in the feature space. The main criterion to judge the quality of clustering is the intra class variance (Bock, 1998). With this criterion we can assess how homogenous the classes are. If the intra class variance is small, we consider the clustering as good with respect to this criterion. But we also need to take into account the selectivity between the classes. This criterion is based on the distances between all centroids to the next nearest centroid. This leads us to the average minimal distance criterion (Zerbst, 2001). To see how the new parameter ρ works we made some clusterings. The number of classes is set to 25. The parameter ρ will be vary.

ρ	SSW	AvgMinDist
0 – 0.9	68.4	548.91
0.95	73.6	643.96
0.99	73.6	643.96
0.999	65.5	592.10
0.9999	37.1	511.63
0.999925	35.0	495.52
0.99995	31.1	509.40
0.999975	26.6	488.23
0.99998	24.6	504.70
0.999985	23.8	538.05
0.99999	24.8	548.18
0.999999	24.0	1313.54
0.9999999	24.0	1313.54
≤ 0.99999999	23.0	1223.00

Tab. 4: Parameters of the clustering with AMLA by 25 classes and several different ρ.

As seen from Tab. 4 there are no essential differences, when the parameter ρ varies between 0 and 0.9. That means that the frequency wouldn't be used at all in this image clustering. Even for ρ between 0.95 and 0.99 the results are substantial the same. At least in the area between 0.9999 and 1 there will be observable different results.

The intra class variance varies between 23.0 and 73.6. In comparison with the total variance of 797.6 comes this criterion to the conclusion to judge all clusterings as "good". Even if the difference is not large we can see the following: As ρ becomes large, the intra class variance becomes small. The average minimal distance (AvMiDi) varies between 495.52 and 1313.54. This criterion tends to larger values if the frequency is considered, but the values seem not strongly be linked to ρ as the intra class variance. For different ρ we get greater differences by the AvMiDi than by the intra class variance. Further interpretation can be read by Zerbst (2001). We can't make any conclusions from the determined table values. If necessary we can detect a trend, how different values of ρ influence the clustering. For this reason we'll have a look on the visualization of the class representatives in the feature space.
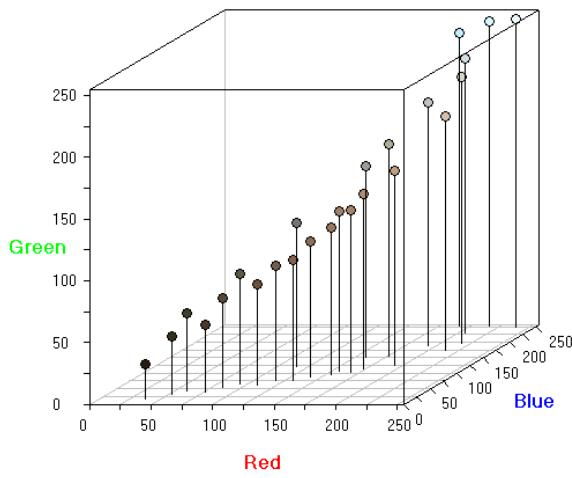
Fig. 3.a: Centroids of the Clustering of Fig. 1 with $\rho \in (0.0 \, ; 0.9)$
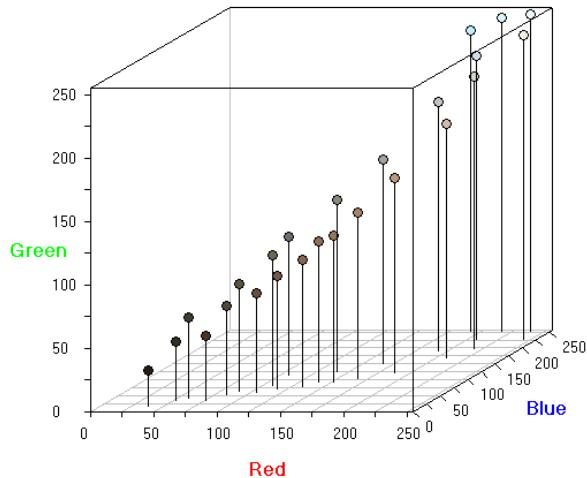
Fig. 3.b: Centroids of the Clustering of Fig. 1 with $\rho \in (0.95 \, ; 0.99)$
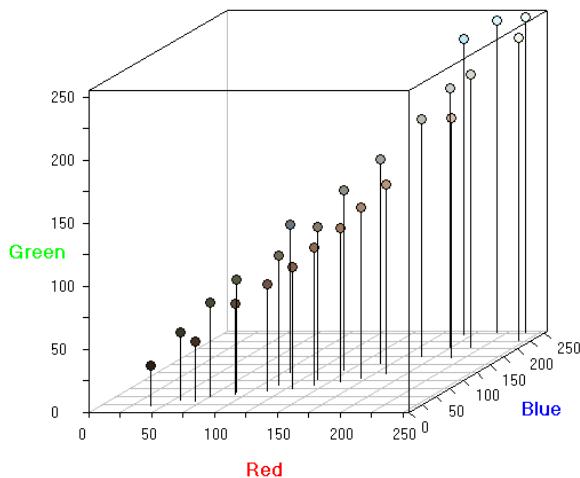
Fig. 3.a: Centroids of the Clustering of Fig. 1 with $\rho = 0.999$
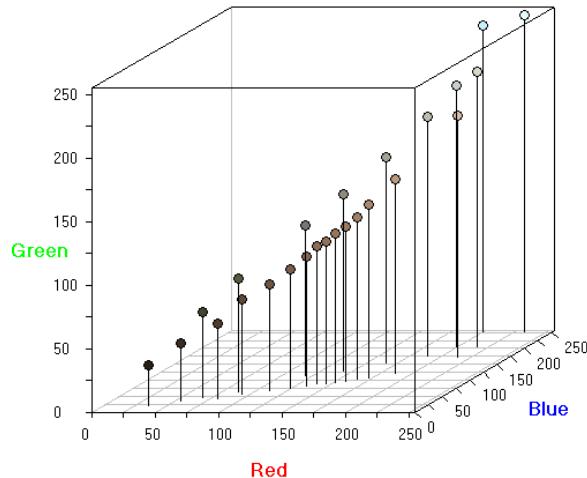
Fig. 3.a: Centroids of the Clustering of Fig. 1 with $\rho = 0.9999$
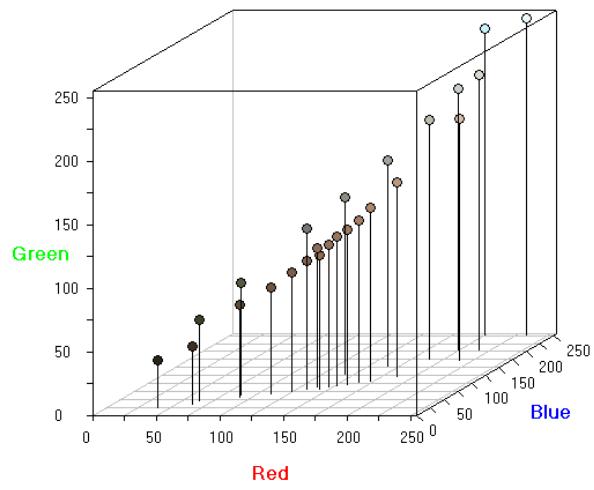
**Fig. 3.c:** Centroids of the Clustering of Fig. 1
with ρ = 0.999925
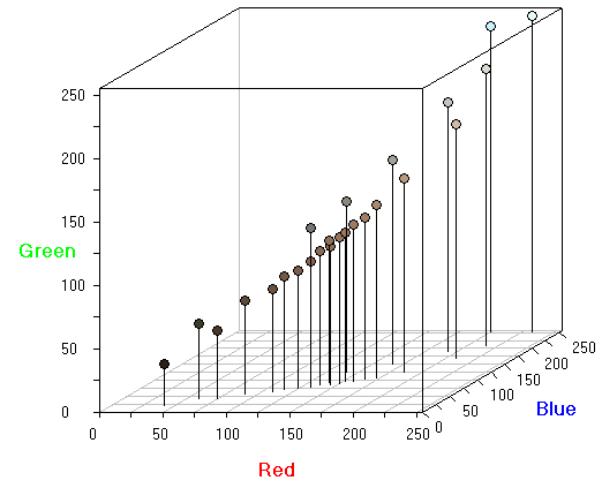
**Fig. 3.d:** Centroids of the Clustering of Fig. 1
with ρ = 0.99995
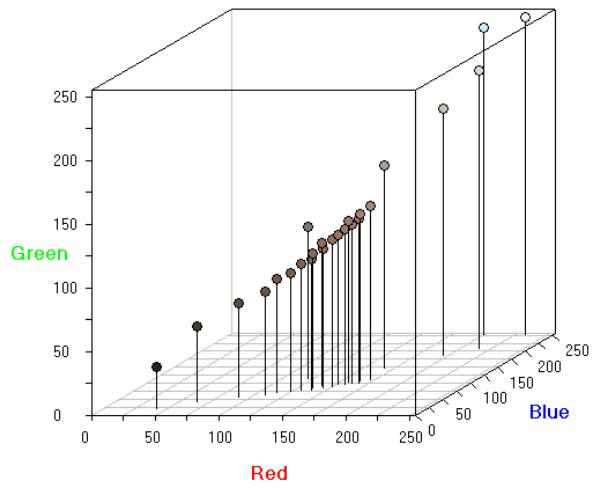
Fig. 3.e: Centroids of the Clustering of Fig. 1
with ρ = 0.999975
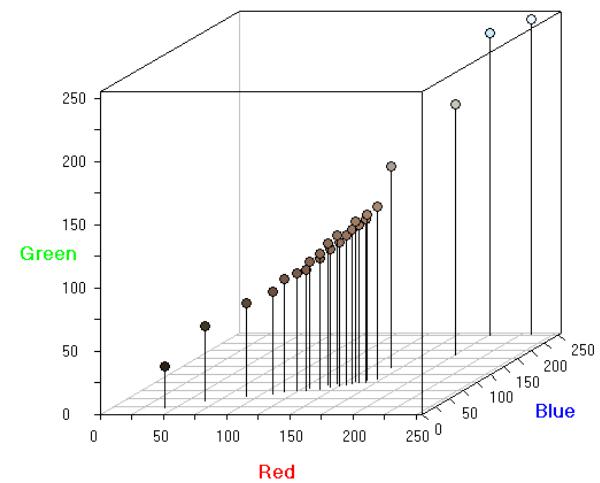
Fig. 3.f: Centroids of the Clustering of Fig. 1
with ρ = 0.99998
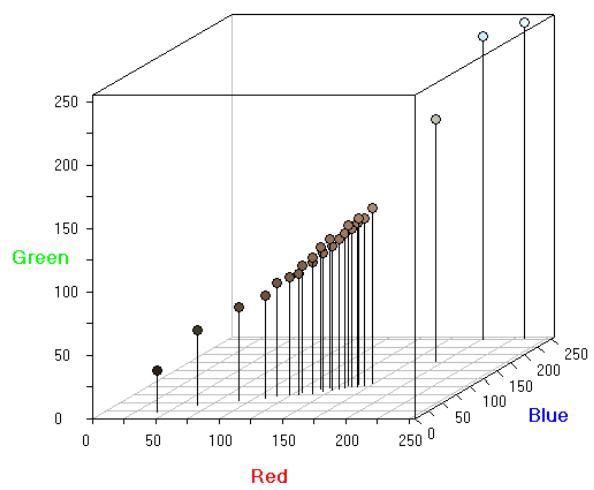
Fig. 3.g: Centroids of the Clustering of Fig. 1
with ρ = 0.999985
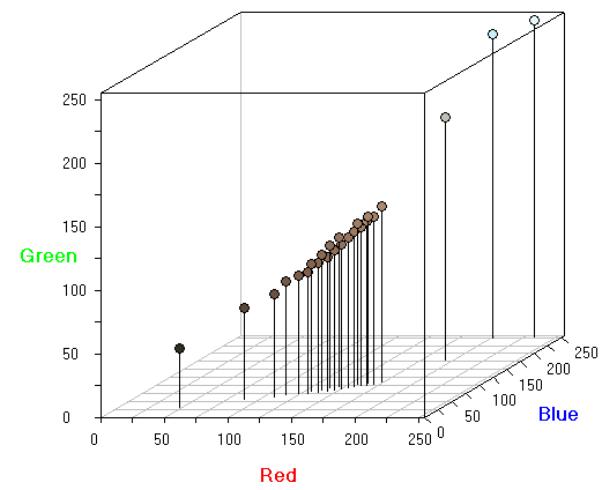
Fig. 3.h: Centroids of the Clustering of Fig. 1
with ρ = 0.99999
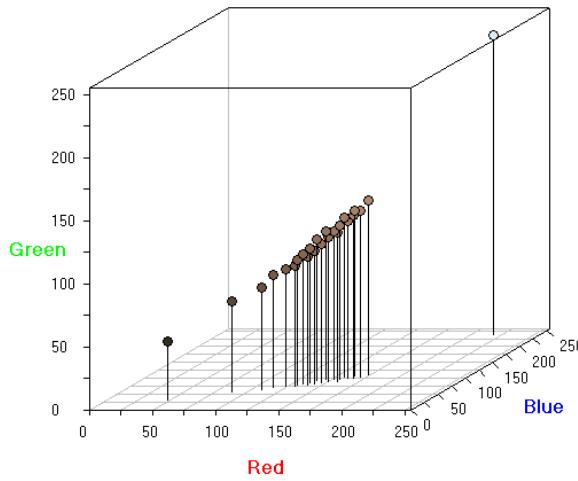
Fig. 3.i: Centroids of the Clustering of Fig. 1 with $\rho \in (0.999999 ; 0.9999999)$
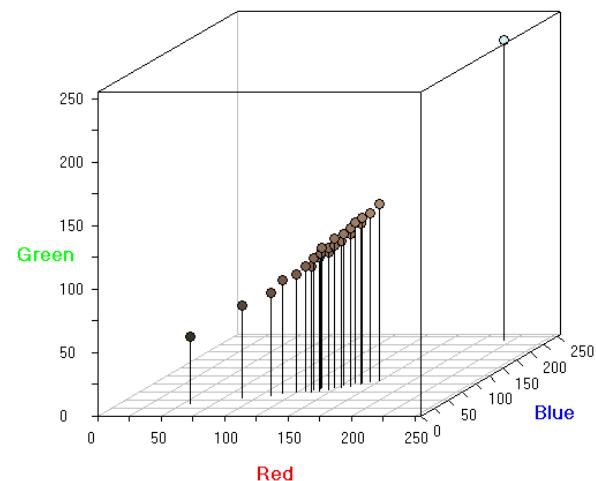
Fig. 3.j: Centroids of the Clustering of Fig. 1 with $\rho \geq 0.99999999$

The consequence of different values for ρ can be seen in Fig. 3.a – Fig. 3.j. Fig. 3.a to Fig. 3.e shows how the centroids are "pressed" towards the line between the points (0,0,0) and (255,255,255). On reason is that most of the frequent data values lie on this line. Values different from this line are much more rare. Fig. 3.f to Fig. 3.j show the "pressed" effect going on. Additionally the segment line where all centroids lie shrinks from (45,25,40) - (175,155,115) to (100,100,50) - (175,150,100). All centroids chosen from a clustering with ρ = 1 show large frequency values.

4 Summary and outlook

The described advanced maximum linkage algorithm and the postulated examples show to following results.

1. AMLA is superior to MLA. Results produced with MLA can be reproduced with AMLA easily.
2. AMLA doesn't consider frequency only theoretical but practical as well. The example shows intuitively how different considerations of the frequency of data influence the clustering. That implies that the frequency has a big influence on clustering.

At the end there are a couple of questions about the how the frequency should be taken into account. Why are the results of AMLA the same for ρ between 0 and 0.95? Why are there substantial differences for values of ρ between 0.9999 and 0.99999999? Does this effect depend on the example?
This question has to be clarified in further work.

Literature

Anderberg, M. R. (1973): *Cluster Analysis for Applications*. Academic Press, New York.

Bock, H.-H. (1998): Clustering and neural networks. In: Rizzi, A.; Vichi, M.; Bock, H.-H. (eds): *Advances in data science and classification*. Springer, Heidelberg, 265-278.

Hartigan, J.A. (1975): *Clustering algorithms.* Wiley, New York, New York.

Johnson, R.A.; Wichern, D.W. (1992): *Applied multivariate statistical analysis.* 3rd-Edition, Prentice Hall, Engelwood Cliffs, California.

Kohonen, T. (2001): *Self-organizing maps, 3rd Edition.* Springer, Heidelberg.

Myers, W.; Patil, G.P.; Taillie, C. (1997): *PHASE formulation of synoptic multivariate landscape data.* Technical report Number 97-1102. Center for statistical ecology and environmental statistics, Department of Statistics, PennState University, Pennsylvania.

Tou, J.T.; Gonzalez, R.C. (1974): *Pattern recognition principles.* Addison-Wesley, Reading, Mass.

Zerbst, M.; Tschiersch, L.; Guimarães, G.; Talbi, M.; Urfer, U. (2001): *On clustering of aerial photographs and high resolution satellite images.* Handed in ENVIRONMETRICS, Canada.

Zerbst, M. (2001): *Die pixelbasierte Clusterung von Luftaufnahmen im Rahmen von Erosionsuntersuchungen.* Dissertation, Universität Dortmund, Dortmund.