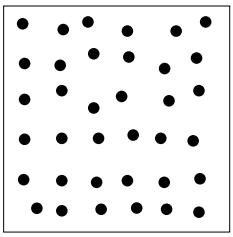

FIGURE 8.23

A data set that is uniformly distributed in the data space.

the points into groups, the groups will unlikely mean anything significant to the application due to the uniform distribution of the data. □

"How can we assess the clustering tendency of a data set?" Intuitively, we can try to measure the probability that the data set is generated by a uniform data distribution. This can be achieved using statistical tests for spatial randomness. To illustrate this idea, let us look at a simple yet effective statistic called the Hopkins statistic.

The **Hopkins Statistic** is a spatial statistic that tests the spatial randomness of a variable as distributed in a space. Given a data set, D, which is regarded as a sample of a random variable, o, we want to determine how far away o is from being uniformly distributed in the data space. We calculate the Hopkins Statistic as follows:

1. Sample n points, $\boldsymbol{p_1}, \ldots, \boldsymbol{p_n}$ from the data space. For each point, $\boldsymbol{p_i}$ $(1 \leq i \leq n)$, we find the nearest neighbor in D, and let x_i be the distance between $\boldsymbol{p_i}$ and its nearest neighbor in D. That is,

$$x_i = \min_{\boldsymbol{v} \in D} \{dist(\boldsymbol{p_i}, \boldsymbol{v})\}. \tag{8.26}$$

2. Sample n points, $\boldsymbol{q_1}, \ldots, \boldsymbol{q_n}$ uniformly from D without replacement. That is, each point in D has the same probability of being included in this sample, and one point can only be included in the sample at most once. For each $\boldsymbol{q_i}$ $(1 \leq i \leq n)$, we find the nearest neighbor of $\boldsymbol{q_i}$ in $D - \{\boldsymbol{q_i}\}$, and let y_i be the distance between $\boldsymbol{q_i}$ and its nearest neighbor in $D - \{\boldsymbol{q_i}\}$. That is,

$$y_i = \min_{\boldsymbol{v} \in D, \boldsymbol{v} \neq \boldsymbol{q_i}} \{dist(\boldsymbol{q_i}, \boldsymbol{v})\}. \tag{8.27}$$

3. Calculate the Hopkins statistic, H, as

$$H = \frac{\sum_{i=1}^{n} x_i^d}{\sum_{i=1}^{n} x_i^d + \sum_{i=1}^{n} y_i^d}, \tag{8.28}$$

where d is the dimensionality of the data set D.

"What does the Hopkins statistic tell us about how likely data set D *follows a uniform distribution in the data space?"* If D is uniformly distributed, then $\sum_{i=1}^{n} y_i^d$ and $\sum_{i=1}^{n} x_i^d$ are close to each other,

and thus H tends to be about 0.5. However, if D is highly skewed, then the points in D are closer to their nearest neighbors than the random points $\boldsymbol{p_1}, \ldots, \boldsymbol{p_n}$ are, and thus $\sum_{i=1}^{n} x_i^d$ shall be substantially larger than $\sum_{i=1}^{n} y_i^d$ in expectation, and H tends to be close to 1.

Example 8.9. Hopkins statistic. Consider a 1-D data set $D = \{0.9, 1, 1.3, 1.4, 1.5, 1.8, 2, 2.1, 4.1, 7, 7.4, 7.5, 7.7, 7.8, 7.9, 8.1\}$ in the data space $[0, 10]$. We draw a sample of four points from D without replacement, say, 1.3, 1.8, 7.5, and 7.9. We also draw a sample of four points uniformly from the data space $[0, 10]$, say, 1.9, 4, 6, 8. Then, the Hopkins statistic can be calculated as

$$H = \frac{|1.9-2| + |4-4.1| + |6-7| + |8-8.1|}{(|1.9-2| + |4-4.1| + |6-7| + |8-8.1|) + (|1.3-1.4| + |1.8-2| + |7.5-7.4| + |7.9-7.8|)}$$
$$= \frac{1.3}{1.3+0.5} = \frac{1.3}{1.8} = 0.72.$$

Since the Hopkins statistic is substantially larger than 0.5 and is close to 1, the data set D has a strong clustering tendency. Indeed, there are two clusters, one around 1.5 and the other one around 7.8. □

In addition to Hopkins statistic, there are some other methods, such as spatial histogram and distance distribution, comparing statistics between a data set under clustering tendency analysis and the corresponding uniform distribution. For example, distance distribution compares the distribution of pairwise distance in the target data set and that in a random uniform sample from the data space.

8.5.2 Determining the number of clusters

Determining the "right" number of clusters in a data set is important, not only because some clustering algorithms like k-means require such a parameter, but also because the appropriate number of clusters controls the proper granularity of cluster analysis. It can be regarded as finding a good balance between *compressibility* and *accuracy* in cluster analysis. Consider two extreme cases. What if you were to treat the entire data set as a cluster? This would maximize the compression of the data, but such a cluster analysis has no value. In contrast, treating each object in a data set as a cluster gives the finest clustering resolution (i.e., most accurate due to the zero distance between an object and the corresponding cluster center). In some methods like k-means, this even achieves the minimum cost. However, having one object per cluster does not enable any data summarization.

Determining the number of clusters is far from easy, often because the "right" number is ambiguous. Figuring out the right number of clusters often depends on the distribution's shape and scale in the data set, as well as the clustering resolution required by the user. There are many possible ways to estimate the number of clusters.

For example, a simple method is to set the number of clusters to about $\sqrt{\frac{n}{2}}$ for a data set of n points. In expectation, each cluster has $\sqrt{2n}$ points. Section 8.2.2 introduces the Calinski-Harabasz index, which estimates the number of clusters for k-means.

Let us look at two more alternative methods.

The **elbow method** is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other. However, the marginal

effect of reducing the sum of within-cluster variances may drop if too many clusters are formed, because splitting a cohesive cluster into two gives only a small reduction. Consequently, a heuristic for selecting the right number of clusters is to use the turning point in the curve of the sum of within-cluster variances with respect to the number of clusters.

Technically, given a number, $k > 0$, we can form k clusters on the data set in question using a clustering algorithm like k-means, and calculate the sum of within-cluster variances, $var(k)$. We can then plot the curve of var with respect to k. The first (or most significant) turning point of the curve suggests the "right" number.

More advanced methods can determine the number of clusters using information criteria or information theoretic approaches. Please refer to the bibliographic notes for further information (Section 8.8).

The "right" number of clusters in a data set can also be determined by **cross-validation**, a technique often used in classification (Chapter 6). First, we divide the given data set, D, into m parts. Next, we use $m - 1$ parts to build a clustering model, and use the remaining part to test the quality of the clustering. For example, for each point in the test set, we can find the closest centroid. Consequently, we can use the sum of the squared distances between all points in the test set and the closest centroids to measure how well the clustering model fits the test set. For any integer $k > 0$, we repeat this process m times to derive clusterings of k clusters, using each part in turn as the test set. The average of the quality measure is taken as the overall quality measure. We can then compare the overall quality measure with respect to different values of k and find the number of clusters that best fits the data.

8.5.3 Measuring clustering quality: extrinsic methods

Suppose you have assessed the clustering tendency of a given data set. You may have also tried to predetermine the number of clusters in the set. You can now apply one or multiple clustering methods to obtain clusterings of the data set. "*How good is the clustering generated by a method, and how can we compare the clusterings generated by different methods?*"

Extrinsic vs. intrinsic methods

We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available. Here, *ground truth* is the ideal clustering that is often built using human experts.

If ground truth is available, it can be used by the **extrinsic methods**, which compare the clustering against the ground truth and measure. If the ground truth is unavailable, we can use the **intrinsic methods**, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of "cluster labels." Hence, extrinsic methods are also known as *supervised methods*, whereas intrinsic methods are *unsupervised methods*.

In this section, we focus on extrinsic methods. We will discuss intrinsic methods in the next section.

Desiderata of extrinsic methods

When the ground truth is available, we can compare it with a clustering to assess the quality of the clustering. Thus the core task in extrinsic methods is to assign a score, $Q(\mathcal{C}, \mathcal{C}_g)$, to a clustering, $\mathcal{C}$, given the ground truth, $\mathcal{C}_g$. Whether an extrinsic method is effective largely depends on the measure, Q, it uses.

In general, a measure Q on clustering quality is effective if it satisfies the following four essential criteria:

- **Cluster homogeneity.** This requires that the purer the clusters in a clustering are, the better the clustering. Suppose that the ground truth says that the objects in a data set, $D = \{a, b, c, d, e, f, g, h\}$, can belong to three categories. Objects a and b are in category L_1, objects c and d belong to category L_2, and the others are in category L_3. Consider clustering, $\mathcal{C}_1 = \{\{a, b, c, d\}, \{e, f, g, h\}\}$, wherein a cluster $\{a, b, c, d\} \in \mathcal{C}_1$ contains objects from two categories, L_1 and L_2. Also consider clustering $\mathcal{C}_2 = \{\{a, b\}, \{c, d\}, \{e, f, g, h\}\}$, which is identical to $\mathcal{C}_1$ except that $\mathcal{C}_2$ is split into two clusters containing the objects in L_1 and L_2, respectively. A clustering quality measure, Q, respecting cluster homogeneity should give a higher score to $\mathcal{C}_2$ than $\mathcal{C}_1$, that is, $Q(\mathcal{C}_2, \mathcal{C}_g) > Q(\mathcal{C}_1, \mathcal{C}_g)$.
- **Cluster completeness.** This is the counterpart of cluster homogeneity. Cluster completeness requires that for a clustering, if any two objects belong to the same category according to the ground truth, then they should be assigned to the same cluster. Cluster completeness requires that a clustering should assign objects belonging to the same category (according to the ground truth) to the same cluster. Continue our previous example. Suppose clustering $\mathcal{C}_3 = \{\{a, b\}, \{c, d\}, \{e, f\}, \{g, h\}\}$. $\mathcal{C}_3$ and $\mathcal{C}_2$ are identical except that $\mathcal{C}_3$ split the objects in category L_3 into two clusters. Then, a clustering quality measure, Q, respecting cluster completeness should give a higher score to $\mathcal{C}_2$, that is, $Q(\mathcal{C}_2, \mathcal{C}_g) > Q(\mathcal{C}_1, \mathcal{C}_g)$.
- **Rag bag.** In many practical scenarios, there is often a "rag bag" category containing objects that cannot be merged with other objects. Such a category is often called "miscellaneous," "other," and so on. The rag bag criterion states that putting a heterogeneous object into a pure cluster should be penalized more than putting it into a rag bag. Consider a clustering $\mathcal{C}_1$ and a cluster $C \in \mathcal{C}_1$ such that all objects in C except for one, denoted by $\boldsymbol{o}$, belong to the same category according to the ground truth. Consider a clustering $\mathcal{C}_2$ identical to $\mathcal{C}_1$ except that $\boldsymbol{o}$ is assigned to a cluster $C' \neq C$ in $\mathcal{C}_2$ such that C' contains objects from various categories according to ground truth, and thus is noisy. In other words, C' in $\mathcal{C}_2$ is a rag bag. Then, a clustering quality measure Q respecting the rag bag criterion should give a higher score to $\mathcal{C}_2$, that is, $Q(\mathcal{C}_2, \mathcal{C}_g) > Q(\mathcal{C}_1, \mathcal{C}_g)$.
- **Small cluster preservation.** If a small category is split into small pieces in a clustering, those small pieces may likely become noise and thus the small category cannot be discovered from the clustering. The small cluster preservation criterion states that splitting a small category into pieces is more harmful than splitting a large category into pieces. Consider an extreme case. Let D be a data set of $n + 2$ objects such that, according to ground truth, n objects, denoted by $\boldsymbol{o_1}, \ldots, \boldsymbol{o_n}$, belong to one category and the other two objects, denoted by $\boldsymbol{o_{n+1}}, \boldsymbol{o_{n+2}}$, belong to another category. Suppose clustering $\mathcal{C}_1$ has three clusters, $C_1^1 = \{\boldsymbol{o_1}, \ldots, \boldsymbol{o_n}\}$, $C_2^1 = \{\boldsymbol{o_{n+1}}\}$, and $C_3^1 = \{\boldsymbol{o_{n+2}}\}$. Let clustering $\mathcal{C}_2$ have three clusters, too, namely $C_1^2 = \{\boldsymbol{o_1}, \ldots, \boldsymbol{o_{n-1}}\}$, $C_2^2 = \{\boldsymbol{o_n}\}$, and $C_3^2 = \{\boldsymbol{o_{n+1}}, \boldsymbol{o_{n+2}}\}$. In other words, $\mathcal{C}_1$ splits the small category and $\mathcal{C}_2$ splits the big category. A clustering quality measure Q preserving small clusters should give a higher score to $\mathcal{C}_2$, that is, $Q(\mathcal{C}_2, \mathcal{C}_g) > Q(\mathcal{C}_1, \mathcal{C}_g)$.

Categories of extrinsic methods

The ground truth may be used in different ways to evaluate clustering quality, which lead to different extrinsic methods. In general, the extrinsic methods can be categorized according to how the ground truth is used as follows.

- **The matching-based methods** examine how well the clustering results match the ground truth in partitioning the objects in the data set. For example, the purity methods assess how a cluster matches only those objects in one group in the ground truth.
- **The information theory–based methods** compare the distribution of the clustering results and that of the ground truth. Entropy or other measures in information theory are often employed to quantify the comparison. For example, we can measure the conditional entropy between the clustering results and the ground truth to measure whether there exists dependency between the information of the clustering results and the ground truth. The higher the dependency, the better the clustering results.
- **The pairwise comparison–based methods** treat each group in the ground truth as a class and then check the pairwise consistency of the objects in the clustering results. The clustering results are good if more pairs of objects of the same class are put into the same cluster, less pairs of objects of different classes are put into the same cluster, and less pairs of objects of the same class are put into different clusters.

Next, let us use some examples to illustrate the above categories of extrinsic methods.

Matching-based methods

The matching-based methods compare clusters in the clustering results and the groups in the ground truth. Let us use an example to explain the ideas.

Suppose a clustering method partitions a set of objects $D = \{o_1, \ldots, o_n\}$ into clusters $\mathcal{C} = \{C_1, \ldots, C_m\}$. The ground truth $\mathcal{G}$ also partitions the same set of objects into groups $\mathcal{G} = \{G_1, \ldots, G_l\}$. Let $C(o_x)$ and $G(o_x)$ $(1 \leq x \leq n)$ be the cluster-id and the group-id of object o_x in the clustering results and the ground truth, respectively.

For a cluster C_i $(1 \leq i \leq m)$, how well C_i matches group G_j in the ground truth can be measured by $|C_i \cap G_j|$, the larger the better. $\frac{|C_i \cap G_j|}{|C_i|}$ can be regarded as the purity of cluster C_i, where G_j matching C_i maximizes $|C_i \cap G_j|$. The purity of the whole clustering results can be calculated as the weighted sum of the purity of the clusters. That is,

$$purity = \sum_{i=1}^{m} \frac{|C_i|}{n} \max_{j=1}^{l} \{\frac{|C_i \cap G_j|}{|C_i|}\} = \frac{1}{n} \sum_{i=1}^{m} \max_{j=1}^{l} \{|C_i \cap G_j|\}. \tag{8.29}$$

The higher the purity, the purer are the clusters, that is, the more objects in each cluster belong to the same group in the ground truth. When the purity is 1, each cluster either matches a group perfectly or is a subset of a group. In other words, no two objects belong to two groups are mixed in one cluster. However, it is possible that multiple clusters partition a group in the ground truth.

Example 8.10. Purity. Consider the set of objects $D = \{a, b, c, d, e, f, g, h, i, j, k\}$. The clustering ground truth and two clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$ output by two methods are shown in Table 8.1.

The purity of clustering $\mathcal{C}_1$ is calculated by $\frac{1}{11} \times (4 + 2 + 4 + 1) = \frac{11}{11} = 1$ and that of clustering $\mathcal{C}_2$ is $\frac{1}{11}(2 + 3 + 1) = \frac{6}{11}$. In terms of purity, $\mathcal{C}_1$ is better than $\mathcal{C}_2$. Please note that, although $\mathcal{C}_1$ has purity 1, it splits G_1 in the ground truth into two clusters, C_1 and C_2. □

There are some other matching based methods further refine the measurement of matching quality, such as maximum matching and using F-measure.

Table 8.1 A set of objects, the clustering ground truth, and two clusterings.

Object	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*
Ground truth $\mathcal{G}$	G_1	G_1	G_1	G_1	G_1	G_1	G_2	G_2	G_2	G_2	G_3
Clustering $\mathcal{C}_1$	C_1	C_1	C_1	C_1	C_2	C_2	C_3	C_3	C_3	C_3	C_4
Clustering $\mathcal{C}_2$	C_1	C_1	C_2	C_2	C_2	C_3	C_1	C_2	C_2	C_1	C_3

Information theory–based methods

A clustering assigns objects to clusters and thus can be regarded as a compression of the information carried by the objects. In other words, a clustering can be regarded as a compressed representation of a given set of objects. Therefore we can use information theory to compare a clustering and the ground truth as representations. This is the general idea behind the information theory–based methods.

For example, we can measure the amount of information needed to describe the ground truth given the distribution of a clustering output by a method. Better the clustering results approach the ground truth, less amount information is needed. This leads to a natural approach using conditional entropy.

Concretely, according to information theory, the entropy of a clustering $\mathcal{C}$ is

$$H(\mathcal{C}) = -\sum_{i=1}^{m} \frac{|C_i|}{n} \log \frac{|C_i|}{n},$$

and the entropy of the ground truth is

$$H(\mathcal{G}) = -\sum_{i=1}^{l} \frac{|G_i|}{n} \log \frac{|G_i|}{n}.$$

The conditional entropy of $\mathcal{G}$ given cluster C_i is

$$H(\mathcal{G}|C_i) = -\sum_{j=1}^{l} \frac{|C_i \cap G_j|}{|C_i|} \log \frac{|C_i \cap G_j|}{|C_i|}.$$

The conditional entropy of $\mathcal{G}$ given clustering $\mathcal{C}$ is

$$H(\mathcal{G}|\mathcal{C}) = \sum_{i=1}^{m} \frac{|C_i|}{n} H(\mathcal{G}|C_i) = -\sum_{i=1}^{m} \sum_{j=1}^{l} \frac{|C_i \cap G_j|}{n} \log \frac{|C_i \cap G_j|}{|C_i|}.$$

In addition to the simple conditional entropy, more sophisticated information theory–based measures may be used, such as normalized mutual information and variation of information.

Taking the case in Table 8.1 as an example, we can calculate

$$H(\mathcal{G}|\mathcal{C}_1) = -(\frac{4}{11} \log \frac{4}{4} + \frac{2}{11} \log \frac{2}{2} + \frac{4}{11} \log \frac{4}{4} + \frac{1}{11} \log \frac{1}{1}) = 0$$

and

$$H(\mathcal{G}|\mathcal{C}_2) = -(\frac{2}{11}\log\frac{2}{4} + \frac{2}{11}\log\frac{2}{4} + \frac{3}{11}\log\frac{3}{5} + \frac{2}{11}\log\frac{2}{5} + \frac{1}{11}\log\frac{1}{2} + \frac{1}{11}\log\frac{1}{2})$$
$$= 0.297.$$

Clustering $\mathcal{C}_1$ has better quality than $\mathcal{C}_2$ in terms of conditional entropy. Again, although $H(\mathcal{G}|\mathcal{C}_1) = 0$, conditional entropy cannot detect the issue that $\mathcal{C}_1$ splits the objects in G_1 into two clusters.

Pairwise comparison–based methods

The pairwise comparison–based methods treat each group in the ground truth as a class. For each pair of objects $o_i, o_j \in D$ $(1 \leq i, j \leq n, i \neq j)$, if they are assigned to the same cluster/group, the assignment is regarded as positive, and otherwise, negative. Then, depending on assignments of o_i and o_j into clusters $C(o_i)$, $C(o_j)$, $G(o_i)$, and $G(o_j)$, we have four possible cases.

	$C(o_i) = C(o_j)$	$C(o_i) \neq C(o_j)$
$G(o_i) = G(o_j)$	true positive	false negative
$G(o_i) \neq G(o_j)$	false positive	true negative

Using the statistics on pairwise comparison, we can assess the quality of the clustering results approaching the ground truth. For example, we can use the Jaccard coefficient, which is defined as

$$J = \frac{\text{true positive}}{\text{true positive} + \text{false negative} + \text{false positive}}.$$

Many other measures can be built based on the pairwise comparison statistics, such as Rand statistic, fowlkes-Mallows measure, BCubed precision, and recall. The pairwise comparison results can be further used to conduct correlation analysis. For example, we can form a binary matrix $\boldsymbol{G}$ according to the ground truth, where element $v_{ij} = 1$ if $G(o_i) = G(o_j)$, and otherwise 0. A binary matrix $\boldsymbol{C}$ can also be constructed in a similar way based on a clustering $\mathcal{C}$. We can analyze the element-wise correlation between the two matrixes and use the correlation to measure the quality of the clustering results. Clearly, the more correlated the two matrixes, the better the clustering results.

8.5.4 Intrinsic methods

When the ground truth of a data set is not available, we have to use an intrinsic method to assess the clustering quality. Unable to reference any external supervision information, the intrinsic methods have to come back to the fundamental intuition in clustering analysis, that is, examining how compact clusters are and how well clusters are separated. Many intrinsic methods take the advantage of a similarity or distance measure between objects in the data set.

For example, the **Dunn index** measures the compactness of clusters by the maximum distance between two points that belong to the same cluster, that is, $\Delta = \max_{C(o_i)=C(o_j)}\{d(o_i, o_j)\}\}$. It measures the degree of separation among different clusters by the minimum distance between two points that belong to different clusters, that is $\delta = \min_{C(o_i)\neq C(o_j)}\{d(o_i, o_j)\}$. Then, the Dunn index is simply the ration $DI = \frac{\delta}{\Delta}$. The larger the ratio, the farther away the clusters are separated comparing to the compactness of the clusters.

The Dunn index uses the extreme distances to measure the cluster compactness and intercluster separation. The measures δ and Δ may be affected by the outliers. Many methods consider the average situations. The **silhouette coefficient** is such a measure. For a data set, D, of n objects, suppose D is partitioned into k clusters, $C_1, \ldots, C_k$. For each object $\boldsymbol{o} \in D$, we calculate $a(\boldsymbol{o})$ as the average distance between $\boldsymbol{o}$ and all other objects in the cluster to which $\boldsymbol{o}$ belongs. Similarly, $b(\boldsymbol{o})$ is the minimum average distance from $\boldsymbol{o}$ to all clusters to which $\boldsymbol{o}$ does not belong. Formally, suppose $\boldsymbol{o} \in C_i$ $(1 \leq i \leq k)$. Then,

$$a(\boldsymbol{o}) = \frac{\sum_{\boldsymbol{o}' \in C_i, \boldsymbol{o} \neq \boldsymbol{o}'} dist(\boldsymbol{o}, \boldsymbol{o}')}{|C_i| - 1} \tag{8.30}$$

and

$$b(\boldsymbol{o}) = \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{\boldsymbol{o}' \in C_j} dist(\boldsymbol{o}, \boldsymbol{o}')}{|C_j|} \right\}. \tag{8.31}$$

The **silhouette coefficient** of $\boldsymbol{o}$ is then defined as

$$s(\boldsymbol{o}) = \frac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{\max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}. \tag{8.32}$$

The value of the silhouette coefficient is between -1 and 1. The value of $a(\boldsymbol{o})$ reflects the compactness of the cluster to which $\boldsymbol{o}$ belongs. The smaller the value, the more compact the cluster. The value of $b(\boldsymbol{o})$ captures the degree to which $\boldsymbol{o}$ is separated from other clusters. The larger $b(\boldsymbol{o})$ is, the more separated $\boldsymbol{o}$ is from other clusters. Therefore when the silhouette coefficient value of $\boldsymbol{o}$ approaches 1, the cluster containing $\boldsymbol{o}$ is compact and $\boldsymbol{o}$ is far away from other clusters, which is the preferable case. However, when the silhouette coefficient value is negative (i.e., $b(\boldsymbol{o}) < a(\boldsymbol{o})$, this means that, in expectation, $\boldsymbol{o}$ is closer to the objects in another cluster than to the objects in the same cluster as $\boldsymbol{o}$. In many cases, this is a bad situation and should be avoided.

To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set. The silhouette coefficient and other intrinsic measures can also be used in the elbow method to heuristically derive the number of clusters in a data set by replacing the sum of within-cluster variances.

8.6 Summary

- A **cluster** is a collection of data objects that are *similar* to one another within the same cluster and are *dissimilar* to the objects in other clusters. The process of grouping a set of physical or abstract objects into classes of *similar* objects is called **clustering**.
- Cluster analysis has extensive **applications**, including business intelligence, image pattern recognition, Web search, biology, and security. Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution or as a preprocessing step for other data mining algorithms operating on the detected clusters.
- Clustering is a dynamic field of research in data mining. It is related to **unsupervised learning** in machine learning.