Chapter 2

Decomposition

In different research fields, decomposition usually describes the process of splitting a problem in smaller problems with less complexity. As was already motivated in section 1.2, a suitable reduction of a cluster problem can be achieved via a grouping of nearly maximally homogeneous objects and a representation of each group by a single object with compressed frequency value. If this kind of partitioning of the data set V exhibits a certain homogeneity property, we will call it a decomposition. After giving a general definition, we will introduce a special type of decomposition, the so called *approximate box decomposition*. Here the objects are pre-grouped in a way that they build a special subspace in Ω that has the shape of a multidimensional box if Ω is a metric space. We will develop a theory for an efficient reduction of cluster problems via representative clustering based on decomposition and we will present a basic reduction algorithm that will be refined in chapter 4. Finally we will show how an approximate box decomposition can be used to derive an efficient cluster description based on a minimal number of so called *discriminating attributes*.

2.1 General Definition

Let $V = \{v_1, \ldots, v_n\} \subset \Omega$ be any data set in Ω with frequency function f and homogeneity function h.

Definition 2.1.1 *Assume $n_k \in \mathbf{N}$ with $n_k \leq n$ and $\epsilon \in \mathbf{R}_0^+$ with $\epsilon \leq h_{max}(V)$. We call $\Theta := \{\Theta_1, \ldots, \Theta_{n_k}\}$ an ϵ-decomposition of (V, h) with partitions Θ_s, if*

$$\bigcup_{s=1}^{n_k} \Theta_s = V \ , \quad \Theta_s \neq \emptyset \, , \quad \Theta_s \cap \Theta_p = \emptyset \quad \text{for } 1 \leq s < p \leq n_k$$

$$\text{and} \quad h(v, w) \geq h_{max}(V) - \epsilon \ \text{ for all } v, w \in \Theta_s, s = \{1, \ldots, n_k\}.$$

We further call

$$\vartheta_{f,h}(\Theta) := \frac{1}{f(V)} \sum_{s=1}^{n_k} \frac{1}{f(\Theta_s)} \sum_{v \in \Theta_s} \sum_{w \in \Theta_s} (h_{max}(V) - h(v,w)) f(v) f(w) \rightarrow \min$$

the decomposition error of Θ with respect to f and h.

Since $0 \leq h(v,w) \leq h_{max}(V)$ for all $v, w \in \Omega$, any n_k-clustering of V is an ϵ-decomposition of (V,h) with $\epsilon = h_{max}(V)$. The following Lemma guarantees $\vartheta_{f,h}(\Theta) \in [0, h_{max}(V)]$ for any ϵ-decomposition of (V,h):

Lemma 2.1.2 *Let Θ any ϵ-decomposition of (V,h), then we have: $\vartheta_{f,h}(\Theta) \leq \epsilon$.*

Proof: We have $(h_{max}(V) - h(v,w)) \leq \epsilon$ for all $v, w \in \Theta_s$ and therefore

$$\begin{aligned}
\vartheta_{f,h}(\Theta) &\leq \frac{1}{f(V)} \sum_{s=1}^{n_k} \frac{1}{f(\Theta_s)} \sum_{v \in \Theta_s} \sum_{w \in \Theta_s} \epsilon f(v) f(w) \\
&= \frac{\epsilon}{f(V)} \sum_{s=1}^{n_k} \frac{1}{f(\Theta_s)} \sum_{v \in \Theta_s} f(v) \sum_{w \in \Theta_s} f(w) \\
&= \frac{\epsilon}{f(V)} \sum_{s=1}^{n_k} \frac{1}{f(\Theta_s)} \sum_{v \in \Theta_s} f(v) f(\Theta_s) \\
&= \frac{\epsilon}{f(V)} \sum_{s=1}^{n_k} \sum_{v \in \Theta_s} f(v) = \frac{\epsilon}{f(V)} \sum_{s=1}^{n_k} f(\Theta_s) = \frac{\epsilon}{f(V)} f(V) = \epsilon
\end{aligned}$$

□

We will refer to Θ as a decomposition of V, if there exists a homogeneity function h and an $\epsilon \in [0, h_{max}(V)]$ so that Θ is an ϵ-decomposition of (V,h).

If we use the homogeneity measure $h = h_d$ (see Lemma 1.1.2) based on a distance function d, one easily checks that we have $h_{max} = 1$ and

$$\vartheta_{f,h}(\Theta) = \frac{1}{f(V)} \frac{1}{\max_{\widetilde{v},\widetilde{w} \in V} d(\widetilde{v},\widetilde{w})^2} \sum_{s=1}^{n_k} \frac{1}{f(\Theta_s)} \sum_{v \in \Theta_s} \sum_{w \in \Theta_s} d(v,w)^2 f(v) f(w).$$

Therefore, in this special case, we can use algorithms that try to optimize the sum-of-squares cost function to compute a decomposition for given n_k with minimal decomposition error. Figure 2.1 shows two possible decompositions with $n_k := 6$ partitions Θ_s for our example of a geometric cluster problem in R^2 using the Euclidean distance function $d = d_{euclid}$. The decomposition on the left hand side has been computed automatically via a simple hierarchical optimization method and leads to $\epsilon = 0.137$ and $\vartheta_{f,h}(\Theta) = 0.019$. The decomposition on the right hand side has been additionally optimized manually and leads to $\epsilon = 0.135$ and $\vartheta_{f,h}(\Theta) = 0.018$. Obviously ϵ is only a very rough upper bound of the decomposition error.

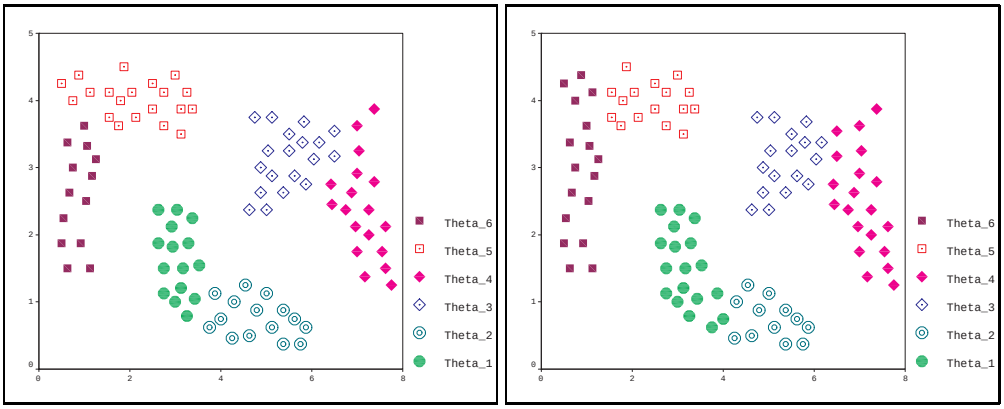

Figure 2.1: **Example: Two possible decompositions with six partitions in R^2.**

2.2 Approximate box decomposition

In the following we call any subset $B \subset \Omega$ a *box* in Ω, if there exist non-void subsets $B_1, \ldots, B_q$ with $B_j \subset A_j$ and $B = \bigotimes_{j=1}^{q} B_j$. We set $\mathrm{BOX}(\Omega) := \{B \mid B \,\mathrm{box\,in}\, \Omega\}$.

Definition 2.2.1 *Assume $n_k \in \mathbf{N}$ with $n_k \leq n$. We call (Θ,Δ) an approximate box decomposition of V with respect to f, whenever $\Theta := \{\Theta_1, \ldots, \Theta_{n_k}\}$ is a decomposition of V and Δ is a set of n_k boxes $\Delta_1, \ldots, \Delta_{n_k} \in \mathrm{BOX}(\Omega)$ so that $\mathrm{overlap}_f(\Delta) \approx 0$ and $f(\Theta_s \cap \Delta_s) > 0$ for $s = 1, \ldots, n_k$. The value $\mathrm{overlay}_f(\Theta, \Delta) \in]0, 1]$ indicates how good Δ approximates Θ.*

Herein we use the terms *overlap* and *overlay* in the following way:

Definition 2.2.2 *Let $\mathcal{M} := \{M_1, \ldots, M_k\}$ be any set of $n_k \in N$ subsets of Ω with $f(M_s) > 0$ for $s = 1, \ldots, n_k$. Let Θ be a decomposition of V with n_k partitions Θ_s. Then the overlay of Θ and $\mathcal{M}$ with respect to f is given by*

$$\mathrm{overlay}_f(\Theta, \mathcal{M}) := \frac{1}{f(V)} \sum_{s=1}^{n_k} f(M_s \cap \Theta_s), \tag{2.1}$$

whereas the overlap of $\mathcal{M}$ with respect to f is given by

$$\mathrm{overlap}_f(\mathcal{M}) := \sum_{s=1}^{k} \frac{f(M_s \cap \bigcup_{p \neq s} M_p)}{f(\bigcup_{p=1}^{k} M_p)}. \tag{2.2}$$

If $\mathrm{overlay}_f(\Theta, \Delta) = 1$, we call (Θ, Δ) a perfect box decomposition of V, Note that if $\Delta(V) := \{\Delta_1 \cap V, \ldots, \Delta_{n_k} \cap V\}$ is a decomposition of V, $(\Delta(V), \Delta)$ is always a perfect box decomposition.

Figure 2.2 presents two approximate box decompositions based on the decompositions shown in Figure 2.1. On the left hand side, the six boxes does not approximate the decomposition perfectly, because two boxes overlap each other and four points are not covered, i.e. there is an insufficient overlay. On the right hand side of Figure 2.2, the decomposition is approximated perfectly with six boxes. Note that for the automatically computed decomposition, shown on the left hand side of Figure 2.1, no perfect approximation with six boxes is possible at all.

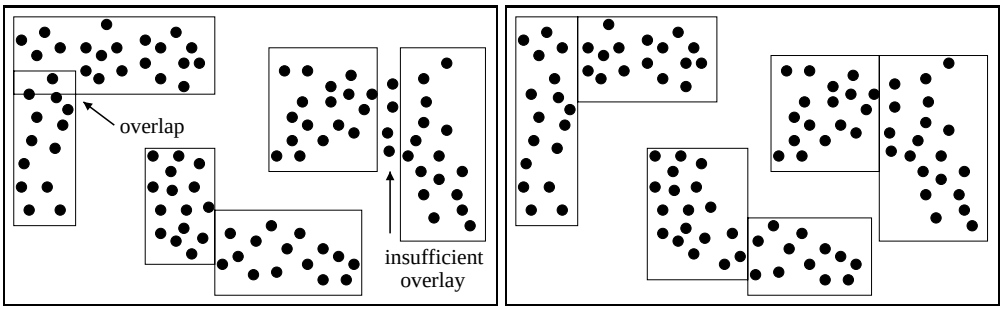

Figure 2.2: **Example: Approximate box decomposition ($n_k = 6$) in R^2.** Left hand side: Approximate box decomposition with insufficient overlay and overlap. Right hand side: Perfect box decomposition.

Example: Uniform box decomposition

We can always construct a perfect box decomposition: For $j \in \{1, \ldots, q\}$ choose any $m_j \in \mathbf{N}$ and any disjoint non-void subsets $B_{1,j}, \ldots, B_{m_j,j} \subset A_j$ so that $\bigcup_{i=1}^{m_j} B_{i,j} = A_j$. Set $m := \prod_{j=1}^{q} m_j$ and for any index tuple $(i_1, \ldots, i_q)$ with $1 \leq i_j \leq m_j$ choose an unique number $p = p(i_1, \ldots, i_q) \in \{1, \ldots, m\}$ and define $\Delta_p := \bigotimes_{j=1}^{q} B_{i_j,j}$. Obviously we have $\Delta_p \in \mathrm{BOX}(\Omega)$ for each $p \in \{1, \ldots, m\}$.

If we set $I(V) := \{p \,|\, \Delta_p \cap V \neq \emptyset\}$ and $\Delta_{I(V)} := \{\Delta_p \,|\, p \in I(V)\}$, then one easily checks that $(\Delta_{I(V)}(V), \Delta_{I(V)})$ is a perfect box decomposition of V because $\Delta_{I(V)}(V) := \{\Delta_p \cap V \,|\, p \in I(V)\}$ is a decomposition of V. Since the construction of Δ_p is uniform in the sense that each attribute of Ω is divided into m_j disjoint subsets, we call $(\Delta_{I(V)}(V), \Delta_{I(V)})$ an uniform box decomposition of Ω.

Note that the construction of the decomposition $\Delta_{I(V)}(V)$ is independent of the homogeneity function h and so the decomposition error is not guaranteed to be small. Further remember that, with increasing q, the number m grows exponentially, even if we split each attribute in only two subsets, i.e. if we set $m_j := 2$ for $j = 1, \ldots, q$. For example $q = 20$ leads to $m > 10^6$. So we usually have $m > n$ and therefore $|I(V)| \approx n$. But this makes an uniform box decomposition unsuitable for a reduction of high-dimensional cluster problems.

Figure 2.3 shows an example of an uniform box decomposition for our geometric cluster problem in R^2.

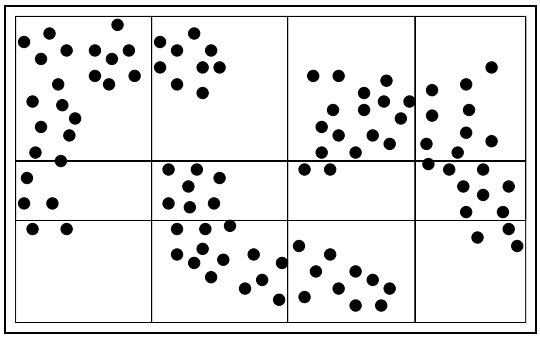

Figure 2.3: **Example: Uniform box decomposition in R^2.**

In chapter 3 we will present an adaptive method based on self-organized neural networks that allows to compute approximate box decompositions without the described shortages of an uniform procedure.

2.3 Decomposition based representative clustering

In section 1.2 we motivated the basic idea of a cluster problem reduction via representative clustering. We have presented a simple way to compute representatives $w_i \in \Omega$ with compressed frequency value $\check{f}(w_i)$. Further, we have shown that an optimal clustering of the representatives corresponds to an optimal clustering of the original data set V, if the homogeneity function h meets a local maximum and a global correspondence condition for all objects that are compressed to the same representative. Unfortunately this often leads to an unsatisfactory problem reduction, i.e. too many representatives are needed. The described conditions seems to be too strong for practical applications.

In this section we will develop a theory for cluster problem reduction via decomposition based representative clustering, without using any conditions for h. The objects are grouped together so that they are building partitions of a decomposition of the data set V. For the computation of an optimal k-cluster set of the representative set W, the original homogeneity function h is replaced by a compressed function $\check{h}$. We will show that if the decomposition is suitably fine, i.e. the decomposition error is small, this k-cluster set can be extended to an optimal k-cluster set of V with respect to f and h.

Definition 2.3.1 *Assume $n_k \in \mathbf{N}$ with $n_k \leq n$. Let $W := \{w_1, \ldots, w_{n_k}\} \subset V$ any subset of V and let Θ any decomposition of V with n_k partitions Θ_s.*
(i) We call W a codebook of Θ, if $w_s \in \Theta_s$ for $s = 1, \ldots, n_k$. We will refer to the data objects w_s as representatives or codebook vectors.
(ii) Let W any codebook of Θ, then we call the function $\check{f} : \Omega \longrightarrow \mathbf{R}_0^+$ with

$$\check{f}(w_s) := f(\Theta_s) \text{ for } s = 1, \ldots, n_k \text{ and } \check{f}(v) := 0 \text{ for } v \in \Omega \setminus W,$$

the compression of f on W. We set $\check{f}(M) := \sum_{w \in M} \check{f}(m)$ for any subset $M \subset \Omega$.
(iii) Let W any codebook of Θ, then we call the function $\check{h}_f : \Omega \longrightarrow [0, 1]$ with

$$\check{h}_f(w_s, w_p) := \frac{1}{\check{f}(w_s)\check{f}(w_p)} \sum_{v \in \Theta_s} \sum_{w \in \Theta_p} h(v, w) f(v) f(w) \text{ for } s, p = 1, \ldots, n_k$$

and $\check{h}_f(v, w) := 0$ for $v, w \in \Omega \setminus W$, the compression of h on W with respect to f.
(iv) For any k-cluster set $\mathcal{C} := \{C_1, \ldots, C_k\}$ of V, set $C_s(W) := C_s \cap W$. Then we call $\mathcal{C}(W) := \{C_1(W), \ldots, C_k(W)\}$ the compression of $\mathcal{C}$ on W.
(v) For any k-cluster set $\mathcal{C} := \{C_1, \ldots, C_k\}$ of a codebook W of Θ, we define $\hat{\mathcal{C}} := \{\hat{C}_1, \ldots, \hat{C}_k\}$ with $\hat{C}_s := \bigcup_{w_p \in C_s} \Theta_p$ and call $\hat{\mathcal{C}}$ the extension of $\mathcal{C}$ on V.

Lemma 2.3.2 *Assume $n_k \in \mathbf{N}$ with $k \leq n_k \leq n$ and let Θ be any decomposition of V with n_k partitions Θ_s and a codebook W. Then we have:*
(a) The compression $\check{f}$ is a frequency function for W and the compression $\check{h}_f$ is a homogeneity function for W.
(b) If $\mathcal{C}$ is a k-cluster set of W then the extension $\hat{\mathcal{C}}$ is a k-cluster set of V.

Proof: (a) and (b) follow directly from Definition 2.3.1. □

A decomposition is fine enough for a given k-cluster set, if each partition belongs to only one cluster:

Definition 2.3.3 *Let $\mathcal{C} := \{C_1, \ldots, C_k\}$ be any k-cluster set of V. Further assume $n_k \in \mathbf{N}$ with $k \leq n_k \leq n$ and let $\Theta := \{\Theta_1, \ldots, \Theta_{n_k}\}$ be any decomposition of V. We call Θ a covering of $\mathcal{C}$, if there exist non-void disjoint index subsets $I_1, \ldots, I_k$ with $\bigcup_{s=1}^{k} I_s = \{1, \ldots, n_k\}$ so that $C_s = \bigcup_{p \in I_s} \Theta_p$.*

Obviously $\Theta_V := \{\{v\} \,|\, v \in V\}$ and $\Theta_{\mathcal{C}} := \mathcal{C}$ are trivial coverings of $\mathcal{C}$. But there exists also non-trivial coverings if $\mathcal{C}$ meets a stronger version of the optimal cluster assumption (see section 1.2):

Lemma 2.3.4 *Let $\mathcal{C}$ be any k-cluster set of V and $\epsilon \in \mathbf{R}_0^+$ with $\epsilon < h_{max}(V)$. If we have $(v \in C \implies w \in C)$ for any cluster $C \in \mathcal{C}$ and all $v, w \in V$ with $h(v, w) \geq h_{max}(V) - \epsilon$, then any ϵ-decomposition Θ of (V, h) is a covering of $\mathcal{C}$.*

Proof: Let $n_k \in \mathbf{N}$ with $k \leq n_k \leq n$ and $\Theta := \{\Theta_1, \ldots, \Theta_{n_k}\}$ be any ϵ-decomposition of (V, h). For any cluster $C_s \in \mathcal{C}$ set $I_s := \{p \,|\, \Theta_p \cap C_s \neq \emptyset\}$. Then we have $\bigcup_{s=1}^{k} I_s = \{1, \ldots, n_k\}$ and $C_s \subset \bigcup_{p \in I_s} \Theta_p$. Obviously we are ready, if we show:

$$(\forall p \in I_s)\ \Theta_p \subset C_s.$$

But this follows directly: Since $p \in I_s$ there exists an object $v \in \Theta_p \cap C_s$. Then for all $w \in \Theta_p$ we have $h(v, w) \geq h_{max}(V) - \epsilon$ and therefore also $w \in C_s$. □

The next Lemma shows that the weighted intra-cluster homogeneity of any k-cluster set $\mathcal{C}$ of V and its compression on W are equal if there exists any covering of $\mathcal{C}$. We will use this fact in combination with Lemma 2.3.6 within the proof of the basic Theorem 2.3.7.

Lemma 2.3.5 *Let $\mathcal{C} := \{C_1, \ldots, C_k\}$ be any k-cluster set of V and Θ be any covering of $\mathcal{C}$ with n_k partitions Θ_p and a codebook $W := \{w_1, \ldots, w_{n_k}\}$. Then the compression $\mathcal{C}(W)$ is a k-cluster set of W with $\Gamma_{\check{f},\check{h}_f}(\mathcal{C}(W)) = \Gamma_{f,h}(\mathcal{C})$.*

Proof: Obviously $\mathcal{C}(W)$ is a k-cluster set, if $C_s(W) \neq \emptyset$ for $s = 1, \ldots, k$. But this follows immediately from the fact that Θ is a covering of $\mathcal{C}$ with codebook W. Further it follows that the index subsets $I_1, \ldots, I_k$ with $I_s := \{p \,|\, w_p \in C_s\}$ are non-void and disjoint and that we have $C_s = \bigcup_{p \in I_s} \Theta_p$.
Since $f(C_s) = \check{f}(C_s(W))$, this yields:

$$\begin{aligned}
\Gamma_{f,h}(\mathcal{C}) &= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{f(C_s)}\sum_{v\in C_s}\sum_{w\in C_s} h(v,w)f(v)f(w)\\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{f(C_s)}\sum_{p_1\in I_s}\sum_{p_2\in I_s}\sum_{v\in\Theta_{p_1}}\sum_{w\in\Theta_{p_2}} h(v,w)f(v)f(w)\\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{f(C_s)}\sum_{p_1\in I_s}\sum_{p_2\in I_s} \check{h}_f(w_{p_1},w_{p_2})\check{f}(w_{p_1})\check{f}(w_{p_2})\\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{\check{f}(C_s(W))}\sum_{p_1\in I_s}\sum_{p_2\in I_s} \check{h}_f(w_{p_1},w_{p_2})\check{f}(w_{p_1})\check{f}(w_{p_2})\\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{\check{f}(C_s(W))}\sum_{w_{p_1}\in C_s(W)}\sum_{w_{p_2}\in C_s(W)} \check{h}_f(w_{p_1},w_{p_2})\check{f}(w_{p_1})\check{f}(w_{p_2})\\
&= \Gamma_{\check{f},\check{h}_f}(\mathcal{C}(W))
\end{aligned}$$

□

The covering property of a decomposition can be transmitted to its extension:

Lemma 2.3.6 *Let Θ be any covering of $\widetilde{\mathcal{C}}$ with n_k partitions Θ_p and a codebook $W := \{w_1, \ldots, w_{n_k}\}$. If $\mathcal{C} := \{C_1, \ldots, C_k\}$ is a k-cluster set of W, then Θ is a covering of the extension $\hat{\mathcal{C}}$ of $\mathcal{C}$ on V.*

Proof: Set $J_s := \{p \,|\, w_p \in C_s\}$ for $s = 1, \ldots, k$. Since $\mathcal{C}$ is a k-cluster set of W, we have $J_s \neq \emptyset$, $J_s \cap J_p = \emptyset$ for $1 \leq s < p \leq k$ and $\bigcup_{s=1}^{k} J_s = \{1, \ldots, n_k\}$. By definition of $\hat{\mathcal{C}}$, we further have $\hat{C}_s := \bigcup_{w_p \in C_s} \Theta_p = \bigcup_{p \in J_s} \Theta_p$ and therefore Θ is a covering of $\hat{\mathcal{C}}$. □

Using the previous lemmata we can proof the basic theorem of decomposition based representative clustering:

Theorem 2.3.7 *Let $\widetilde{\mathcal{C}} := \{\widetilde{C}_1, \ldots, \widetilde{C}_k\}$ be any optimal k-cluster set of (V, f, h). Further let Θ be any covering of $\widetilde{\mathcal{C}}$ with n_k partitions Θ_p and a codebook W. If $\mathcal{C}$ is an optimal k-cluster set of $(W, \check{f}, \check{h}_f)$, then the extension $\hat{\mathcal{C}}$ is an optimal k-cluster set of (V, f, h).*

Proof: (i) Let $\widetilde{\mathcal{C}}(W) := \{\widetilde{C}_1(W), \ldots, \widetilde{C}_k(W)\}$ with $\widetilde{C}_s(W) := \widetilde{C}_s \cap W$ be the compression of $\widetilde{\mathcal{C}}$. Since Θ is an covering of $\widetilde{\mathcal{C}}$, we can apply Lemma 2.3.5 and yield:

$$\Gamma_{f,h}(\widetilde{\mathcal{C}}) \;=\; \Gamma_{\check{f},\check{h}_f}(\widetilde{\mathcal{C}}(W)).$$

(ii) Let $\hat{\mathcal{C}}(W) := \{\hat{C}_1(W), \ldots, \hat{C}_k(W)\}$ with $\hat{C}_s(W) := \hat{C}_s \cap W$ be the compression of $\hat{\mathcal{C}}$. Then one easily checks that $\hat{\mathcal{C}}(W) = \mathcal{C}$. Since Lemma 2.3.6 guarantees that Θ is a covering of $\hat{\mathcal{C}}$, we can again apply Lemma 2.3.5 and yield:

$$\Gamma_{f,h}(\hat{\mathcal{C}}) \;=\; \Gamma_{\check{f},\check{h}_f}(\mathcal{C}).$$

(iii) Since $\mathcal{C}$ is an optimal k-cluster set of $(W, \check{f}, \check{h}_f)$ and $\widetilde{\mathcal{C}}$ is an optimal k-cluster set of (V, f, h), we have

$$\Gamma_{\check{f},\check{h}_f}(\mathcal{C}) \;\geq\; \Gamma_{\check{f},\check{h}_f}(\widetilde{\mathcal{C}}(W)) \;\text{ and }\; \Gamma_{f,h}(\widetilde{\mathcal{C}}) \;\geq\; \Gamma_{f,h}(\hat{\mathcal{C}}).$$

Using $(i) - (iii)$ we get

$$0 \;\geq\; \Gamma_{f,h}(\hat{\mathcal{C}}) \;-\; \Gamma_{f,h}(\widetilde{\mathcal{C}}) \;=\; \Gamma_{f,h}(\hat{\mathcal{C}}) \;-\; \Gamma_{\check{f},\check{h}_f}(\widetilde{\mathcal{C}}(W)) \;\geq\; \Gamma_{f,h}(\hat{\mathcal{C}}) \;-\; \Gamma_{\check{f},\check{h}_f}(\mathcal{C}) \;=\; 0$$

and therefore $\Gamma_{f,h}(\hat{\mathcal{C}}) \;=\; \Gamma_{f,h}(\widetilde{\mathcal{C}})$. Since $\widetilde{\mathcal{C}}$ is an optimal k-cluster set, this guarantees that $\hat{\mathcal{C}}$ is also optimal. □

From Theorem 2.3.7 we can derive a basic algorithm for the reduction of cluster problems via representative clustering based on decomposition:

Basic reduction algorithm

Suppose we want to compute an optimal k-cluster set of a data set V with respect to a frequency function f and a homogeneity function h.
(1) To reduce the complexity of the cluster problem, we have to compute first a decomposition $\Theta := \{\Theta_1, \ldots, \Theta_{n_k}\}$ of V and a codebook W so that Θ is an covering of an optimal k-cluster set of (V, f, h).
(2) Next we compute an optimal *representative clustering*, i.e. an optimal k-cluster set $\mathcal{C}$ of $(W, \check{f}, \check{h}_f)$.
(3) Finally we have to extend $\mathcal{C}$ on V. The resulting $\hat{\mathcal{C}}$ is an optimal k-cluster set of (V, f, h).

Obviously such an algorithm makes only sense if in step (1) the optimal k-cluster set has not to be known a priori and the number n_k is much smaller than the number n of objects in V.

Using the optimal cluster assumption (see section 1.2) and Lemma 2.3.4, we can suppose that for sufficiently small ϵ, each ϵ-decomposition of V is a covering of each optimal k-cluster set of (V, f, h). This motivates the following assumption:

Covering assumption

If a decomposition Θ of V is sufficiently fine, i.e. if $\vartheta_{f,h}(\Theta)$ is small, then there exists a nearly optimal k-cluster set of (V, f, h) so that Θ is a covering of it.

Obviously the fineness of Θ corresponds with the number of partitions n_k. Therefore we need a method that — given an upper bound of n_k — tries to compute a maximally fine decomposition, while using only a minimal number of partitions. In chapter 3 we will present such a method based on KOHONEN'S Self-Organizing Maps (SOM). Since the choice of the upper bound for n_k is rather arbitrary, in chapter 4 we will refine our basic reduction algorithm to a multilevel algorithm that iterates the steps (1) and (2) until a sufficiently fine decomposition and corresponding optimal representative clustering is found.

Example: Representative clustering of a geometric cluster problem in R^2

We will give a short demonstration of our basic reduction algorithm by our example of a geometric cluster problem in R^2.

Since $h_{max}(V) = 1$, any ϵ-decomposition Θ with $\epsilon = 0.05$ should be fine enough to use it within our algorithm. Figure 2.4 shows a suitable ϵ-decomposition of the 100 points in the data set V with $n_k = 10$ partitions.

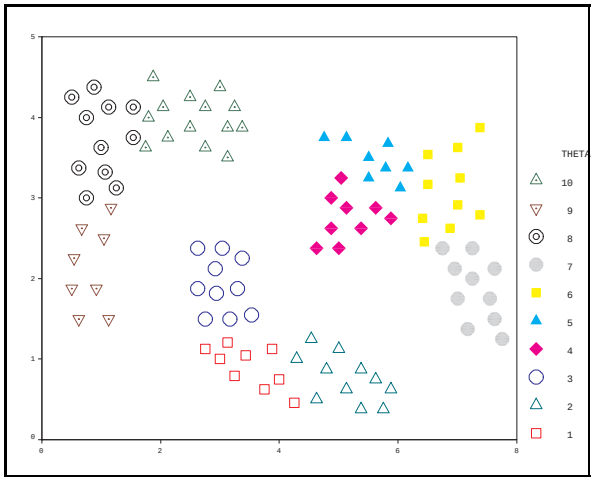

Figure 2.4: **Example: Covering with $n_k = 10$ partitions of 3-cluster set in R^2.**

Now we have to choose any codebook $W := \{w_1, \ldots, w_{10}\}$ of Θ and to compute the compressed functions $\check{f}$ and $\check{h}$ according to Definition 2.3.1.

One easily checks that $\{\{w_1, w_2, w_3\}, \{w_4, w_5, w_6, w_7\}, \{w_8, w_9, w_{10}\}\}$ is an optimal 3-cluster set of $(W, \check{f}, \check{h})$. An extension on V directly leads to the three clusters C_1, C_2 and C_3 (see Figure 1.1). Note that the 3-cluster set $\mathcal{C} := \{C_1, C_2, C_3\}$ meets the condition $(v \in C \implies w \in C)$ for any cluster $C \in \mathcal{C}$ and all $v, w \in V$ with $h(v, w) \geq h_{max}(V) - \epsilon$. Therefore Lemma 2.3.4 guarantees that our decomposition Θ is a covering of $\mathcal{C}$, i.e. that it was fine enough.

Decomposition clustering

Instead of clustering codebook vectors, we can also cluster a decomposition itself: Let $\Theta := \{\Theta_1, \ldots, \Theta_{n_k}\}$ be any decomposition of V. Then Θ can be interpreted as a data set in $\hat{\Omega} := \wp(\Omega)$, where $\wp(\Omega) := \{M \mid M \subset \Omega\}$ denotes the power set of Ω. We can extend the frequency function f and the homogeneity function h on subsets of Ω:

Definition 2.3.8
(a) We call $\hat{f} : \wp(\Omega) \longrightarrow \mathbf{N}$ with $\hat{f}(M) := \sum_{v \in M} f(v)$ for any subset $M \subset \Omega$, the set extension of f. We set $\hat{f}(\mathcal{M}) := \sum_{M \in \mathcal{M}} \hat{f}(M)$ for $\mathcal{M} \subset \wp(\Omega)$.
(b) We call $\hat{h}_f : \wp(\Omega) \times \wp(\Omega) \longrightarrow [0, 1]$, with

$$\hat{h}_f(V_1, V_2) := \begin{cases} \frac{1}{\hat{f}(V_1)\hat{f}(V_2)} \sum_{v \in V_1} \sum_{w \in V_2} h(v, w) f(v) f(w) & \text{if } V_1 \cap V, V_2 \cap V \neq \emptyset \\ 0 & \text{else} \end{cases}$$

for any subsets $V_1, V_2 \subset \Omega$, the set extension of h with respect to f.

Note that we have $0 \leq \hat{h}_f(V_1, V_2) \leq 1$ and $\hat{h}_f(V_1, V_2) = \hat{h}_f(V_2, V_1)$ for any non-void subsets $V_1, V_2 \subset V$.

The following Theorem guarantees that the computation of an optimal k-cluster set of $(W, \check{f}, \check{h})$ is equivalent to the computation of an optimal k-cluster set of $(\Theta, \hat{f}, \hat{h})$, if Θ is any decomposition of V with codebook W. This makes it possible to replace the clustering of codebook vectors by a direct clustering of the corresponding partitions of the decomposition within step (2) of the basic reduction algorithm.

Theorem 2.3.9 *Let $W := \{w_1, \ldots, w_{n_k}\}$ be any codebook of Θ.*
(i) Let $\mathcal{C} := \{C_1, \ldots, C_k\}$ be any k-cluster set of Θ. Then there exist k non-void disjoint index subsets I_s with $\bigcup_{s=1}^{k} I_s = \{1, \ldots, n_k\}$ so that $C_s = \{\Theta_p \,|\, p \in I_s\}$. If we set $\check{C}_s(W) := \{w_p \,|\, p \in I_s\}$, then $\check{\mathcal{C}}(W) := \{\check{C}_1(W), \ldots, \check{C}_k(W)\}$ is a k-cluster set of W with $\Gamma_{\hat{f},\hat{h}_f}(\mathcal{C}) = \Gamma_{\check{f},\check{h}_f}(\check{\mathcal{C}}(W))$.
(ii) Let $\mathcal{C} := \{C_1, \ldots, C_k\}$ be any k-cluster set of W. If we set $I_s := \{p \,|\, w_p \in C_s\}$, then the index subsets $I_1, \ldots, I_k$ are non-void and disjoint with $\bigcup_{s=1}^{k} I_s = \{1, \ldots, n_k\}$. The extension $\hat{\mathcal{C}}(\hat{\Omega}) := \{\hat{C}_1(\hat{\Omega}), \ldots, \hat{C}_k(\hat{\Omega})\}$ with $\hat{C}_s(\hat{\Omega}) := \{\Theta_p \,|\, p \in I_s\}$ is a k-cluster set of Θ with $\Gamma_{\check{f},\check{h}_f}(\mathcal{C}) = \Gamma_{\hat{f},\hat{h}_f}(\hat{\mathcal{C}}(\hat{\Omega}))$.

Proof: Since (ii) follows analogously, we only show (i):

$$
\begin{aligned}
\text{(a)}\ \Gamma_{\hat{f},\hat{h}_f}(\mathcal{C}) &= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{\hat{f}(C_s)}\sum_{V_1\in C_s}\sum_{V_2\in C_s}\hat{h}_f(V_1,V_2)\hat{f}(V_1)\hat{f}(V_2)\\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{\sum_{p\in I_s} f(\Theta_p)}\sum_{p_1\in I_s}\sum_{p_2\in I_s}\hat{h}_f(\Theta_{p_1},\Theta_{p_2})\hat{f}(\Theta_{p_1})\hat{f}(\Theta_{p_2})\\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{\sum_{p\in I_s}\check{f}(w_p)}\sum_{p_1\in I_s}\sum_{p_2\in I_s}\sum_{v\in\Theta_{p_1}}\sum_{w\in\Theta_{p_2}}h(v,w)f(v)f(w)\\
&= \frac{1}{k}\sum_{s=1}^{k}\frac{1}{\check{f}(\check{C}_s(W))}\sum_{p_1\in I_s}\sum_{p_2\in I_s}\check{h}_f(w_{p_1},w_{p_2})\check{f}(w_{p_1})\check{f}(w_{p_2})\\
&= \Gamma_{\check{f},\check{h}_f}(\check{\mathcal{C}}(W))
\end{aligned}
$$

□

We will use this equivalence of representative clustering and decomposition clustering in the discussion of our main Theorem 4.3.9 in chapter 4.

2.4 Efficient cluster description via approximate box decomposition

In this section we will describe, how approximate box decompositions can be used to generate efficient cluster descriptions according to section 1.3.

2.4.1 Computation of membership rules

We can easily determine cluster membership rules for a k-cluster set $\mathcal{C}$, if we have an approximate box decomposition of V that is a covering of $\mathcal{C}$:

Lemma 2.4.1 *Assume $n_k \in \mathbf{N}$ with $k \leq n_k \leq n$. Let $\mathcal{C} := \{C_1, \ldots, C_k\}$ be any k-cluster set of V and $\Theta := \{\Theta_1, \ldots, \Theta_{n_k}\}$ be any covering of $\mathcal{C}$ with non-void disjoint index subsets $I_1, \ldots, I_k$ so that $C_s = \bigcup_{p \in I_s} \Theta_p$. Further suppose the existence of any $\Delta := \{\Delta_1, \ldots, \Delta_{n_k}\}$ so that (Θ,Δ) is an approximate box decomposition of V with respect to f.*
(i) For $p \in \{1, \ldots, n_k\}$ there exist for each $j \in \{1, \ldots, q\}$ a subset $B_{p,j} \subset A_j$ so that $\Delta_p = \bigotimes_{j=1}^{q} B_{p,j}$.
(ii) Set $\mathcal{B}_p := \{B_{p,1}, \ldots, B_{p,q}\}$ for $p \in \{1, \ldots, n_k\}$ and define $r_{\mathcal{B}_p} : \Omega \longrightarrow \{0, 1\}$ with

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

If $p \in I_s$ and $f(\Delta_p \setminus C_s) = 0$, then $r_{\mathcal{B}_p}$ is a membership rule for cluster C_s.
(iii) If $f(\Delta_p \setminus C_s) = 0$ for all $p \in I_s$ and $C_s \subset \bigcup_{p \in I_s} \Delta_p$, then $r_s := \{r_{\mathcal{B}_p} \mid p \in I_s\}$ is a complete membership rule set of cluster C_s.

Proof: (i) Follows directly from $\Delta_p \in \mathrm{BOX}(\Omega)$.
(ii) We have

$$f(\Delta_p \setminus C_s) = 0 \iff \Delta_p \cap V \subset C_s$$

and therefore

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

(iii) From (ii) follows that $r_{\mathcal{B}_p}$ is a membership rule of C_s for each $p \in I_s$. Since $C_s \subset \bigcup_{p \in I_s} \Delta_p$, we have

$$v \in C_s \implies (\exists p \in I_s)\, v \in \Delta_p \iff (\exists p \in I_s)\, r_{\mathcal{B}_p}(v) = 1\,.$$

□

Note that the condition $f(\Delta_p \setminus C_s) = 0$ is only violated if boxes from different clusters overlap each other. Therefore this condition is weaker than the condition $\mathrm{overlap}_f(\Delta) = 0$.

Membership rule set algorithm

From Lemma 2.4.1 we can derive an algorithm to compute complete membership rule sets that are nearly minimal for a k-cluster set $\mathcal{C}$:
(1) Compute an approximate box decomposition (Θ, Δ) of V so that Θ is a covering of $\mathcal{C}$, Δ fits the conditions of Lemma 2.4.1 and $n_k \ll n$.
(2) Construct the n_k membership rules $r_{\mathcal{B}_p}$ as described in Lemma 2.4.1. Since for each cluster a minimally complete membership rule set must contain at least one rule, we need at least k membership rules to describe a k-cluster set $\mathcal{C}$. If the difference of n_k and k is not to large, the complete membership rule sets r_s are nearly minimal.

Example: Complete membership rule set for a 3-cluster set in R^2 based on approximate box decomposition.

For our geometrically based cluster problem in R^2 with $k = 3$, Figure 2.5 shows an approximate box decomposition (Ω, Δ) that covers the optimal 3-cluster set. Obviously the overlap between the boxes causes no problems and therefore we can use $\Delta := \{\Delta_1, \ldots, \Delta_{n_k}\}$, with boxes $\Delta_p = B_{p,1} \times B_{p,2}$ and subsets $B_{p,j} \subset \mathbf{R}$ according to Table 2.1, to determine minimal membership rule set for the optimal 3-cluster set $\{C_1, C_2, C_3\}$.

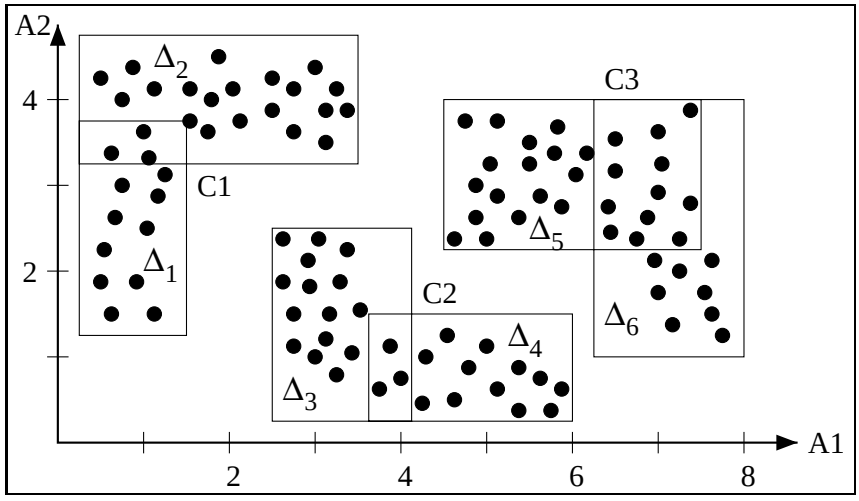

Figure 2.5: **Example: Approximate box decomposition that is a covering of a 3-cluster set in R^2.** Unproblematic overlap between boxes of the same cluster.

If we define the membership rules $r_{\mathcal{B}_p}$ as described in Lemma 2.4.1, then $r_1 := \{r_{\mathcal{B}_1}, r_{\mathcal{B}_2}\}$ (respectively $r_2 := \{r_{\mathcal{B}_3}, r_{\mathcal{B}_4}\}$, $r_3 := \{r_{\mathcal{B}_5}, r_{\mathcal{B}_6}\}$) is a complete membership rule set of cluster C_1 (respectively C_2, C_3). One easily checks that r_1, r_2 and r_3 are minimal.

p	$B_{p,1}$	$B_{p,2}$
1	[0.25, 1.5]	[1.25, 3.75]
2	[0.25, 3.5]	[3.25, 4.75]
3	[2.5, 4.125]	[0.25, 2.5]
4	[3.625, 6]	[0.25, 1.5]
5	[4.25, 7.5]	[2.25, 4]
6	[6.25, 8]	[1, 4]

Table 2.1: **Example: Approximate box decomposition that is a covering of a** 3**-cluster set in** R^2**.**

Instead of Δ we could also use the box decomposition that is shown on the right hand side of Figure 2.2. But note that the approximate box decomposition on the left hand side leads to an incomplete membership rule set for cluster C_3. The uniform box decomposition from Figure 2.3 is also suitable, but the corresponding membership rule sets are not minimal.

2.4.2 Discriminating attributes

Since we are interested in efficient cluster descriptions, we have not only to determine complete membership rule sets, we have also to reduce them as much as possible (see section 1.3). Therefore we have to identify the *discriminating attributes* of the cluster problem, i.e. the attributes that are necessary to determine the cluster membership of each data object.

Let $V = \{v_1, \ldots, v_n\} \subset \Omega$ be any data set in Ω with frequency function f and homogeneity function h. Further let $\mathcal{C} := \{C_1, \ldots, C_k\}$ be any k-cluster set of V and $\Theta := \{\Theta_1, \ldots, \Theta_{n_k}\}$ be any covering of $\mathcal{C}$ with non-void disjoint index subsets $I_1, \ldots, I_k$ so that $\bigcup_{s=1}^{k} I_s = \{1, \ldots, n_k\}$ and $C_s = \bigcup_{p \in I_s} \Theta_p$ for $s = 1, \ldots, k$. Remember that for any index subset $J \in \{1, \ldots, q\}$, $v(J)$ denotes the projection of $v \in \Omega$ on $\Omega(V)$, where $\Omega(V)$ is spanned by the attributes A_j with $j \in J$. Remember further that we have defined $M(J) := \{v(J) \,|\, v \in M\}$ for any subset $M \subset \Omega$.

Definition 2.4.2 *Let $J \subset \{1,\ldots,q\}$ be any non-void index subset and denote by $J^c := \{1,\ldots,q\} \setminus I$ its complement.*
(a) We call the attribute set $\mathcal{A}(J^c) := \{A_j \,|\, j \in J^c\}$ redundant for $\mathcal{C}$ if we have:

$$v \in C_s \iff v(J) \in \bigcup_{p \in I_s} \Theta_p(J) \ \text{ forall } v \in V.$$

(b) We call the attribute set $\mathcal{A}(J^c)$ maximally redundant for $\mathcal{C}$ if there exists no subset $\widetilde{J} \subset \{1,\ldots,q\}$ so that $\mathcal{A}(\widetilde{J}^c)$ is redundant for $\mathcal{C}$ and $|J| > |\widetilde{J}|$.
(c) We call attribute A_i an univariate discriminating attribute of $\mathcal{C}$, if $\mathcal{A}(\{j\})$ is not redundant for $\mathcal{C}$.
(d) We call the attributes $A_j \in \mathcal{A}(J)$ multivariate discriminating attributes of $\mathcal{C}$ if $\mathcal{A}(J^c)$ is maximally redundant for $\mathcal{C}$.

The following Lemma is an extension of Lemma 2.4.1:

Lemma 2.4.3 *Suppose there exist any $\Delta := \{\Delta_1,\ldots,\Delta_{n_k}\}$ so that (Θ,Δ) is an approximate box decomposition of V with respect to f. Choose any $s \in \{1,\ldots,k\}$ and any $p \in I_s$. Define $r_{\mathcal{B}_p}$ according to Lemma (2.4.1) and suppose further that $f(\Delta_p \setminus C_s) = 0$, then we have:*
The function $r_{\mathcal{B}_p(J)}$ with $\mathcal{B}_p(J) := \{B_{p,1}(J),\ldots,B_{p,q}(J)\}$ and

$$B_{p,j}(J) := \begin{cases} B_j & \text{if } j \in J \\ A_j & \text{else} \end{cases}, \ \text{for } j \in \{1,\ldots,q\},$$

is a J-reduced membership rule for cluster C_s if $\mathcal{A}(J^c)$ is redundant for $\mathcal{C}$.

Proof: We have

$$f(\Delta_p \setminus C_s) = 0 \iff \Delta_p \cap V \subset C_s = \bigcup_{p \in I_s} \Theta_p \iff \Delta_p(J) \subset \bigcup_{p \in I_s} \Theta_p(J)$$

and therefore

$$r_{\mathcal{B}_p(V)}(v) = 1 \implies v(J) \in \Delta_p(J) \subset \bigcup_{p \in I_s} \Theta_p(J) \iff v \in C_s.$$

□

Analogously to Lemma 2.4.1 one easily checks that $\{r_{\mathcal{B}_p(J)} \,|\, p \in I_s\}$ is a J-reduced complete membership rule set of cluster C_s, if $f(\Delta_p \setminus C_s) = 0$ for all $p \in I_s$ and $C_s \subset \bigcup_{p \in I_s} \Delta_p$. Moreover if $\mathcal{A}(J^c)$ is maximally redundant, $\{r_{\mathcal{B}_p(J)} \,|\, p \in I_s\}$ is optimally reduced.

Discriminating attributes identification algorithm

Suppose that $\mathcal{C}$ is any optimal k-cluster set of (V, f, h) and that there exist any $\Delta := \{\Delta_1, \ldots, \Delta_{n_k}\}$ so that (Θ,Δ) is an approximate box decomposition of V with respect to f. Then the following algorithm can be used to determine the multivariate discriminating attributes of $\mathcal{C}$:
(1) Choose $0 < \delta \ll 1$. Set $J_{opt} := \{1, \ldots, q\}$ and $\delta_{opt} := 0$.
(2) Let $J \subset \{1, \ldots, q\}$ be any index subset of minimal size so that

$$\text{overlap}_f(\hat{\Delta}(J) \leq \text{overlap}_f(\Delta) + \delta,$$

where $\hat{\Delta}(J) := \{\hat{\Delta}_1(J), \ldots, \hat{\Delta}_{n_k}(J)\}$ with

$$\hat{\Delta}_p(J) \subset \Omega \text{ and } v \in \hat{\Delta}_p(J) \iff v(J) \in \Delta_p(J) \text{ for all } v \in \Omega.$$

(3) If $|J| < q$, then goto step (5).
(4) If $\delta_{opt} = 0$, then goto step (7), else stop.
(5) If $\mathcal{A}(J^C)$ is not redundant for $\mathcal{C}$, then decrease δ and goto step (2).
(6) If $|J| < |J_{opt}|$, then set $J_{opt} := J$ and $\delta_{opt} := \delta$, else stop.
(7) If $|J| > 1$, then increase δ and goto step (2), else stop.

For cluster problems with a special type of homogeneity function, that exhibits a stochastic property, in chapter 4 we are going to present a method that allows to proof quickly if $\mathcal{A}(J^C)$ is redundant for $\mathcal{C}$.

Example: Discriminating attributes of cluster problem with unknown number of clusters

If we look again at our simple example from section 1.4, we can easily identify the discriminating attributes corresponding to the optimal k-cluster sets for differently chosen k.

Obviously for the clusterings $\mathcal{C}(1) - \mathcal{C}(4)$ we need for each $v \in V$ only the value for attribute A_1 to determine the cluster membership.

Formally spoken, if we set $J := 1$ and choose $k \in \{1, \ldots, 4\}$, then we have for each cluster $C \in \mathcal{C}(k)$ and for all $v \in V$:

$$v \in C_s \iff v(J) \in C_s(J).$$

Since $\Theta := \mathcal{C}(k)$ is always a trivial covering of k-cluster set $\mathcal{C}(k)$, the attribute set $\mathcal{A}(J^c) = \{A_2\}$ is redundant. Further it is maximally redundant, because it is not possible that a redundant attribute set contains all attributes. Therefore $A_1 \in \mathcal{A}(J)$ is a multivariate discriminating attribute of $\mathcal{C}(k)$, $k = 1, \ldots, 4$.

To illustrate the working of the suggested identification algorithm, we use it to determine the discriminating attributes of $\mathcal{C} := \mathcal{C}(2)$:

- At the beginning we set $\Theta := \mathcal{C}$ and $\Delta := \{\Delta_1, \Delta_2\}$, with boxes $\Delta_1 := B_{1,1} \times B_{1,2} := [0.5, 2] \times [0.5, 3]$ and $\Delta_2 := B_{2,1} \times B_{2,2} := [5.5, 6] \times [1.5, 2.5]$. Then (Θ, Δ) is an approximate box decomposition of V.

- In step (1) we choose a small δ, e.g., $\delta := 0.01$. We set $J_{opt} := \{1, \ldots, q\}$ and $\delta_{opt} := 0$.

- Obviously in step (2) it is enough to investigate $J_1 := \{1\}$ and $J_2 := \{2\}$. Extending the projections $\Delta_s(J_1) := B_{s,1}$ and $\Delta_s(J_2) := B_{s,2}$ we got $\hat{\Delta}_s(J_1) := B_{s,1} \times \mathbf{R}$ and $\hat{\Delta}_s(J_2) := \mathbf{R} \times B_{s,2}$ for $s = 1, 2$. This leads to $\mathrm{overlap}_f(\hat{\Delta}(J)) := 0$ and $\mathrm{overlap}_f(\hat{\Delta}(J_2)) := 0.56$. Since we have $\mathrm{overlap}_f(\Delta) = 0$, we set $J := J_1$.

- At step (3) we have $|J| = 1 < 2 = q$ and therefore we jump to step (5).

- Now we have to prove, if $\mathcal{A}(J^C) = \{A_2\}$ is redundant. This is the case and we go to step (6).

- Since $|J| = 1 < 2 = |J_{opt}|$, we set $J_{opt} := J$ and $\delta_{opt} := \delta$.

- At step (7) we stop, because $|J| = 1$. The result of the algorithm is $J_{opt} := 1$ and determines A_1 as the only multivariate discriminating attribute of $\mathcal{C}$. One easily checks, that δ_{opt} is a kind of quality indicator of the computation. If δ_{opt} is sufficiently small, we can be confident that we have identified the correct multivariate discriminating attributes of clustering $\mathcal{C}$.