3 Approaches to Cluster Analysis

Many data mining methods rely on some concept of the similarity between pieces of information encoded in the data of interest. Various names have been applied to these clustering methods, depending largely on the field of application in data science. For example, in biology the term "numerical taxonomy" is used [Thorel et al., 1990], in psychology the term Q analysis is sometimes employed, market researchers often talk about "segmentation" [Arimond/Elfessi, 2001] and in the artificial intelligence literature, unsupervised pattern recognition is the favored label [Everitt et al., 2001, p. 4]. The corresponding methods can be either data-driven or need-driven. The latter, called also constraint clustering [Tung et al., 2001] aims at organizing the true structure to meet certain application requirements such as energy aware sensor networks, privacy preservation, and market segmentation [Ge et al., 2007, p. 320]. An overview of constrained clustering algorithms can be found in [Basu et al., 2008].

Here, however, the focus is placed on data-driven[10] methods, in which patterns present in the data are used to identify homogeneous groups of objects [Arabie et al., 1996, p. 8 ff.]. Consequently, the term *cluster analysis* is used to refer to a step in the knowledge discovery process (chapter 2, Figure 2.5.). Let it be assumed that in Figure 3.1 (top left), the first data set (I) contains two variables[11]. The division of this homogeneous data set into different patterns would be called dissection [Everitt et al., 2001, p. 7]. By contrast, *natural clusters* do not require dissection; instead, they are clearly separated in the data [Duda et al., 2001, p. 539; Theodoridis/Koutroumbas, 2009, pp. 579, 600], as shown in the second data set (II) in Figure 3.1 (top right).

No generally accepted definition of clusters exists in the literature [Hennig et al., 2015, p. 705]. Additionally, Kleinberg showed for a set of three simple properties (scale-invariance, consistency and richness), that there is no clustering function[12] satisfying all three [Kleinberg, 2003]. By concentrating on distance and density based *structures*[13], this work restricts clusters to "natural" clusters (see section 2) and therefore omits the axiom of richness where all partitions should be achievable. Consequently, only natural clusters, in which objects are similar within clusters and dissimilar between clusters [Bouveyron et al., 2012], are considered here. For example, the distance distribution in the input space can be bimodal, indicating a distinction between the inter- versus intracluster distances: in data set I in Figure 3.1 (bottom left), no large intercluster distances exist and the distribution of the distances is unimodal, whereas in data set II in Figure 3.1 (bottom right), the distribution of the distances is bimodal because data set II contains two natural clusters with a large intercluster distance. Another example is the case in which the number of data points in one *elementary volume* ($d\vec{v}$) of the input space is higher than that in another elementary volume $d\vec{v}$, which can be estimated using a nonparametric technique for density estimation (e.g., kernel density estimation). In a third example, local proximities can be defined as structures based on neighborhoods $H_j(k, \Gamma, M)$ (see chapter 2.2.1).

[10] The progress in an "algorithmic activity" is enforced by data w.r.t. patterns (as opposite to intuition or personal experience, e.g. through the setting of parameters).

[11] In fact, this figure shows a CCA projection of the leukemia data set (see chapter 9).

[12] "[A]ny function f that takes a set S of n points with pairwise distances between them, and returns a partition of S" [Kleinberg, 2003, p 2].

[13] They can be described as patterns identified based on discontinuity.

M. C. Thrun, *Projection-Based Clustering through Self-Organization and Swarm Intelligence*, https://doi.org/10.1007/978-3-658-20540-9_3

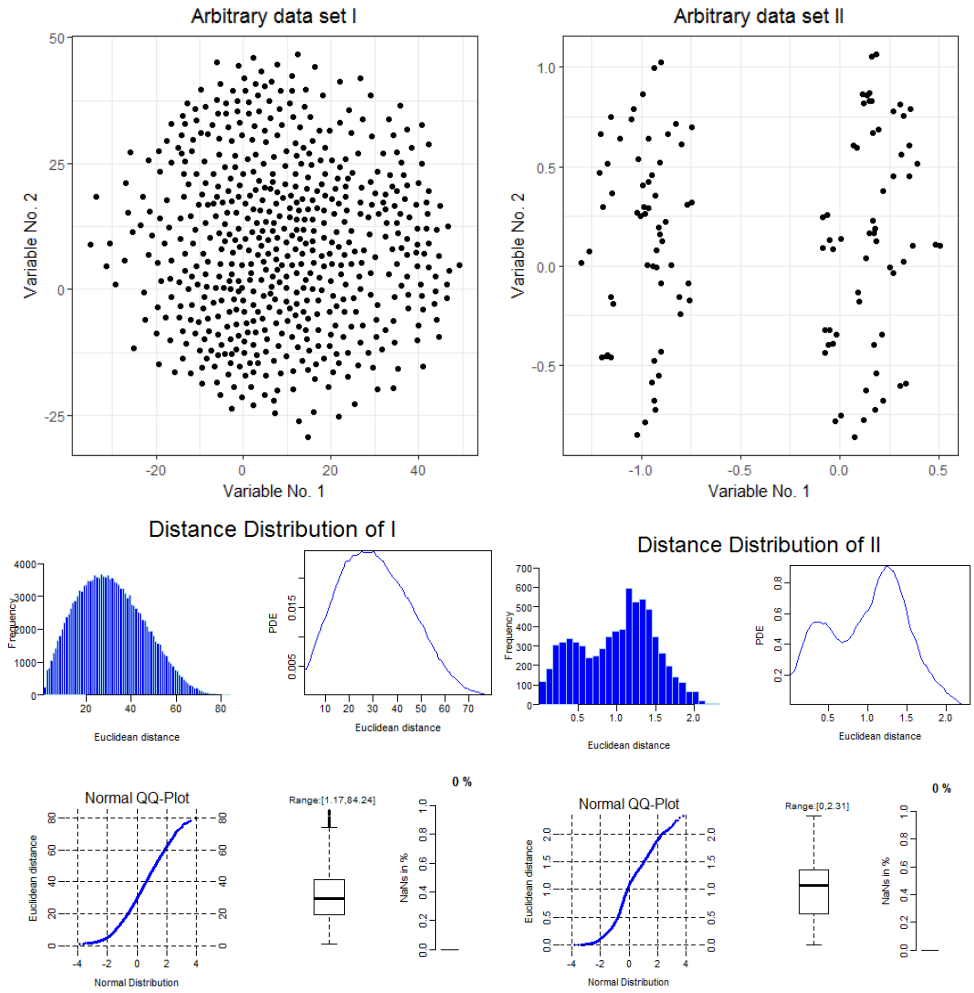

Figure 3.1: Data set I is an approximately homogeneous data set with patterns that form no natural clusters (left, top). The distance distribution in this case is not bimodal (left, bottom). Data set II contains two natural clusters with a large intercluster distance (right, top). The distance distribution is bimodal here (right, bottom). See Figure 12.2 or supplement B for a high-dimensional example. Distance distributions was generated using the AdaptGauss CRAN package [Thrun/Ultsch, 2015; Ultsch et al., 2015].

3.1 Common Clustering Methods

Clustering methods can be broadly divided into two groups: hierarchical and partitional methods [Jain, 2010]. Partitional clustering methods simultaneously divide a set of data points into subsets. Because we are concentrating on *natural clusters,* overlapping clustering is not considered here. It should be remarked that the choice of the clustering algorithm to be used is more important than the choice of the distance calculation [Jain/Dubes, 1988, p. 140].

A prominent example of a partitional clustering method is the well-known *k-means* method of [MacQueen, 1967] (originally from [Steinhaus, 1956]). It proceeds as follows: Once the number

of clusters has been chosen, a random initialization of cluster centers, called centroids, is performed in the input space. Then, the nearest data points to each centroid are assigned to that centroid. After the mapping of the data points, the centroids are moved such that the distances from the assigned points to their corresponding centroids are minimized. This process is performed repeatedly. Figure 3.2 illustrates four iterations of the process. In summary, k-means centroids are average points rather than individual data points. Details about the algorithm can be found in [Hennig et al., 2015, p. 68ff].

By contrast, the clustering method called partitioning around medoids (PAM), introduced in [L. Kaufman/Rousseeuw, 1990], minimizes the sum of the distances from the data points within a cluster to one chosen data point in the same cluster, called the medoid [Mirkin, 2005, p. 181]. In other words, the average distance between a medoid and a subset of data points in the same cluster is minimized. Aside from the change from centroids to medoids, the algorithm can be formulated analogously to k-means [Mirkin, 2005, p. 182].

Hierarchical clustering algorithms are based on the "representation of data as a hierarchy of clusters nested over set-theoretic inclusion" [Mirkin, 2005, p. 112]. In the agglomerative approach, such an algorithm begins with each data point in its own cluster and successively merges the most similar pairs of clusters to form a cluster hierarchy[14].

A typical visual representation of this process is called a dendrogram (Figure 3.3). A dendrogram is a tree showing a hierarchical structure of distance-based connections between subsets of points. The similarity between points or groups of points depends on the algorithm. [Bock, 1974] demonstrated (see chapter 2 for details) that for every dendrogram, an ultrametric space can be constructed in which the triangle inequality is redefined as

$D(l,j) \leq max\,(D(l,m), D(m,j))$.

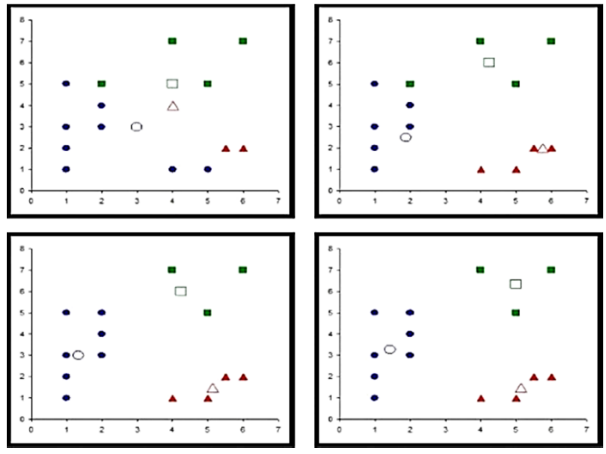

Figure 3.2: Steps of iteration using the k-means algorithm. After a random initialization of three centroids the nearest data points are assigned to each centroid. Then the centroids are moved to minimalize the distances.

[14] The divisive approach is not considered here (see [Mirkin, 2005, p. 113 ff] for details).

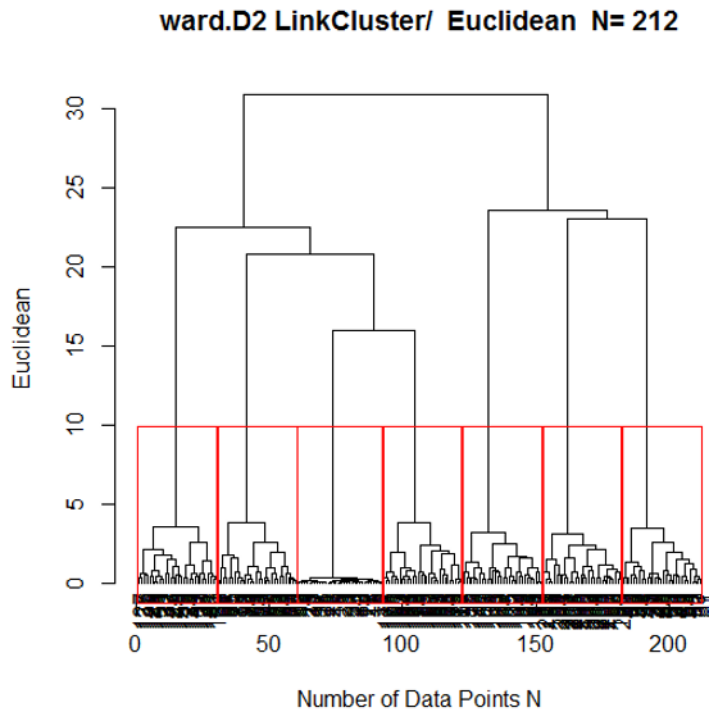

Figure 3.3: Dendrogram of the Hepta data set based on the Ward algorithm. Large changes in fusion levels of the ultrametric portion of the Euclidean distance in the Ward algorithm (y-axis) indicate the best cut. Seven clusters are indicated by red boxes at the y-axis value of 10. If only small changes in the fusion levels exist, it indicates that the algorithm is not able to find a cluster structure.

One of the most common hierarchical clustering algorithms is called *single linkage* (SL) [Florek et al., 1951; Sokal/Sneath, 1963], in which the clustering process is agglomerative [Jain et al., 1999]. In SL, the similarity between two subsets of data points is defined as the minimum distance between data points in these subsets [Duda et al., 2001, p. 553].

Let $\widetilde{D}$ be the distance between two clusters $c_1 \subset I$ and $c_2 \subset I$, and let $D(l,j)$ be the distance between two data points in the input space I; then, SL is defined based on (see [Hennig et al., 2015, p. 9]) $\widetilde{D}(c_1, c_2) = \min_{l \in c_1, j \in c_2} D(l,j)$.

In graph theory terminology, this process generates a tree [Duda et al., 2001, p. 553]. If it is allowed to continue until all subsets of points are linked, the result is a (minimal) spanning tree (MST) [Duda et al., 2001, pp. 553, 554; Jain/Dubes, 1988, p. 70]. Of all common algorithms developed before 1968, only SL satisfies all conditions of a "theoretically valid" clustering (see [Jardine/Sibson, 1968] for details).

Another hierarchical clustering algorithm that will be used here is called the *Ward* algorithm [Ward Jr, 1963]. In the Ward algorithm, the similarity between two subsets of points is based on an optimal value of an objective function, which commonly is the sum of squared errors (*SE)*.

Let $c_r \subset I$ and $c_q \subset I$ be two clusters such that $r, q \in \{1, \dots, k\}$ and $c_r \cap c_q = \{\}$ for $r \neq q$, and let the data points in the clusters be denoted by $j_i \in c_q$ and $l_i \in c_r$, with the cardinality of the sets being $k = |c_q|$ and $p = |c_r|$ and with

$$m(c_q) = \frac{1}{k}\sum_{i=1}^{k} j_i \text{ and } m(c_r) = \frac{1}{p}\sum_{i=1}^{p} l_i;$$

then, the *SE* is defined as (see [Theodoridis/Koutroumbas, 2009, pp. 661-663])

$$SE = \frac{k * p}{k + p} \sum_{i=1}^{n} \left(m(c_k) - m(c_p)\right)^2$$

In Figure 3.3, the ultrametric property of the Ward algorithm is represented in a dendrogram (for further details, see [Duda et al., 2001, p. 557; Everitt et al., 2001, p. 68ff; Jain/Dubes, 1988]). If the values on the y axis "for the levels are roughly evenly distributed throughout the range of possible values, then there is no principled argument that any particular number of clusters is better or more natural than another" [Duda et al., 2001, p. 551]. "Large changes in fusion levels are taken to indicate the best cut" [Everitt et al., 2001, p. 76]. The cut depicted in Figure 3.3 generates a clustering consisting of seven clusters of roughly equal size.
The next clustering method used in this work is called spectral clustering.

> *"[It] is a class of graph-based techniques that unravel the structure properties of a graph using information conveyed by the spectral decomposition [eigendecomposition [see [Goodfellow et al., 2016, pp. 42-44]]] of an associated [Laplacian] matrix. The elements of this matrix code the underlying similarities among nodes [data points] of the graph" [Theodoridis/Koutroumbas, 2009, p. 772].*

> *"The K principal eigenvectors of the Laplacian matrix provide a mapping of the objects into K dimensions. To obtain clusters, the resulting K-dimensional vectors are clustered by standard methods, usually K-means. There are various interpretations of this. […]. For these [Euclidean] data, spectral clustering acts as a remarkably robust linkage method." [Hennig et al., 2015, p. 10].*

There is a close resemblance between spectral clustering and manifold learning methods [Theodoridis/Koutroumbas, 2009, p. 779]. Here, the clustering algorithm of [Ng et al., 2002] is used to take advantage of the open-source implementation of this method that is available in the R language [R Development Core Team, 2008].
"Clustering via mixtures of parametric probability models is sometimes in the literature referred to as 'model-based clustering'" [Hennig et al., 2015, p. 10]. With the clustering algorithm of [Fraley/Raftery, 2006] in mind, here, this clustering method is called the *mixture of Gaussians* (MoG) method. The MoG method uses the *expectation maximization* (EM) algorithm (for further details on the EM algorithm, see [Bishop, 2006]).

> *The EM algorithm is "an algorithm of alternating maximization applied to the likelihood function for a mixture of distributions model. At each iteration, EM is performed according to the following steps: (1) Expectation: Given parameters of the mixture P_k and individual density functions a_k, find posterior probabilities for observations to belong to individual clusters g_{ik} […]. (2) Maximization: given posterior probabilities g_{ik}, find parameters P_k, a_k maximizing the likelihood function" [Mirkin, 2005, p. 178].*

The MoG method suffers "from the well-known curse of dimensionality [Bellman, 1957], which is mainly due to the fact that model-based clustering methods are over-parametrized in high-dimensional spaces" [Bouveyron/Brunet-Saumard, 2014, p. 53]. To solve this problem, "for model based clustering, variable selection can be tackled within a Bayesian framework" [Bouveyron et al., 2012]. In the case of the MoG clustering method, the optimal model can be

calculated according to the Bayesian information criterion [Aho et al., 2014] for parameterized Gaussian mixtures that are EM initialized using hierarchical agglomeration [Fraley/Raftery, 2002, pp. 10-12].

> *"In each hierarchical agglomeration, each stage of merging corresponds to a unique number of clusters, and a unique partition of data. A given partition can be transformed into indicator variables […] which can then be used as conditional probabilities in an M-step of EM for parameter estimation, initializing an EM iteration" [Fraley/Raftery, 2002, p. 11]. Here, the R package mclust is used [Fraley/Raftery, 2006].*

3.2 Structure of Natural Clusters

> *"Clusters can be of arbitrary shapes (structures) and sizes in a multidimensional pattern space. Each clustering criterion imposes a certain structure on the data, and if the data happen to conform to the requirements of a particular criterion, the true clusters are recovered. Only a small number of independent clustering criteria can be understood both mathematically and intuitively. Thus the hundreds of criterion functions proposed in the literature are related and the same criterion appears in several disguises"* [Jain/Dubes, 1988, p. 91].

This section analyzes common clustering algorithms from the perspective of structures, whereas in various other sources, the clustering criterion or objective function has been understood only intuitively. Here, it is argued that the main argument of Jain and Dubes has received overall consent from the clustering community: Different clustering methods tend to implicitly assume different structures of clusters [Duda et al., 2001, pp. 537, 542, 551; Everitt et al., 2001, pp. 61, 177; Handl et al., 2005; Theodoridis/Koutroumbas, 2009, pp. 862, 896; Ultsch/Lötsch, 2016].

3.2.1 Types of Structures Sought by Clustering Algorithms

The argument of Handl et al. is partially adopted here, in which natural clusters are considered to exhibit two types of structures, called compact and connected structures [Handl et al., 2005], as depicted in Figure 3.4. Clusters with compact structures show small variations in their intracluster distances; connected structures are based on the idea of neighborhoods of data points [Handl et al., 2005]. Here, a compact structure is considered to be mainly defined by inter- versus intracluster distances, whereas a connected structure is primarily defined by neighborhoods H of data. Using the definitions presented in section 2.2.1, neighborhoods can be identified based on graph theory. This can result in connected structures consisting of either unidirectional or direction-based neighborhoods.

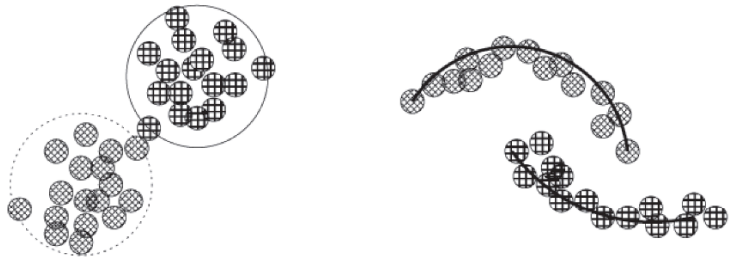

Figure 3.4: Two types of cluster structures, compact (left) and connected (right), taken from [Handl et al., 2005]. Here, a compact structure is considered to be mainly defined by intra- versus intercluster distances, whereas a connected structure is primarily defined based on neighborhoods $H_j\ (k, \Gamma, M)$ and the density of the data.

An example of an algorithm that seeks compact clusters is the k-means clustering algorithm, which imposes a spherical cluster structure [Duda et al., 2001, p. 542; Handl et al., 2005, p. 3202; Hennig et al., 2015, p. 61; Mirkin, 2005, p. 108; Theodoridis/Koutroumbas, 2009, p. 742] such that the clusters cannot be too elongated [L. R. Kaufman/Rousseeuw, 2005, p. 117]. This cluster structure can be found in a data set if "the data points are actually normally distributed" (…) because "the sample mean tends to fall in the region where the samples are most densely concentrated" [Duda et al., 2001, p. 537]. The k-means algorithm is sensitive to noise and outliers [Theodoridis/Koutroumbas, 2009, p. 744]. "This drawback […] gave rise to the k-medoids algorithms […]." The PAM algorithm is less sensitive to outliers. Because of its strong similarity to the k-means algorithm, it is assumed here that PAM also yields a compact spherical cluster structure.

Examples of algorithms that seek connected clusters include density-based methods such as DBscan [Ester et al., 1996] and SL [Handl et al., 2005]. Because SL searches for nearest neighbors [Cormack, 1971, p. 331], it tends to produce connected and chain-like structures [Duda et al., 2001, p. 554; Everitt et al., 2001, p. 67; Hartigan, 1981; Jain/Dubes, 1988, pp. 64-65; Theodoridis/Koutroumbas, 2009, p. 660]. A nearest neighbor is also a Delaunay neighbor (Figure 3.4), leading to a direction-based connected structure of clusters. Spectral clustering is based on graph theory and consequently searches for connected structures [Ng et al., 2002, p. 5] of clusters with "chain-like or other intricate structures" [Duda et al., 2001, p. 582]. This indicates that such an algorithm also searches for direction-based connected clusters (see also [Hennig et al., 2015, p. 10]). "They [spectral clustering methods] are well-suited for the detection of arbitrarily shaped clusters, but can lack robustness when there is little spatial separation between the clusters" [Handl et al., 2005, p. 3202].

The Ward algorithm is sensitive to outliers and tends to find compact clusters of equal size [Everitt et al., 2001, p. 61, Tab. 1] that are ellipsoidal in structure [Ultsch/Lötsch, 2016]. The MoG method uses a mixture-of-distributions approach, which leads to connected clusters. Contrary to [Handl et al., 2005], it is argued here that the MoG method should be able to separate clusters that are non-linear separable (e.g., Chainlink [Ultsch/Vetter, 1995]). Jains and Dubes report that "fitting a mixture density model to patterns" creates clusters with hyper-ellipsoidal shapes [Jain/Dubes, 1988, p. 92]. [Handl et al.] report that the MoG method is very effective for well-separated clusters [Handl et al., 2005, p. 3202].

In the case of self-organizing mapping (SOM)[15], the structures have been reported to be of "very general shapes" [Duda et al., 2001, p. 582; Ultsch/Lötsch, 2016]. Similarly to the emergent SOM (ESOM)/U-matrix clustering method [Ultsch et al., 2016a], the Databionic swarm (DBS) method that is discussed later in this work also uses the concept of emergence[16], through which novel properties can arise in a system. Emergence leads to clusters whose structures are not predefined.

To summarize, the cluster structures that are theoretically sought by various methods are visualized in Figure 3.5. It should be noted that clustering methods that search for clusters with connected structures should also be able to find compact clusters as long as the distance between

[15] However, for k-means-SOM of the batch type, spherical or well-separated structures have been reported [Handl et al., 2005, p. 3202] (see the SOM section in chapter 4 for the differences between ESOM and k-means-SOM).

[16] Definition, see chapter 7.3, p. 81-82

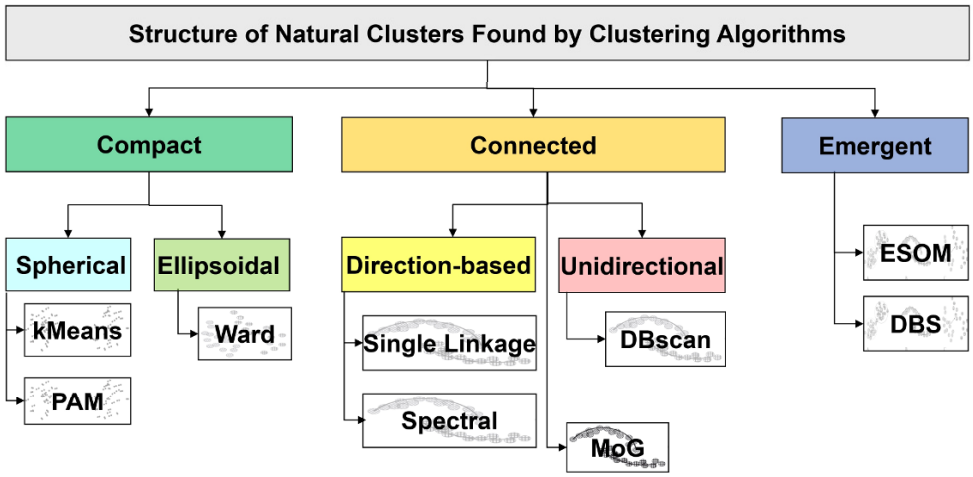

Figure 3.5: Overview of the cluster structures that common clustering algorithms tend to find. It is based on the literature, except for the MoG algorithm[17], for which an educated guess is made. The subgroup of DBscan clustering is characterized based on arguments presented in section 3.2.1, for the definition of emergent see chapter 7.3.

clusters is large or the density between clusters is very low (see also [Handl et al., 2005, p. 3202]); e.g., "single-linkage clusters detect high-density clusters if there is a low enough valley separating them" [Hartigan, 1981]. However, methods that search for compact and spherical structures cannot be expected to find connected structures.

3.2.2 *Quality of Clustering*

> *"[The quality of clustering is measured using a] "procedure for validating a cluster structure […]. This can be based on an internal index, an external index or resampling. An internal index scores the degree of correspondence between the data and the cluster structure. An external index compares the cluster structure with a structure given externally. A resampling is used to see whether the cluster structure is stable with respect to data change" [Mirkin, 2005, p. 205] (see also [Jain/Dubes, 1988, p. 161ff]).*

Internal and external indices are also often called *intrinsic* or *extrinsic* indices, respectively; here, they are referred to as *supervised* or *unsupervised* indices, respectively.
The simplest example of a supervised index is the accuracy, which is defined as follows:

$$\text{Accuracy } [\%] = \frac{[\text{No. of true positives}]}{[\text{No. of cases}]} \qquad (3.1)$$

In Eq. 3.1, the number of true positives is the number of labeled data points for which the label defined by a prior classification is identical to the label defined after the clustering process.
To determine either the number of clusters or the clustering quality, two approaches are generally possible. Covariance matrices can be calculated, or the intra- versus intercluster distances can be compared to evaluate the homogeneity versus heterogeneity of the clusters. In the literature, a sufficient overview of 15-30 indices has already been provided [Charrad et al., 2012; Dimitriadou et al., 2002], and these indices will not be further discussed here. A special type of unsupervised indices, referred to as quality measures for projection methods, will be separately

[17] Also known as model-based clustering.

introduced in chapter 6. Two unsupervised indices and corresponding visualizations are presented in the following sections.

3.2.2.1 Heatmaps

A heatmap is an example of an unsupervised index. For the ordering of the data points in heatmaps, dendrograms are often used. They enable the visualization of high-dimensional information and dissimilarity matrices without projecting them into a lower-dimensional space. Their use strongly depends on the sequence of the observations. For cluster validation, it is desirable to plot observations that are in the same cluster together [Hennig et al., 2015].

> *"[A heatmap] consists of a rectangular tiling, with each tile shaded on a color scale to represent the value of the corresponding element of the data set. The rows (columns) of the tiling are ordered such that similar rows (columns) [in the sense that they are in the same cluster] are near each other" [Wilkinson/Friendly, 2012]. "The cluster heat map is a rectangular tiling of a data matrix with cluster trees appended to its margins. Within a relatively compact display area, it facilitates inspection of joint cluster structure" [Wilkinson/Friendly, 2009].*

Unlike in [Wilkinson/Friendly, 2009; Fig. 1], in Figure 3.7, the dendrogram between the variables is disregarded and only the nxn heat map of the distance matrix is shown.

3.2.2.2 Silhouette plots

The Silhouette plot is a common unsupervised index for visual evaluation of a clustering [L. R. Kaufman/Rousseeuw, 2005].

> *"A score function s: $X \rightarrow [-1, 1]$ evaluates the positioning of data objects inside their assigned cluster. Let a(x) denote the average distance between x and all other objects of the same cluster, and b(x) denotes the smallest average distance between x and all objects of another cluster. The silhouette score follows as $(x) = \frac{b(x)-a(x)}{max\{a(x),b(x)\}}$. Silhouette scores similar to 1 indicate objects that have been assigned to an appropriate cluster, whereas −1 indicates objects that have been badly classified. Silhouette scores similar to 0 indicate objects that lie in between clusters. Each cluster is represented by one silhouette, showing which objects lie within the cluster and which objects merely hold an intermediate position. The entire clustering is displayed by plotting all silhouettes into a single diagram, from which the quality of the clusters can be compared" [Herrmann, 2011, pp. 91-92].*

A reasonable clustering is characterized by a silhouette width of greater than 0.5, and an average width below 0.2 should be interpreted as indicating a lack of any substantial cluster structure [Everitt et al., 2001, p. 105]. However, it is evident that silhouette scores assume clusters that are spherical or Gaussian in shape [Herrmann, 2011, pp. 91-92].

3.3 Problems with Clustering Methods

To illustrate several problems encountered when using common clustering methods, a domain expert measured genetic data for subjects who were known either to be healthy or to have one of 3 subtypes of leukemia. Here, a typical knowledge discovery task could be to identify patterns in the cancer subtypes based on the four diagnoses leading to the prior classification**.**

> *"[I]t is a common practice among researchers to employ a variety of different clustering techniques to analyse a dataset, and to use visual inspection[18] and prior biological knowledge to select what is considered the most 'appropriate' result" [Handl et al., 2005, pp. 3202-3203].*

Consequently, the first step would be to confirm that the structure defined by the classification distinguishing the healthy patients from the non-healthy ones does indeed exist in this data set.

[18] The application of visual inspection will be reported in chapter 6, Fig. 1, resulting in arbitrary projections.

The data set used as an example to illustrate the general problem described above contains data representing 7747 variables for 554 subjects (see chapter 9 for details). Of the subjects, 109 are healthy, 15 have acute promyelocytic leukemia (APL), 266 have chronic lymphocytic leukemia (CLL), and 164 have acute myeloid leukemia (AML). There is a possibility that some subjects might be misclassified, but a future publication will address this diagnostic.

The heatmap and the silhouette plot presented in Figure 3.7 and 3.6 show that this data set is defined by discontinuities because the intracluster distances are small and the intercluster distances large. Hence, the leukemia data set is a high-dimensional data set with natural clusters that are specified by the illness status and defined by discontinuities[19].

Table 3.1 shows the accuracies of common clustering algorithms computed by comparing the clustering results with the prior classification made available by the domain expert. The default settings were used for all algorithms, and the number of clusters was assumed to be four. The MoG algorithm cannot be applied without first using dimensionality reduction methods because the dimensionality of the data set is too high. Only one algorithm (Ward) is able to fully reproduce the prior classification. However, a classification should typically be reproduced using more than one algorithm, and the reproduction of a classification with 100% accuracy is unusual.

This example illustrates that "Clustering algorithms will create clusters whether the data are naturally clustered or purely random" [Jain/Dubes, 1988, p. 201] and "By imposing a predefined shape on the clusters, classical algorithms occasionally suggest a cluster structure in homogenously distributed data or assign points to incorrect clusters" [Ultsch/Lötsch, 2016].

To summarize, the unsupervised indices, namely, the heatmap and the silhouette plot, agree with the prior classification provided by the domain expert, whereas the external index of accuracy and the projections of the data[5] disagree with the domain expert. The question arises whether this data set contains natural clusters and, if so, how the structure of these natural clusters can be correctly identified or how the optimal clustering (or projection) algorithm can be chosen for the knowledge discovery task. This work will propose approaches and solutions to these problems.

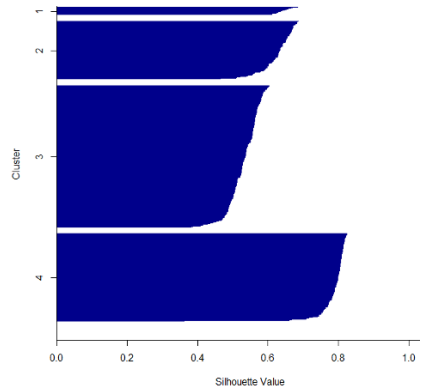

Figure 3.6: Silhouette plot of the leukemia data set indicates a cluster structure.

[19] It should be remarked that common data-driven methods as well as the heatmap and Silhouette plot do not reproduce the (sub) classification(s) of AML (like FAB subtypes) or CLL of research in this area, e.g. [Bene et al., 1995; Bennett et al., 1985; Vardiman et al., 2009; Haferlach et al., 2010], for CLL [Rosenwald et al., 2001].

Table 3.1: Accuracy results for common clustering algorithms.
No result could be calculated for the MoG algorithm (also known as model-based clustering).

Algorithm	Ward	SL	k-means	MoG	PAM	Spectral
Accuracy in %	100	80.1	76.53	*Not Computable*	78.3	59.0

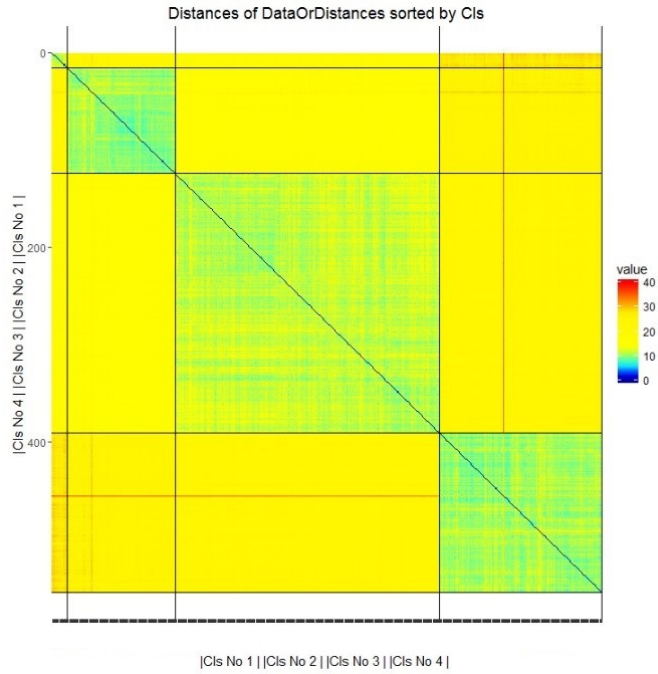

Figure 3.7: The heatmap of the leukemia data set with at least one outlier (red line). The intracluster distances are distinctively smaller than the intercluster distances. Cls1 =APL, Cls2= healthy, Cls3=CLL, Cls4=AML.