Glossary

**Basic clustering** The clustering performed for the purpose of generating a single data partition for the subsequent consensus clustering. See Chap. 7 and the term “consensus clustering”.

**Class imbalance problem** The challenge to data mining tasks where the data have multiple classes (or true clusters) in varying sizes. See Chaps. 2, 5, and 6.

**Cluster analysis** The data analysis task that attempts to partition data objects into multiple clusters (or groups) without using external information, such that objects in a cluster are more similar to each other than to objects in different clusters. See Chap. 1.

**Cluster validity** Using external or internal validation measures to evaluate clustering results in a quantitative and objective way. See Chap. 5 and the term “external validation measure”.

**Consensus clustering** Also known as cluster ensemble or clustering aggregation, an NP-complete combinatorial optimization problem that aims to find a single clustering from multi-source basic clusterings such that this single clustering matches all the basic clusterings as much as possible. See Chap. 7.

**Data smoothing** A technique that adds a small positive real number to all data objects such that the sparsity of high-dimensional data, such as text corpora, can be eliminated. See Chap. 4.

**External validation measure** A category of cluster validity measures that evaluates clustering results by comparing them to the true cluster structures defined by external information such as class labels. See Chap. 5.

**Fuzzy c-means** A type of prototype-based fuzzy clustering algorithms that acts like K-means clustering but allows a data object to belong to two or more clusters with a membership grade between zero and one. See Chap. 3 and the term “K-means”.

**Information-theoretic K-means** The K-means algorithm using the Kullback-Leibler divergence (KL-divergence) as the distance function. See Chap. 4 and the term “K-means”.

**K-means** A type of prototype-based clustering algorithms that assigns data objects to closest clusters by computing the distances between the data objects and the centroids of the clusters. It can be also viewed as a special case of fuzzy c-means when the fuzzy factor tends to one. See Chap. 1 and the term “fuzzy c-means”.

**Local clustering** A data decomposition technique that performs clustering on a subset of data, e.g. the major class of data. See Chap. 6.

**Measure normalization** The issue that attempts to normalize the cluster validity measures into a small value range such as [0,1] or [−1,1], for the purpose of comparing clustering quality. See Chap. 5.

**Point-to-centroid distance** The only family of distance functions that fits directly K-means clustering with centroids of arithmetic means. See Chap. 3 and the term “K-means”.

**Rare class analysis** The task of classification analysis on highly imbalanced data with the emphasis on identifying positive instances of rare classes. It plays a vital role in many important real-life applications, such as network intrusion detection, credit-card fraud detection, and facility fault detection. See Chap. 6.

**Resampling** A technique that draws randomly with or without replacement from the available data for generating a smaller (under-sampling) or a larger (over-sampling) subset of that data. See Chap. 6.

**Spherical K-means** The K-means algorithm using the cosine similarity as the proximity function. See Chap. 4 and the term “K-means”.

**Uniform effect** The effect of K-means that tends to partition data objects into clusters in uniform sizes. This is a negative effect when applying K-means for class imbalance data. See Chap. 2 and the term “K-means”.

**Variable neighborhood search** An optimization meta-heuristic which exploits systematically the idea of neighborhood change, both in the descent to local minima and in the escape from the valleys that contain them. See Chap. 4.

**Zero-value dilemma** The problem in computing KL-divergence between data objects and centroids when there exist zero-value features in the high-dimensional feature space. See Chap. 4.