Chapter 1

Cluster Analysis in High-Dimensional Data

Clustering can be loosely defined as partitioning a set of objects into a given number k of disjoint subsets, so called clusters, so that the homogeneity between objects within each cluster is strong. Instead of homogeneity, the terms relationship or similarity are used synonymously in the literature.

Obviously, the definition given above does only make sense together with a measure for the homogeneity between objects. In this case any possible set of k clusters has a certain quality, depending on the measured homogeneity between all objects within each cluster.

One easily checks that the number of ways to partition a set of n objects in k disjoint non-void subsets is given by [18]:

$$\mathcal{K}(n,k) := \frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}(-1)^i(k-i)^n. \tag{1.1}$$

The function $\mathcal{K}(n,k)$ grows exponentially fast in n. Already in a very small set of objects the number of possible partitionings in k disjoint subsets is staggering, e.g., for $n = 100$ objects, there are $\mathcal{K}(100,2) \approx 10^{30}$ ways to partition them in two subsets. It can be shown that the problem to compute a set of k clusters of high quality is NP-complete [33]. Therefore fast solutions usually can only be achieved by using heuristic algorithms.

In addition to the identification of clusters, one is also interested in their description, i.e. in rules that allow to determine the cluster membership of each object, based on its properties. Especially in the case of high-dimensional data, where the objects have a high number of properties, such rules have to be efficient in the sense that their number is as small as possible and that they depend on a minimal number of properties only.

Given the above terminology, we define *cluster analysis in high-dimensional data* as the process of fast identification and efficient description of clusters. The clusters have to be of high quality with regard to a suitably chosen homogeneity measure.

1.1 Modeling

In the following we suggest a general model for cluster problems, supposing that the measure for the relationship between objects is given explicitly. It will be shown that the model — in contrast to other models suggested in the literature that are designed for geometric cluster problems — is usable for different fields of applications, because it is not only suitable for a geometrically based modeling, but also for dynamic cluster problems.

Let $\mathcal{A} := \{A_1, \ldots, A_q\}$ be a set of not necessarily ordered domains and define $\Omega := \bigotimes_{j=1}^{q} A_j := \{(a_1, \ldots, a_q)^T \,|\, a_j \in A_j\,, j = 1, \ldots, q\}$. We will refer to $A_1, \ldots, A_q$ as the *attributes* of Ω and to q as the *dimension* of Ω. Each finite subset $V = \{v_1, \ldots, v_n\} \subset \Omega$, $n \geq 2$, is called a *data set* in Ω and for each *data object* $v_i := (v_{i,1}, \ldots, v_{i,q})^T \in V$, the value $v_{i,j} \in A_j$ denotes the *property* of v_i for attribute A_j. We will further call each function $f : \Omega \longrightarrow \mathbf{R}_0^+$ with $f(v) = 0 \iff v \notin V$ a *frequency function* for the data set V and we define $f(M) := \sum_{v \in M} f(v)$ for any subset $M \subset \Omega$.

Suppose now that there exists a function $h : \Omega \times \Omega \longrightarrow [0, 1]$ so that $h(v, w) = h(w, v)$ for any $v, w \in V$. Then h will be called a *homogeneity function* for the data set V. We set $h_{max}(V) := \max_{v,w \in V} h(v, w)$ and call two objects $v_1, v_2 \in V$ maximally homogeneous, if $h(v_1, v_2) = h_{max}(V)$.

Based on given functions f and h the problem of clustering V in a given number k of subsets can be stated in the following general way:

Definition 1.1.1 *Let $k \in \{1, \ldots, n\}$ and $\mathcal{C} := \{C_1, \ldots, C_k\}$ any set of k non-void subsets $C_s \subset V$.*
(i) *If $\bigcup_{s=1}^{k} C_s = V$ and $C_s \cap C_t = \emptyset$ for $1 \leq s < t \leq k$, then we call $\mathcal{C}$ a k-cluster set of the data set V.*
(ii) *Let $\mathcal{C}$ any k-cluster set of V. If $\mathcal{C}$ maximizes the weighted intra-cluster homogeneity*

$$\Gamma_{f,h}(\mathcal{C}) := \frac{1}{k} \sum_{s=1}^{k} \frac{1}{f(C_s)} \sum_{v \in C_s} \sum_{w \in C_s} h(v, w) f(v) f(w) \;\rightarrow\; \max, \tag{1.2}$$

then we call $\mathcal{C}$ an optimal k-cluster set of (V, f, h).

1.1.1 Geometric cluster problems

Many of the traditional clustering methods, including the famous k*-means* method [46], have in common that they are geometrically driven, i.e. they suppose that Ω can be modeled as a metric space, e.g., $\Omega \subset \mathbf{R}^q$, and that the relationship between objects is given by a *distance function* $d : \Omega \longrightarrow \mathbf{R}_0^+$, satisfying the following requirements for all $v, w, z \in \Omega$:

$$\begin{array}{ll} (D1) & d(v,w) \geq 0 \\ (D2) & d(v,v) = 0 \\ (D3) & d(v,w) = d(w,v) \\ (D4) & d(v,w) \leq d(v,z) + d(z,w). \end{array}$$

In the case that $\Omega \subset \mathbf{R}^q$, the *Euclidean distance* function is often used:

$$d_{euclid}(v,w) := \|v - w\| := \sqrt{(v-w)^T(v-w)}\ , v, w \in \mathbf{R}^q.$$

The basic idea of almost all geometrically driven cluster methods is the identification of a k-cluster set $\mathcal{C} := \{C_1, \ldots, C_k\}$ so that $\sum_{s=1}^{k} \operatorname{cost}(C_s)$ is minimized, where $\operatorname{cost} : \wp(\Omega) \longrightarrow \mathbf{R}_0^+$ is a cost function based on the distance function. The methods differ in the choice of the cost and the distance function and the several possible optimization strategies lead to different cluster algorithms. Many popular algorithms try to minimize the *sum-of-squares* cost function [20]:

$$\operatorname{cost}(C_s) := \frac{1}{f(C_s)} \sum_{v \in C_s} \sum_{w \in C_s} d(v,w)^2 f(v) f(w) \ \rightarrow \ \min\ .$$

The corresponding cluster problem can be formulated within our general definition:

Lemma 1.1.2 *Let Ω be a metric space with a distance function $d : \Omega \longrightarrow \mathbf{R}_0^+$. Further let $V := \{v_1, \ldots, v_n\} \subset \Omega$, $n \geq 2$, be any finite data set in Ω and $f : V \longrightarrow \mathbf{R}_0^+$ be any frequency function for V. Finally suppose that $\mathcal{C}$ is any k-cluster set of V.*
(a) Then $h_d : \Omega \times \Omega \longrightarrow [0,1]$, with

$$h_d(v,w) := 1 - \frac{d(v,w)^2}{(\max_{\widetilde{v},\widetilde{w} \in V} d(\widetilde{v},\widetilde{w}))^2}\ \ , v, w \in \Omega.$$

is a homogeneity function for V.
(b) $\mathcal{C}$ is an optimal k-cluster set of (V, f, h), if and only if

$$\sum_{s=1}^{k} \frac{1}{f(C_s)} \sum_{v \in C_s} \sum_{w \in C_s} d(v,w)^2 f(v) f(w) \ \rightarrow \ \min\ .$$

Proof: (a) h_d is well defined, because $h_d(v,w) \in [0,1]$ for all $v, w \in \Omega$. Since d is a distance function, i.e. $d(v,w) = d(w,v)$ for any $v, w \in \Omega$, one further checks that $h_d(v,w) = h_d(w,v)$ and therefore h_d is a homogeneity function.
(b) Since $\max_{\widetilde{v},\widetilde{w} \in V} d(\widetilde{v},\widetilde{w})$, $f(V)$ are constant and positive values, we have:

$$
\begin{aligned}
& \min \sum_{s=1}^{k} \frac{1}{f(C_s)} \sum_{v \in C_s} \sum_{w \in C_s} d(v,w)^2 f(v) f(w) \\
\Longleftrightarrow \quad & \min \sum_{s=1}^{k} \frac{1}{f(C_s)} \sum_{v \in C_s} \sum_{w \in C_s} \frac{d(v,w)^2}{(\max_{\widetilde{v},\widetilde{w} \in V} d(\widetilde{v},\widetilde{w}))^2} f(v) f(w) \\
\Longleftrightarrow \quad & \max\ f(V) - \sum_{s=1}^{k} \frac{1}{f(C_s)} \sum_{v \in C_s} \sum_{w \in C_s} \frac{d(v,w)^2}{(\max_{\widetilde{v},\widetilde{w} \in V} d(\widetilde{v},\widetilde{w}))^2} f(v) f(w) \\
\Longleftrightarrow \quad & \max \sum_{s=1}^{k} \left(f(C_s) - \frac{1}{f(C_s)} \sum_{v \in C_s} \sum_{w \in C_s} \frac{d(v,w)^2}{(\max_{\widetilde{v},\widetilde{w} \in V} d(\widetilde{v},\widetilde{w}))^2} f(v) f(w) \right) \\
\Longleftrightarrow \quad & \max \sum_{s=1}^{k} \frac{1}{f(C_s)} \left(f(C_s)^2 - \sum_{v \in C_s} \sum_{w \in C_s} \frac{d(v,w)^2}{(\max_{\widetilde{v},\widetilde{w} \in V} d(\widetilde{v},\widetilde{w}))^2} f(v) f(w) \right) \\
\Longleftrightarrow \quad & \max \sum_{s=1}^{k} \frac{1}{f(C_s)} \sum_{v \in C_s} \sum_{w \in C_s} \left(1 - \frac{d(v,w)^2}{(\max_{\widetilde{v},\widetilde{w} \in V} d(\widetilde{v},\widetilde{w}))^2} \right) f(v) f(w) \\
\Longleftrightarrow \quad & \max \frac{1}{k} \sum_{s=1}^{k} \frac{1}{f(C_s)} \sum_{v \in C_s} \sum_{w \in C_s} h_d(v,w) f(v) f(w).
\end{aligned}
$$

□

If $d = d_{euclid}$, then the sum-of-squares cost function is equivalent to the cost function used by algorithms based on the k-means method:

Lemma 1.1.3 *Let $C \subset V \subset \mathbf{R}^q$ any non-void subset of V and $f : \Omega \longrightarrow \mathbf{R}_0^+$ any frequency function for the data set V. Then we have*

$$
\sum_{v \in C} \|v - \bar{m}_C\|^2 f(v) = \frac{1}{2} \frac{1}{f(C)} \sum_{v \in C} \sum_{w \in C} \|v - w\|^2 f(v) f(w),
$$

where

$$
\bar{m}_C := \frac{1}{f(C)} \sum_{v \in C} f(v) v
$$

denotes the centroid of C.

Proof:

$$
\begin{aligned}
& \sum_{v \in C} \|v - \bar{m}_C\|^2 f(v) \\
= \quad & \sum_{v \in C} v^T v f(v) - 2 \left(\sum_{v \in C} f(v) v^T \right) \bar{m}_C + \sum_{v \in C} f(v) \bar{m}_C^T \bar{m}_C \\
= \quad & \sum_{v \in C} v^T v f(v) \; - \; f(C) \bar{m}_C^T \bar{m}_C \\
= \quad & \frac{1}{f(C)} \left(\sum_{v \in C} f(C) v^T v f(v) - f(C)^2 \bar{m}_C^T \bar{m}_C \right) \\
= \quad & \frac{1}{f(C)} \left(\sum_{v \in C} \sum_{w \in C} v^T v f(v) f(w) - \sum_{v \in C} \sum_{w \in C} v^T w f(v) f(w) \right) \\
= \quad & \frac{1}{2} \frac{1}{f(C)} \left(2 \sum_{v \in C} \sum_{w \in C} v^T v f(v) f(w) - 2 \sum_{v \in C} \sum_{w \in C} v^T w f(v) f(w) \right) \\
= \quad & \frac{1}{2} \frac{1}{f(C)} \sum_{v \in C} \sum_{w \in C} \big(v^T v f(v) f(w) - 2 v^T w f(v) f(w) + w^T w f(w) f(v) \big) \\
= \quad & \frac{1}{2} \frac{1}{f(C)} \sum_{v \in C} \sum_{w \in C} \|v - w\|^2 f(v) f(w)
\end{aligned}
$$

□

A combination of Lemma 1.1.2 and Lemma 1.1.3 guarantees that geometric cluster problems, where the k-means method is suitable, can always be formulated within the suggested general model. Figure 1.1 shows a simple example of such a cluster problem in R^2 with $k = 3$. In the following sections, we will use this example for demonstration purposes.

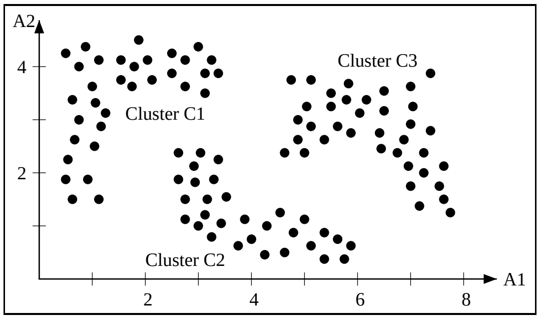


Figure 1.1: **Example: Clustering of data set in R^2 with $k = 3$.**

1.1.2 Dynamic cluster problems

Recently new cluster methods have been suggested using homogeneity measures not derived from a distance function or a more general data model [1, 5, 36]. The reason for this conceptual change is the emergence of new fields of application for cluster analysis, like e.g., the clustering of web-pages or of genomic data, where a geometrically driven modeling is often not suitable.

One of these new fields of application is the the analysis of dynamic systems. Here, an interesting problem is the identification of metastable sets of states, i.e. sets of states with a high probability that the dynamic system moves between states within the same set and a low probability of transitions between states of different sets. Although the state space of a dynamic system might be modeled as a geometric space, it is not advisable to equate metastable sets with geometrically based clusters inside this space: The dynamics between different states may not only depend on their geometric similarity. In the following we transform the identification of metastable sets of states of a dynamic system in a dynamic cluster problem, which will be described within our general model.

Let Ω be the set of all possible states of a dynamic system and choose any representative trajectory $X(1), \dots, X(T) \in \Omega$. Set $V := \{X(t) \,|\, t = 1, \dots, T\}$ and define a frequency function $f := \Omega \longrightarrow \mathbf{R}_0^+$ via $f(v) := |\{t \,|\, X(t) = v, \}|$, where $|M|$ denotes the number of elements in a finite set M. Further define for any $v, w \in V$:

$$S(v,w) := \frac{|\{t \,|\, X(t) = v, X(t+1) = w\}|}{f(v)} \tag{1.3}$$

so that $S(v, w)$ is the conditional probability of transitions from state v to state w in a single step. We can directly extend S on subsets of V, if we define for any non-void subsets $V_1, V_2 \subset V$:

$$\hat{S}(V_1, V_2) := \sum_{v \in V_1} \sum_{w \in V_2} \frac{f(v) S(v,w)}{f(V_1)}. \tag{1.4}$$

One easily checks that $\hat{S}(V_1, V_2)$ is the conditional probability of the dynamic system being in a state of set V_1 to move to a state of set V_2 in a single step.

The identification of k metastable sets of states of a dynamic system corresponds to the computation of k disjoint subsets $C_s \subset V$ so that $\hat{S}(C_s, C_s) \approx 1$ for $s = 1, \dots, k$. Since this is equivalent to a maximization of $\sum_{s=1}^{k} \hat{S}(C_s, C_s)$, the identification of k metastable sets is equivalent to the identification of an optimal k-cluster set for (V, f, h_S) where h_S is a suitable homogeneity function:

Lemma 1.1.4 *Define* $h_S : \Omega \times \Omega \longrightarrow [0,1]$ *via*

$$h_S(v,w) := \begin{cases} \frac{1}{2}\left(\frac{S(v,w)}{f(w)} + \frac{S(w,v)}{f(v)}\right) & \text{if } v,w \in V \\ 0 & \text{else} \end{cases}$$

Then h_S *is a homogeneity function of* V.

Proof: Since $0 \leq |\{t \,|\, X(t) = v, X(t+1) = w\}| \leq f(v)$ for all $v, w \in V$, we have $S(v,w) \in [0,1]$. Therefore h_S is well defined and one easily checks that $h_S(v,w) = h_S(w,v)$ for any $v, w \in V$. □

Lemma 1.1.5 *For any* k*-cluster set* $\mathcal{C}$ *of* V *the weighted intra-cluster homogeneity with respect to* f *and* h_S *is given by*

$$\Gamma_{f,h_S}(\mathcal{C}) = \frac{1}{k}\sum_{s=1}^{k} \hat{S}(C_s, C_s).$$

Proof:

$$\begin{aligned}
\Gamma_{f,h_S}(\mathcal{C}) &= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{f(C_s)}\sum_{v\in C_s}\sum_{w\in C_s} h_S(v,w)f(v)f(w) \\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{f(C_s)}\sum_{v\in C_s}\sum_{w\in C_s}\frac{1}{2}\left(f(v)S(v,w) + f(w)S(w,v)\right) \\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{f(C_s)}\frac{1}{2}\left(\sum_{v\in C_s} f(v)\sum_{w\in C_s} S(v,w) + \sum_{w\in C_s} f(w)\sum_{v\in C_s} S(w,v)\right) \\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{f(C_s)}\sum_{v\in C_s} f(v)\sum_{w\in C_s} S(v,w) \;=\; \frac{1}{k}\sum_{s=1}^{k}\hat{S}(C_s,C_s)
\end{aligned}$$

□

1.2 Problem reduction via representative clustering

A point very critical within the application of algorithms for the identification of clusters in high-dimensional data is the computational complexity, i.e. the correspondence between the time one needs to compute a solution and the number of data objects n, respectively the number of attributes q.

Suppose we have an algorithm that computes an optimal k-cluster set $\mathcal{C}$ of a data set V of size n and dimension q with respect to a frequency function f and a homogeneity function h. One easily checks that we need $\mathcal{O}(n^2)$ values $h(v, w)$ to compute the weighted intra-cluster homogeneity $\Gamma_{f,h}(\mathcal{C})$. This usually makes a direct optimization of $\Gamma_{f,h}(\mathcal{C})$ impossible, if the number n is large. In the literature several heuristic optimization approaches are suggested, but unfortunately, most algorithms are designed for special applications and are therefore not generally usable. Moreover a mathematical justification is very often missing. In the following, we will describe another way to deal with large data sets that is motivated by principles of vector quantization and signal compression (see [35]) and that we will call *representative clustering*.

The reduction of cluster problems to a handier size via representative clustering rests upon the following assumption:

Optimal cluster assumption

Let $\mathcal{C}$ be any optimal k-cluster set of a data set $V \subset \Omega$ with respect to a frequency function f and a homogeneity function h. Then $\mathcal{C}$ assigns nearly maximally homogeneous objects in a predominant portion to the same cluster, i.e. if $C \in \mathcal{C}$ is any cluster and $v, w \in V$ are any data objects with $h(v, w) \leq h_{max}(V) - \epsilon$ for small $\epsilon > 0$, then usually we have: $v \in C \implies w \in C$.

Since each optimal k-cluster set of (V, f, h) maximizes the weighted intra-cluster homogeneity, this assumption should be true for most cluster problems.

Suppose now that the homogeneity function h meets the following two conditions:

- *Local maximum condition:* Objects $v_1, v_2 \in V$ are nearly maximally homogeneous, if they have nearly the same properties.
- *Global correspondence condition:* The homogeneity function h is nearly identical for any two nearly maximally homogeneous objects $v_1, v_2 \in V$:

$$h(v_1, v_2) \approx h_{max}(V) \implies h(v_1, v) \approx h(v_2, v) \text{ for all } v \in V.$$

In the case of geometric cluster problems, the possible homogeneity functions should meet the first condition and usually also the second one. For dynamic cluster problems, it is necessary that the state space Ω is build by a set of attributes. In this case moves between states with identical values for most attributes are usually very frequent, i.e. the local maximum condition holds, and typically, such states have very common dynamic properties, i.e. also the global correspondence condition holds.

If we successively replace objects $v_{i_1}, v_{i_2}, \dots$ that have nearly the same properties by a representative object w_i, e.g., $w_i := v_{i_1}$, and define for w_i a compressed frequency value $\check{f}(w_i) := f(v_{i_1}) + f(v_{i_2}) + \dots$, we come out with a data set $W = \{w_1, w_2, \dots\}$ and a compressed frequency function $\check{f}$ of W.

Let $\mathcal{C} := \{C_1, \dots, C_k\}$ be any optimal k-cluster set of $(W, \check{f}, h)$, then we can extend $\mathcal{C}$ on V, if we define $\hat{\mathcal{C}} := \{\hat{C}_1, \dots, \hat{C}_k\}$ with $\hat{C}_s := \bigcup_{w_i \in C_s} \{v_{i_1}, v_{i_2}, \dots\}$. Obviously $\hat{\mathcal{C}}$ is a k-cluster set of V. The local maximum condition assures that w_i and $v \in \{v_{i_1}, v_{i_2}, \dots\}$ are nearly maximally homogeneous. Therefore the global correspondence condition guarantees:

$$\begin{aligned}
&\Gamma_{\check{f},h}(\mathcal{C}) \\
= \;& \frac{1}{k}\sum_{s=1}^{k} \frac{1}{\check{f}(C_s)} \sum_{w_i \in C_s} \sum_{w_j \in C_s} h(w_i, w_j)\check{f}(w_i)\check{f}(w_j) \\
= \;& \frac{1}{k}\sum_{s=1}^{k} \frac{1}{f(\hat{C}_s)} \sum_{w_i \in C_s} \sum_{w_j \in C_s} h(w_i, w_j) \sum_{v_1 \in \{v_{i_1}, v_{i_2}, \dots\}} f(v_1) \sum_{v_2 \in \{v_{j_1}, v_{j_2}, \dots\}} f(v_2) \\
= \;& \frac{1}{k}\sum_{s=1}^{k} \frac{1}{f(\hat{C}_s)} \sum_{w_i \in C_s} \sum_{v_1 \in \{v_{i_1}, v_{i_2}, \dots\}} \sum_{w_j \in C_s} \sum_{v_2 \in \{v_{j_1}, v_{j_2}, \dots\}} h(w_i, w_j) f(v_1) f(v_2) \\
\approx \;& \frac{1}{k}\sum_{s=1}^{k} \frac{1}{f(\hat{C}_s)} \sum_{v_1 \in \hat{C}_s} \sum_{w_j \in C_s} \sum_{v_2 \in \{v_{j_1}, v_{j_2}, \dots\}} h(v_1, w_j) f(v_1) f(v_2) \\
\approx \;& \frac{1}{k}\sum_{s=1}^{k} \frac{1}{f(\hat{C}_s)} \sum_{v_1 \in \hat{C}_s} \sum_{v_2 \in \hat{C}_s} h(v_1, v_2) f(v_1) f(v_2) \\
= \;& \Gamma_{f,h}(\hat{\mathcal{C}}).
\end{aligned}$$

Suppose now that $\hat{\mathcal{C}}$ is not nearly optimal for (V, f, h). Then the optimal cluster assumption guarantees that there exist objects $v_1, v_2 \in V$ that are assigned to different clusters in $\hat{\mathcal{C}}$, although $h(v_1, v_2)$ is large. But this is a contradiction to the fact that nearly homogeneous objects are replaced by the same representative and therefore are assigned to the same cluster in $\hat{\mathcal{C}}$.

Let $V(j) := \{v_{*,j} \mid v = (v_{*,1}, \dots, v_{*,q})^T \in V\}$ be the projection of V on the attribute A_j. Set $V_\Omega := \bigotimes_{j=1}^{q} V(j) = \{(a_1, \dots, a_q)^T \mid a_j \in V(j)\,,\, j = 1, \dots, q\}$. Obviously we have $V \subset V_\Omega \subset \Omega$ and $n = |V| \le |V_\Omega| \le n^q$. When analyzing high-dimensional data one often observes that V_Ω is rather sparse with respect to V, i.e. the *sparsity factor* $\frac{|V|}{|V_\Omega|}$ is very small. This guarantees that $|W|$ is smaller than n, i.e. we have reduced our cluster problem.

Figure 1.2 shows a reduction of our geometric cluster problem in R^2 via representative clustering in principle.

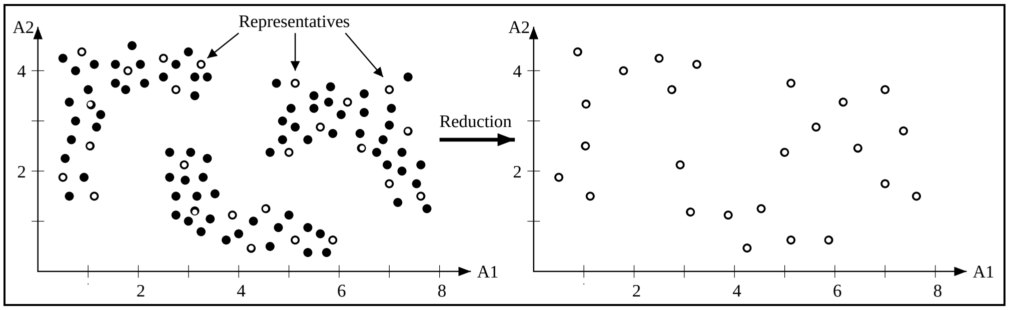


Figure 1.2: **Example: Reduction of geometric cluster problem in R^2.**

A problem reduction via representative clustering is only efficient, if $|W|$ is significantly smaller than the number n. Obviously the number of representatives depends strongly on the criterion that is used for the identification of objects with nearly the same properties. As a brute force approach one could think about using a very weak criterion that allows to replace much objects by the same representative. In this case the local maximum condition only holds, if we call two objects v_1, v_2 nearly maximal homogeneous, even if $h(v_1, v_2)$ is not so high. But then we cannot be sure that their homogeneity in relation to all other objects is nearly identical, i.e. that $h(v_1, v) \approx h(v_2, v)$ holds for all $v \in V$. If the global correspondence condition is violated too often, this usually has negative consequences for the quality of $\hat{\mathcal{C}}$.

In chapter 2 we will describe a concept called *decomposition* that can be used as a basis for the development of methods for an efficient problem reduction via representative clustering. We will replace the global correspondence condition for h by the construction of a compressed homogeneity function $\check{h}$ and define a more convenient condition that guarantees the optimality of $\hat{\mathcal{C}}$, if $\mathcal{C}$ is an optimal k-cluster set of $(W, \check{f}, \check{h})$. Moreover in chapter 4 a multilevel approach is presented that uses decomposition based representative clustering for a fast cluster identification.

1.3 Efficient cluster description

Besides the identification of clusters in high-dimensional data, also their efficient description is very important for most practical applications (see chapter 5). We want to know, which objects are homogeneous and also why they are homogeneous.

Obviously such a description can be achieved via rules that allow to determine the cluster membership of each object, based on its properties, i.e. rules like:

If $v = (v_{*,1}, \dots, v_{*,q})^T \in V$ *has the properties* $v_{*,1} = a_1$ *and* $\dots$ *and* $v_{*,q} = a_q$*, then* v *belongs to cluster* C_s*.*

A description based on such rules has to be consistent, i.e. it contains no rules assigning the same object v to different clusters.

Given any k-cluster set $\mathcal{C} := \{C_1, \dots, C_k\}$ of a data set V in Ω, we can always generate rules for a cluster description in the following trivial way:

Define a function $c_\chi : V \longrightarrow \{1, \dots, k\}$ via

$$c_\chi(v) := \sum_{s=1}^{k} s\, \chi_{C_s}(v) \quad \text{for all } v \in V,$$

where χ_{C_s} denotes the characteristic function of cluster C_s. Then for any object $v_i := (v_{i,1}, \dots, v_{i,q})^T \in V$ we can state a rule r_i:

If $v = (v_{*,1}, \dots, v_{*,q})^T$ *has the properties* $v_{*,1} = v_{i,1}$ *and* $\dots$ *and* $v_{*,q} = v_{i,q}$*, then* v *belongs to cluster* $C_{c_\chi(v_i)}$*.*

Obviously the n rules $r_1, \dots, r_n$ describe the clusters $C_1, \dots, C_k$ consistently, but such a description is surely not efficient. We will demonstrate this by our example of a geometric cluster problem in $\mathbf{R}^2$ (see Fig. 1.1):

Cluster C_1 contains 33 data objects, i.e. we need 33 rules to describe this cluster if we use our trivial approach. If we allow rules that are slightly more complex, one easily checks that the following two rules are sufficient to describe cluster C_1:

If $v = (v_{*,1}, v_{*,2})^T$ *has the properties* $v_{*,1} = a_1$ *and* $v_{*,2} = a_2$ *with* $a_1 \in [0,2]$*,* $a_2 \in [1,5]$*, then* v *belongs to cluster* C_1*.*

If $v = (v_{*,1}, v_{*,2})^T$ *has the properties* $v_{*,1} = a_1$ *and* $v_{*,2} = a_2$ *with* $a_1 \in [2,4]$*,* $a_2 \in [3,5]$*, then* v *belongs to cluster* C_1*.*

This motivates the following definition of cluster membership rules:

Definition 1.3.1 *For any set* $\mathcal{B} := \{B_1, \dots, B_q\}$ *with* $B_j \subset A_j$ *for* $j = 1, \dots, q$*, we call* $r_{\mathcal{B}} : \Omega \longrightarrow \{0,1\}$ *with*

$$r_{\mathcal{B}}(v) := \begin{cases} 1 & \text{if } (\forall j \in \{1, \dots, q\})\ v_{*,j} \in B_j \\ 0 & \text{else} \end{cases} \quad , \quad v := (v_{*,1}, \dots, v_{*,q})^T \in \Omega,$$

a membership rule for cluster C_s*, if*

$$r_{\mathcal{B}}(v) = 1 \implies v \in C_s \quad \text{for all } v \in V.$$

Usually we need a set $r_s := \{r_{s,1}, \ldots, r_{s,m_s}\}$ of $m_s \in \mathbf{N}^+$ membership rules for each cluster C_s, to guarantee that each object $v \in C_s$ is assigned to cluster C_s by at least one rule, i.e. that we have

$$v \in C_s \implies (\exists\, r \in r_s)\, r(v) = 1 \quad \text{for all } v \in V.$$

We call such a set r_s a *complete membership rule set* for cluster C_s.

Based on complete membership rule sets for each cluster C_s, we can easily generate a description of $\mathcal{C}$:

Lemma 1.3.2 *Suppose there exists for each Cluster C_s of $\mathcal{C}$ a complete membership rule set $r_s := \{r_{s,1}, \ldots, r_{s,m_s}\}$. Let $\mathcal{H}_0$ denote the Heaviside function with*

$$\mathcal{H}_0(t) := \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0. \end{cases}$$

Then the function $c_r : V \longrightarrow \{1, \ldots, k\}$ with

$$c_r(v) := \sum_{s=1}^{k} s\, \mathcal{H}_0(-1 + \sum_{j=1}^{m_s} r_{s,j}(v)) \quad \text{for all } v \in V.$$

is a consistent description for $\mathcal{C}$, i.e. we have

$$c_r(v) = s \iff v \in C_s \quad \text{for all } v \in V.$$

Proof: "$\Longleftarrow$": Choose any $s \in \{1, \ldots, k\}$ and any $v \in C_s$. Since r_s is a complete membership rule set, there exists an $t \in \{1, \ldots, m_s\}$ so that $r_{s,t}(v) = 1$. Therefore we have $\mathcal{H}_0(-1 + \sum_{j=1}^{m_s} r_{s,j}(v)) = 1$. Suppose now that there exists another $p \in \{1, \ldots, k\}$ with $p \neq s$ and $\mathcal{H}(-1 + \sum_{j=1}^{m_p} r_{p,j}(v)) = 1$. If this is the case, there must exist a $\widetilde{t} \in \{1, \ldots, m_p\}$ so that $r_{p,\widetilde{t}}(v_i) = 1$. Since $r_{p,\widetilde{t}}$ is a membership rule for Cluster C_p, this implies $v \in C_p$. But this is a contradiction to $v \in C_s$. Therefore we have $c_r(v) = s$.
"$\Longrightarrow$": Choose any $s \in \{1, \ldots, k\}$ and any $v \in V \setminus C_s$. Since $\mathcal{C}$ is a k-cluster set of V there exists a $p \in \{1, \ldots, k\}$ with $p \neq s$ and $v \in C_p$. As already proofed above this guarantees $c_r(v) = p$ and therefore $c_r(v) \neq s$. □

Let $v = (v_{*,1}, \ldots, v_{*,q}) \in V$ be any data object and let $c_r : V \longrightarrow \{1, \ldots, k\}$ be a consistent description of $\mathcal{C}$ with corresponding complete membership rule sets $r_1, \ldots, r_k$. Then the determination of the cluster membership of v is rather simple: Find a membership rule $r_{\mathcal{B}} \in \bigcup_{s=1}^{k} r_s$ with $r_{\mathcal{B}}(v) = 1$, i.e, with $v_{*,j} \in B_j$ for $j = 1, \ldots, q$. Since c_r is consistent, there exists exactly one $s \in \{1, \ldots, k\}$

with $r_{\mathcal{B}} \in r_s$. Therefore data object v belongs to cluster C_s. Note that the existence of more than one membership rule $r \in r_s$ with $r(v) = 1$ is possible.

Obviously descriptions should be efficient in the sense that the corresponding complete membership rule sets $r_s := \{r_{s,1}, \ldots, r_{s,m_s}\}$ are minimal. i.e. the numbers m_s are as small as possible.

Often not all properties of a data object have to be considered to determine its cluster membership. Especially in the case of high-dimensional data, with a great number q of attributes A_j, a description based on a reduced set of attributes is of great interest.

We will illustrate this again by our two-dimensional example. Suppose that we restrict our data set to the data objects of cluster C_1 and cluster C_3. Then the following two rules will be sufficient to describe the clusters:

If $v = (v_{*,1}, v_{*,2})^T$ has the property $v_{*,1} = a_1$ with $a_1 \in [0, 4]$, then v belongs to cluster C_1.

If $v = (v_{*,1}, v_{*,2})^T$ has the property $v_{*,1} = a_1$ with $a_1 \in [4.5, 8]$, then v belongs to cluster C_3.

Obviously we only need attribute A_1 for a description of cluster C_1 and C_3, i.e. attribute A_2 has no influence on the discrimination of both clusters. Note that this is not true, for a description that includes cluster C_2.

We can easily extend our earlier definitions to work with reduced attribute sets:

Let $J := \{j_1, \ldots, j_m\} \subset \{1, \ldots, q\}$ any index subset of length m and let $\mathcal{A}(J) := \{A_j \,|\, j \in J\}$ be a reduced set of attributes of Ω. Set $\Omega(J) := \bigotimes_{j_t \in J} A_{j_t}$ and for $v := (v_{*,1}, \ldots, v_{*,q})^T \in \Omega$ denote by $v(J) := (v_{*,j_1}, \ldots, v_{*,j_m})^T \in \Omega(J)$ the projection on $\Omega(J)$. Further set $M(J) := \{v(J) \,|\, v \in M\} \subset \Omega(J)$ for any subset $M \subset \Omega$.

We can define *J-reduced membership rules* as a special kind of membership rules:

Definition 1.3.3 *Let $r_{\mathcal{B}}$ be any membership rule with $\mathcal{B} := \{B_1, \ldots, B_q\}$ and $B_j \subset A_j$ for $j = 1, \ldots, q$. We call $r_{\mathcal{B}}$ J-reduced, if $B_j = A_j$ for $j \notin J$. Let further r_s be a complete membership rule set of cluster C_s. We call r_s a complete J-reduced membership rule set, if each membership rule $r \in r_s$ is J-reduced.*

There exists an unique projection of any J-reduced membership rule on the subspace $\Omega(J)$:

Lemma 1.3.4 *Let $r_{\mathcal{B}}$ be any J-reduced membership rule with $\mathcal{B} := \{B_1, \ldots, B_q\}$ and $B_j \subset A_j$ for $j = 1, \ldots, q$. Then the function $\bar{r}_{\mathcal{B}} : \Omega(J) \longrightarrow \{0, 1\}$ with*

$$\bar{r}_{\mathcal{B}}(\bar{v}) := \begin{cases} 1 & \text{if } (\forall j \in J)\ v_{*,j} \in B_j \\ 0 & \text{else} \end{cases} \quad , \quad \bar{v} := (v_{*,j_1}, \ldots, v_{*,j_m})^T \in \Omega(J)$$

is the unique projection of $r_{\mathcal{B}}$ on $\Omega(J)$.

Proof: For any $v = (v_{*,1}, \dots, v_{*,q})^T \in \Omega$ we have $v_{*,j} \in A_j = B_j$ for $j \notin J$, and therefore $r_{\mathcal{B}}(v) = \bar{r}_{\mathcal{B}}(v(J))$. □

Analogously to Lemma 1.3.2 we can achieve a description based on the reduced set of attributes $\mathcal{A}(J)$, if there exists for each cluster a complete J-reduced membership rule set:

Lemma 1.3.5 *Let $J \subset \{1, \dots, q\}$ be any index subset of length m. Suppose there exists for each Cluster C_s of $\mathcal{C}$ a complete J-reduced membership rule set $r_s := \{r_{s,1}, \dots, r_{s,m_s}\}$ and $\bar{r}_{s,j}$ denotes the unique projection of the membership rule $r_{s,j}$ on $\Omega(J)$, then the function $c_r : V \longrightarrow \{1, \dots, k\}$ with*

$$c_r(v(J)) := \sum_{s=1}^{k} s\, \mathcal{H}_0(-1 + \sum_{j=1}^{m_s} \bar{r}_{s,j}(v(J))) \quad \text{for all}\, v \in V,$$

is a consistent description for $\mathcal{C}$ based on the reduced attribute set $\mathcal{A}(J)$, i.e. we have

$$c_r(v(J)) = s \iff v \in C_s \quad \text{for all } v \in V.$$

Obviously descriptions should be efficient in the sense that they are based on a maximally reduced attribute set $\mathcal{A}(J)$, i.e. $\mathcal{A}(J)$ should contain as less attributes as possible.

Efficient cluster description algorithm

Using the above definitions, the following general algorithm generates an efficient cluster description for a k-cluster set $\mathcal{C} := \{C_1, \dots, C_k\}$ of a data set $V \in \Omega$:
(1) Find an index subset $J = \{j_1, \dots, j_m\} \subset \{1, \dots, q\}$ of minimal size so that there exists a function $c : V \longrightarrow \{1, \dots, k\}$ with

$$c(v(J)) = s \iff v \in C_s \quad \text{for all } v \in V.$$

(2) Compute for each cluster C_s a minimally complete J-reduced membership rule set $r_s := \{r_{s,1}, \dots, r_{s,m_s}\}$.
(3) Use $r := \{r_1, \dots, r_k\}$ to construct a consistent description c_r of $\mathcal{C}$ based on the reduced attribute set $\mathcal{A}(J)$.

Since we are analyzing high-dimensional data, i.e. the dimension q is large, we obviously need heuristic solutions for step (1) and (2). For the development of suitable methods the concept of decomposition is very helpful: In section 2.4 we will describe techniques for the computation of membership rule sets based on *approximate box decompositions* and we will introduce the concept of *discriminating attributes* that allows the construction of heuristic algorithms to identify optimally reduced attribute sets $\mathcal{A}(J)$.

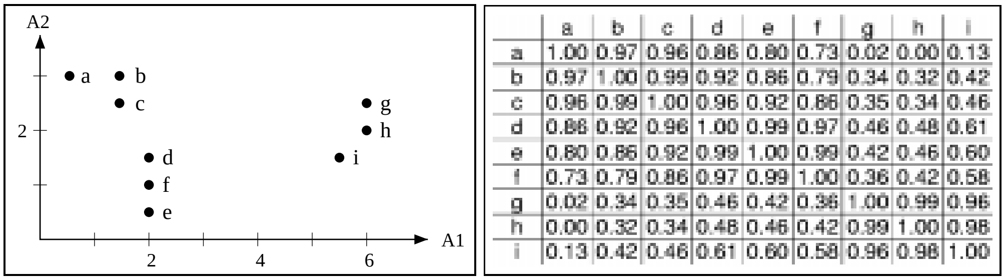
1.4 How many clusters?

Up to now, we have supposed that the number of clusters k is known a priori. But in many real world applications this is not the case. Looking at Eq. (1.1) one easily checks that the number of possible k-cluster sets explodes, if k is a further unknown parameter of the cluster problem. Obviously k is the most important parameter, i.e. with the words of cluster expert J. BEZDEK: *"It is clearly more important to be looking in the right solution space (within k) than it is to be comparing partitions across k because k specifies the number of clusters to look for, while the other parameters control the search for these substructures."* [6].

The definition of a general model for cluster problems with unknown cluster number is still an open problem. Usually it is not suitable to determine a correct number of clusters by computing for different k the optimal k-cluster sets $\mathcal{C}(k)$ and comparing the weighted intra-cluster homogeneities $\Gamma_{f,h}(\mathcal{C}(k))$, because most homogeneity functions tend to prefer extreme clusterings with $k = 1$ or $k = n$.

Example: Cluster problem with unknown number of clusters

We will illustrate this by the following simple example: Suppose we want to compute an optimal clustering of a data set $V = \{a, b, c, d, e, f, g, h, i\} \subset \mathbf{R}^2$ with a frequency function so that $f(v) = 1$ for all $v \in V$. We choose $h = h_d$ (see Lemma 1.1.2) based on the Euclidean distance function $d = d_{euclid}$. Figure 1.3 shows a plot of V and the corresponding homogeneity matrix.



	a	b	c	d	e	f	g	h	i
a	1.00	0.97	0.96	0.86	0.80	0.73	0.02	0.00	0.13
b	0.97	1.00	0.99	0.92	0.86	0.79	0.34	0.32	0.42
c	0.96	0.99	1.00	0.96	0.92	0.86	0.35	0.34	0.46
d	0.86	0.92	0.96	1.00	0.99	0.97	0.46	0.48	0.61
e	0.80	0.86	0.92	0.99	1.00	0.99	0.42	0.46	0.60
f	0.73	0.79	0.86	0.97	0.99	1.00	0.36	0.42	0.58
g	0.02	0.34	0.35	0.46	0.42	0.36	1.00	0.99	0.96
h	0.00	0.32	0.34	0.48	0.46	0.42	0.99	1.00	0.98
i	0.13	0.42	0.46	0.61	0.60	0.58	0.96	0.98	1.00

Figure 1.3: **Example: Cluster problem in R^2 with unknown cluster number k.** Left hand side: Plot of data set V. Right hand side: Homogeneity matrix of V based on Euclidean distance.

In Table 1.1 the optimal k-cluster sets $\mathcal{C}(k)$ of (V, f, h) and their weighted intra-cluster homogeneities $\Gamma_{f,h}(\mathcal{C}(k))$ are presented for different k. Obviously one would expect $k = 2, 3$ or 4 as a correct number of clusters, but a maximization of $\Gamma_{f,h}(\mathcal{C}(k))$ leads always to $k = 1$. Therefore we cannot use $\Gamma_{f,h}(\mathcal{C}(k))$ to judge which k is best.

optimal k-cluster set $\mathcal{C}(k)$	$\Gamma_{f,h}(\mathcal{C}(k))$
$\mathcal{C}(1) := V$	6.17
$\mathcal{C}(2) := \{\{a,b,c,d,e,f\},\{g,h,i\}\}$	4.24
$\mathcal{C}(3) := \{\{a,b,c\},\{d,e,f\},\{g,h,i\}$	2.96
$\mathcal{C}(4) := \{\{a\},\{b,c\},\{d,e,f\},\{g,h,i\}\}$	2.23
$\mathcal{C}(9) := \{\{a\},\{b\},\{c\},\{d\},\{e\},\{f\},\{g\},\{h\},\{i\}\}$	1.00

Table 1.1: **Example: Optimal k-cluster sets of (V, f, h) for different k.**

In the literature [42, 6, 25, 51] several other measures are suggested to determine the validity of a given k-cluster set and so to find the optimal clustering, but all of these measures have the deficit that they first need the computation of optimal k-cluster sets for different k. In the worst scenario this requires the solution of n optimization problems. If n is large, this is a really heroic task.

Another possibility to cope with the problem of the unknown number of clusters might be to determine it in a pre-processing step. Via a projection of the high-dimensional data on a two-dimensional plane, one hopes that the cluster structure is not destroyed through the transformation and the number of clusters can be determined by visual investigation. A very popular tool for such a projection are *multidimensional-scaling* methods [49], e.g., SAMMON'S non-linear mapping algorithm [56]. The deficits of projection methods are obvious: For high-dimensional data it is unlikely that the cluster structure on the two-dimensional plane reflects the original structure. Moreover a visual investigation could be very subjective.

For cluster problems with a special type of homogeneity functions, exhibiting a stochastic property, we will present in chapter 4 a new method based on the theory of *Perron Cluster* analysis that allows the computation of a correct number of clusters. We will show that this method can be easily used together with the suggested multilevel cluster identification approach.