Article type: Advanced Review

Density-based Clustering Preprint

Hans-Peter Kriegel

Ludwig-Maximilians-Universität München

Peer Kröger

Ludwig-Maximilians-Universität München

Jörg Sander

University of Alberta, Edmonton

Arthur Zimek

Ludwig-Maximilians-Universität München

Abstract

Clustering refers to the task of identifying groups or clusters in a data set. In *density-based clustering*, a cluster is a set of data objects spread in the data space over a contiguous region of high density of objects. Density-based clusters are separated from each other by contiguous regions of low density of objects. Data objects located in low-density regions are typically considered noise or outliers.

Introduction

Clustering is the problem of finding a set of groups of similar objects within a data set while keeping dissimilar objects separated in different groups or the group of noise. In general, clustering is seen as an unsupervised learning task. Given the data as a set of objects from a given data space $\mathcal{D} \subset \mathcal{S}$ and a dissimilarity-function $dis : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_0^+$, the task is to learn a meaningful grouping of the data. Often, the data set is a set of d-dimensional real-valued points, i.e., $\mathcal{D} \subset \mathcal{S} = \mathbb{R}^d$, which can be seen as a sample from some unknown

probability density $p(x)$, with dis as the Euclidean or some other form of distance. The meaning of within-group similarity and between-group dissimilarity is different for different families of clustering approaches. From a procedural point of view, many clustering methods seek a partitioning of the data set into a predefined number k of groups where the sum of (squared) pairwise dissimilarities between all cluster objects or the sum of (squared) dissimilarities of all cluster objects w.r.t. some cluster representative (e.g. the mean value) are minimized and the respective values between different clusters are maximized, given some dissimilarity function dis. These assumptions typically result in clusters of convex (or, ideally, spherical) shape (e.g. [28]). From a statistical point of view, such methods correspond to a parametric approach where the unknown density $p(x)$ of the data is assumed to be a mixture of k densities $p_i(x)$, each corresponding to one of the k groups in the data. The $p_i(x)$ are assumed to belong to some parametric family (e.g., Gaussian distributions) with unknown parameters. The parameters are then to be estimated based on the given sample (the data set) (e.g. [11]). In contrast, *density-based* clustering is a non-parametric approach where the clusters are considered to be high density areas of the density $p(x)$. Density-based clustering methods do not require the number of clusters as input parameters, nor do they make assumptions concerning the underlying density $p(x)$ or the variance within the clusters that may exist in the data set. As a consequence, density-based clusters are not necessarily groups of points with a low pairwise within-cluster dissimilarity as measured by a dissimilarity function dis and, thus, do not necessarily have a convex shape but can be arbitrarily shaped in the data space. Intuitively, a density-based cluster is a set of data objects spread in the data space over a contiguous region of high density of objects, separated from other density-based clusters by contiguous regions of low density of objects.

In this article, the intuition motivating the density-based definition of clusters is presented along with an overview of statistical methods and efficient algorithms applicable to large databases. Finding suitable density-thresholds is an inherently difficult problem, alleviation has been proposed by hierarchical methods. Finally, adaptations to different specialized problem settings are discussed.

Background and Motivation

Consider the distribution of data points in Figure 1. Density-based clusters can be imagined as sets of points resulting from a "cut" through (an estimation of) the probability density function for the data at a certain density level: each cut induces separate connected regions in the feature space where the probability density is higher than the cut value; each such region corresponds to a cluster containing all the data points falling into this region. If the level is chosen too low, different clusters will be merged to a single cluster (Figures 1(c) and 1(d)). If the density level is chosen too high, clusters exhibiting a lower density will be lost (Figures 1(g) and 1(h)). In this data set, a good density level can be defined to separate the three original clusters relatively well (Figures 1(e) and 1(f)).

(a) Data set

(b) Probability density function of the data

(c) Two clusters merged

(d) Low density level

(e) Three clusters separated

(f) Good density level

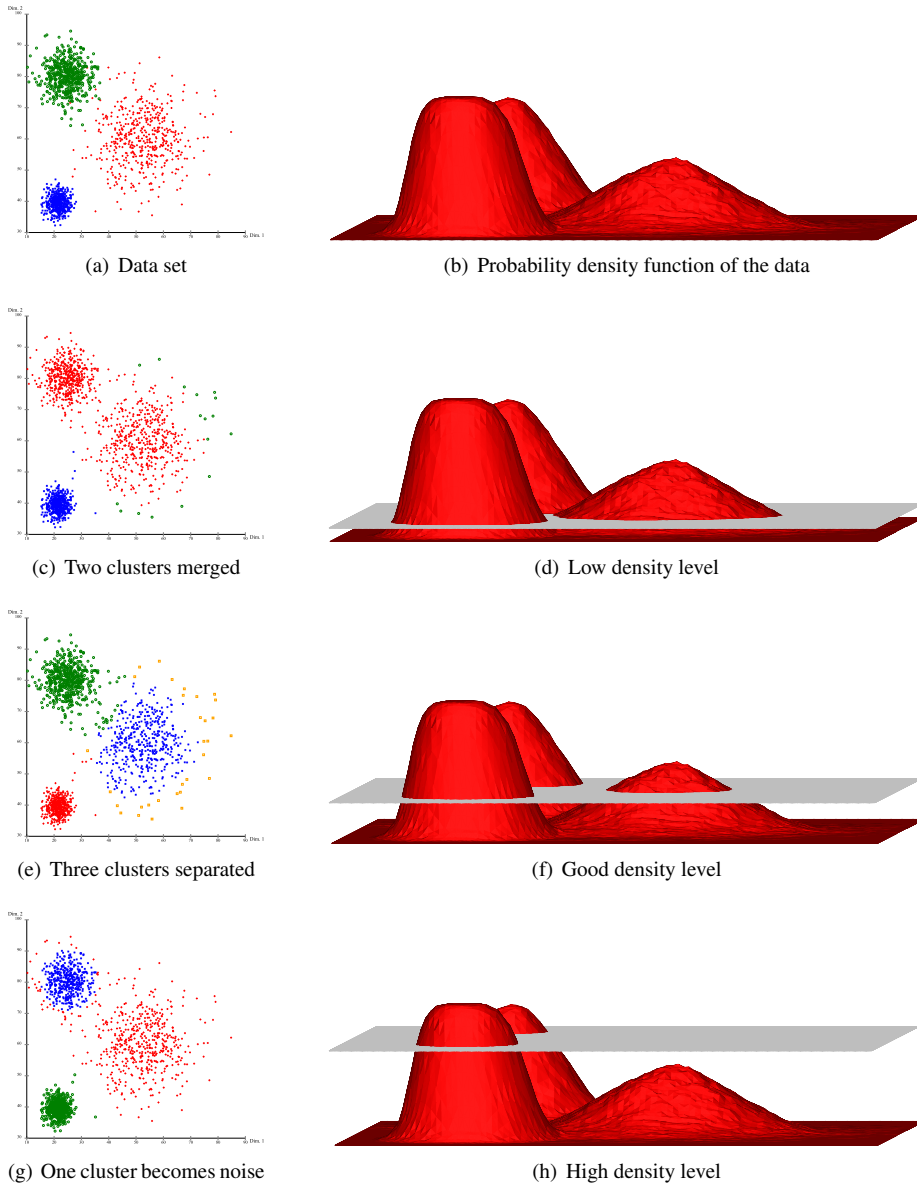(g) One cluster becomes noise

(h) High density level

Figure 1: Density-distributions of data points and density-based clusters for different density levels. Different colors indicate different clusters or noise.
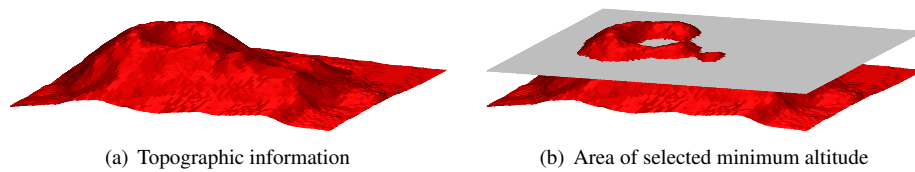
<div align="center">(a) Topographic information   (b) Area of selected minimum altitude</div>

<div align="center">Figure 2: Maunga Whau Volcano (Mt Eden)</div>

Sometimes density-based clusters are also referred to as "natural clusters" since they are particularly suitable for certain nature inspired applications. For example, in spatial data, clusters of points in the space may form along natural structures such as rivers, roads, seismic faults, etc. Figure 2(a) depicts topographic information for Auckland's Maunga Whau Volcano (Mt. Eden). Selecting areas above a certain altitude relates to density-based clustering provided a density level is selected accordingly (Figure 2(b)).

Another natural motivation is often found in zoology [9, 17]. Species diverge, as is commonly assumed, from some common ancestor in different directions as adaptations to different forces of selection. Hence, zoologists define families of different species, where the members probably exhibit not an extraordinary high pairwise similarity to all other members but exhibit a certain similarity to at least some other member of the family [39, 16]. Also, geometrically, evolutionary development can be seen as a tree of paths through the space spanned by the variables describing the phenotypes of organisms. Clusters of such geometry closely resemble the problem described by the density-based clustering approach.

An example for the geometry of variations can be given based on Fisher's Iris data set [15] (well known as an example data set in the domain of classification). It comprises four descriptors of the Iris flower, namely length and width of petals and sepals, respectively. These descriptors are collected for individual flowers of three different species. While the original purpose of these data was to study linear separability, it can serve here to illustrate the typical properties of natural (biological) clusters. The scatter plot given in Figure 3 shows that the occurring natural clusters are typically not spherical in shape, since the extreme individuals within each species are not as similar to each other as they are perhaps to some instance of another species, yet there can be found a chain of similar individuals more or less connecting the individuals within each species. (For another well-known example with discussion see Text-Fig. 4 in [34].)

Statistical Intuition of Density-Based Clustering

An early computational technique following the given intuition of "natural clusters" has become well-known as single-linkage clustering [39]. A hierarchical method has been described in [41]. The original approach of single-linkage was to group all objects below a given distance threshold at a first level, then increase the threshold and repeat the procedure until all objects belong to a group. Though more efficient meth-
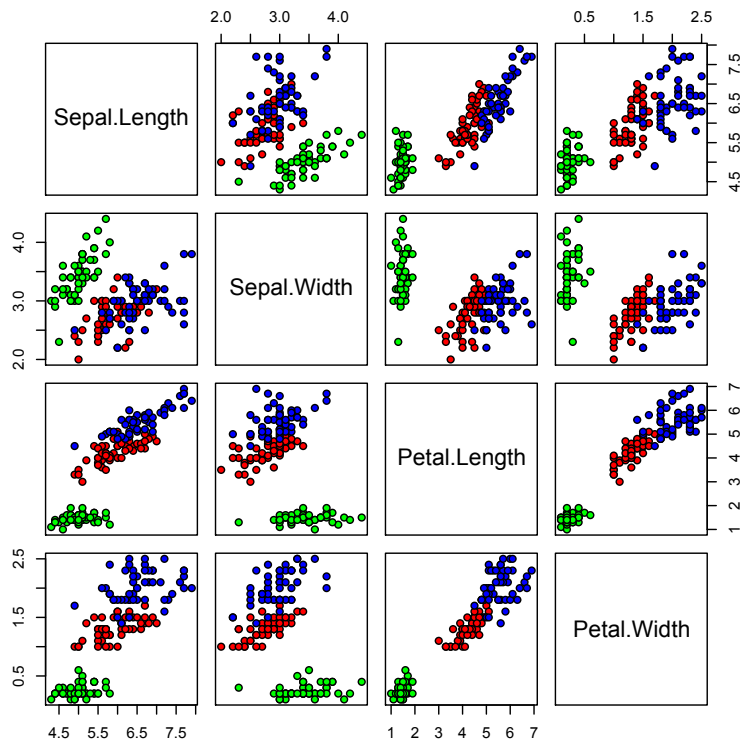
Figure 3: Iris Data (green: *I. setosa*, red: *I. versicolor*, blue: *I. virginica*)

ods have been proposed later [38], single-linkage has been criticized early for the so-called "chaining-effect" (or "single-link effect") [25, 22]. This effect can lead to the undesirable connectivity of different clusters by the existence of a "chain" of single objects between two clusters. Consequently, Wishart proposed a method to remove noise objects (causing the chaining-effect) before clustering the remaining points in a single-linkage manner [42]. This proposal was probably the first approach of computing a density-based clustering. His algorithm for *one level mode analysis* consists of six steps:

1. Selection of a distance threshold r, and a frequency threshold k.

2. Computation of the triangular similarity matrix of all inter-point distances.

3. Evaluation of the frequency k_i of each data point x_i. The frequency is the number of points located within a distance r of point x_i, hence it is a measure of the density.

4. Removal of the non-dense ("noise") points (where $k_i < k$).

5. Clustering of the remaining points by single-linkage.

6. Finally, non-dense points can be allocated to a suitable cluster according to some criterion.

"Some criterion" remains intentionally vague. Wishart gives as an example to include non-dense points in the cluster containing its nearest dense point. This is certainly the most intuitive solution, but other solutions might be possible, including to not assign the dense points to any cluster at all. The latter would also be reasonable since the clusters are defined as to exhibit a certain density. Including the non-dense points weakens the relationship between cluster definition and clusters.

A more general formalization of a density-based cluster was proposed by [18]. Given a density $p(x)$ at each point x, a density threshold λ, and links specified for some pairs of objects, a *density-contour cluster* at level λ is defined as a maximally connected set of points x_i such that $p(x_i) > \lambda$. The links between points may be specified using the distance function dis. For example, two points can be defined as linked if the distance between them does not exceed a threshold r. Hartigan, like Wishart, proposes an algorithm similar to single linkage clustering that computes the maximal connected sets of points having a density greater than the given threshold λ in order to determine density-contour clusters.

Both methods are based upon a similar intuition of what constitutes a cluster. The basic assumption of this intuition is that the data set $\mathcal{D} \subset \mathbb{R}^d$ is a sample from some unknown probability density $p(x)$ and clusters are high density areas of this density $p(x)$. Finding such high density areas of $p(x)$ usually requires two basic ingredients. The first ingredient is a local density estimate at each point. Typically some kernel or nearest neighbor density estimate is used for this issue [35, 31, 27, 29, 12]. The second ingredient is a notion of connection between objects. Typically points are connected if they are within a certain distance ε from each other. Clusters are constructed as

maximal sets of objects which are directly or transitively connected to objects whose density exceeds some threshold λ. The set $\{x|p(x) > \lambda\}$ of all high density objects is called the *density level set* of p at λ. Objects that are not part of such clusters are called *noise* or *outliers*.

The various density based methods proposed in the literature (examples include [43, 10]) mainly differ in the following aspects:

1. How is the density $p(x)$ estimated?

2. How is the notion of connectivity defined?

3. How is the algorithm for finding connected components of the induced graph implemented and is the algorithm supported by suitable data structures to achieve scalability even for large data sets?

In addition, for some methods a cluster consists only of objects whose density exceeds the threshold λ, while other methods also include objects with lower density into a cluster if they are connected to at least one object with density above the threshold λ.

Efficient Density-based Clustering Algorithms

Beside the single-linkage algorithms of Wishart and Hartigan, other algorithms have been proposed to compute density based clustering independently in a different research community. In the context of databases and large data sets, the efficient computability has come into focus and has led to algorithmic variants of computing the density-based clustering model.

The DBSCAN [14] algorithm claims to be scalable to large databases because it allows the use of index structures for density estimation. Given a distance threshold r and a density threshold k (in DBSCAN the threshold is called $minPts$), density of a point x_i is defined, like it is done by Wishart, as the number of points k_i that are within a radius r around x_i. If $k_i > k$, the corresponding point x_i is considered a *core point*. Two points are considered directly connected if they have a distance of less than r. Two points are density connected if they are connected to core points and these core points are in turn density connected. These definitions allow to define the transitive hull of density-connected points, forming density-based clusters.

As an illustration of this concept, points p and m, m and n, n and q in Figure 4 are direct density reachable, respectively, under $minPts = 5$ and the sketched ε. By transitivity, also p and n or q are (indirectly) density reachable, respectively. Hence p and q are density connected via m and n.

The parameters ε and $minPts$ together define a density level. Reconsider Figure 1: there, the clusters shown in Figures 1(c), 1(e), 1(g), respectively, have been derived with DBSCAN and parameters $\varepsilon = 10$, $minPts = 100, 150, 455$. In this example,
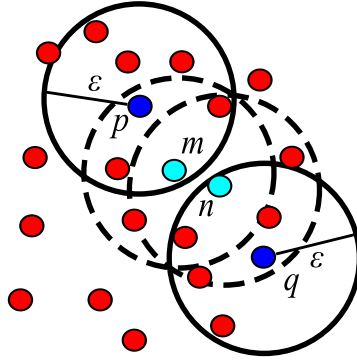
7

Figure 4: Density connectivity.

with a fixed value for ε, the varying density level is expressed with varying values for $minPts$.

A density based cluster is formalized as a maximally connected component of the set of points that have a distance of smaller than r to some core point. Clusters may contain so-called *border points* that do not have the core point property. Objects that are not part of a cluster are *noise points*. In an extreme case, a cluster may only contain one core point and the border points in its neighborhood. In some applications, introducing a minimum number of core points per cluster through an additional parameter can reduce the number of "spurious" clusters that would otherwise be reported.

DBSCAN computes clusters iteratively. It starts a new cluster C with a not yet assigned core point x by assigning all points to C that are connected to x. For determining the connected points of a given point x of a cluster C range queries with radius r are computed for each point that is newly added to C. Each range query has a runtime complexity of $O(N)$ when applying a sequential search, resulting in a total runtime complexity of $O(N^2)$ in the worst case. In moderately dimensional data one could apply appropriate index structures such as an R*-tree to process a range query in $O(\log N)$. Then, the runtime complexity of DBSCAN is in $O(N \log N)$. The runtime complexity for creating such index structures is usually also in $O(N \log N)$ and, thus, does not affect the overall runtime complexity.

DENCLUE [20] uses a kernel density estimator to define density based clusters. In general, any density estimator can be used. Figure 5 depicts an example 2D data set (left) and the resulting density estimation using two different kernels, a Square Wave (uniform) kernel (top) and a Gaussian kernel (bottom). A local maximum of the overall density function is called a *density attractor*. Each point x is associated with the density attractor that is located in the direction of maximum increase in density from x. A density based cluster is defined as a connected component of density attractors with their associated points whose density estimate is above a given threshold λ. In fact, the method of Wishart and DBSCAN can be seen as special cases of DENCLUE using a uniform spherical kernel. While for Wishart's method points that are associ-

8

ated by density attractors but whose density is below λ are not included, these points correspond to border points in DBSCAN and, thus, are included for DBSCAN. The implementation of DENCLUE relies on a Gaussian kernel and a sophisticated data structure for fast local density estimation. In particular, the data space is partitioned by a grid. Each non-empty grid cell c is mapped to a one-dimensional key. This key is stored in a search tree along with further statistics such as number of points in c, pointers to the points in c, and the linear sum of the points in c. The search tree is used for efficiently determining neighboring cells and local density estimates. Although proposing a generalization, the authors of [20] report also a runtime speed-up of DENCLUE over DBSCAN.
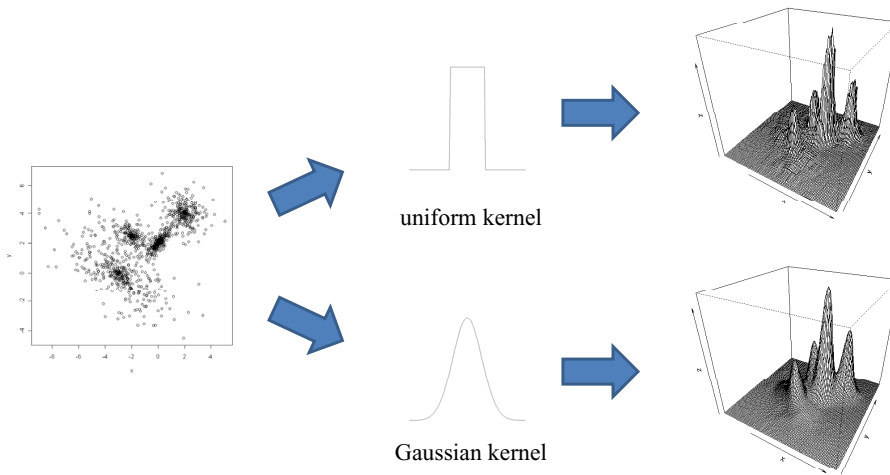

Figure 5: The impact of different kernels on the density estimation.

GDBSCAN [36] is an algorithmic framework that generalizes the notion of density based clusters to the concept of *density connected decomposition* for any type of data. This concept only requires a reflexive and symmetric *neighborhood* relation for any pair of objects that represents direct connectivity, and a *MinWeight* predicate that evaluates *true* for neighborhood sets of core objects and *false* for neighborhood sets of non-core objects. A density connected decomposition consists of the maximally connected components of the set of objects that are in the neighborhood set of some core object. They can be computed with the same algorithmic schema used by the DBSCAN algorithm. Examples of specializations of GDBSCAN, described in [36], include simple forms of region growing, e.g., for clustering images, clustering geometrically extended objects taking their intersection into account, and clustering spatial objects while taking also non-spatial attributes into account.

9

Discovering Clusters of Different Densities

One of the basic problems of computing a density based clustering is to determine a suitable density level (cf. Figure 1). This becomes especially difficult or impossible if clusters in different regions of the data space have considerably different local densities or clusters with different density levels are nested (i.e., there are clusters with higher density levels within a cluster with a lower density level). Clearly, in such cases, the clusters derived from a single density-level cannot completely describe the inherent clustering structure of the data set.

The challenge of finding density-based clusters with very different densities in a data set, which cannot be completely described by a single density level set, has been addressed, e.g., by [13, 32] for point data. In addition, this challenge also motivated hierarchical algorithms to compute clusters at different density levels in one single run [42, 18]. Since connected components from different density levels are either disjoint or the cluster of higher density is completely contained in the cluster of lower density, the result of such hierarchical algorithms can be represented as a tree. Other, more recent approaches for hierarchical density based clustering include OPTICS [4] (which is based on DBSCAN and its density estimation) and the work by Stuetzle [40]. Both algorithms are inspired by single linkage clustering computing a Minimum Spanning Tree of the data where edge weights represent pairwise distances. These distances are smoothed by a density estimator in [4] called *core distance*. The core distance of a point x is the smallest threshold r such that x is still considered a core object by the DBSCAN algorithm (see above), i.e., x has still at least k objects in its neighborhood with radius r. The resulting distance that is used to construct the Minimum Spanning Tree is called *reachability distance*. Taking k as input parameter for smoothing the density estimation, the reachability distance of point x is defined relative to a reference object y as the minimum of the core distance of y and the actual distance between x and y. This distance is not symmetric for $k > 1$ (for $k = 1$ it is equivalent to the distance used for nearest-neighbor computation). Obviously, the smaller the reachability distance of a point is the higher is the density around it. The OPTICS algorithm essentially computes a walk through the Minimum Spanning Tree, the so-called *cluster order*, starting at a randomly chosen point and then visiting in each step that object that has the smallest reachability distance relative to all objects visited so far.

The output of hierarchical algorithms is not a partition of the data but a characterization of the density structure of the data, usually a visualization of the Minimum Spanning Tree called *dendrogram*. From such a characterization, clusters at different density levels can usually be determined. In [4] a different 2D characterization is proposed for the OPTICS algorithm called *reachability plot*. The order in which OPTICS visits the points is displayed along the x-axis while the reachability distance of each point is displayed as a bar along the y-axis (cf. Figure 6). Valleys in this plot indicate clusters. The deeper the valley, the higher the corresponding density-level. The relationship between dendrograms and reachability plots is described in [37]. For $k = 1$ a reachability plot can be transferred into a dendrogram and vice versa.
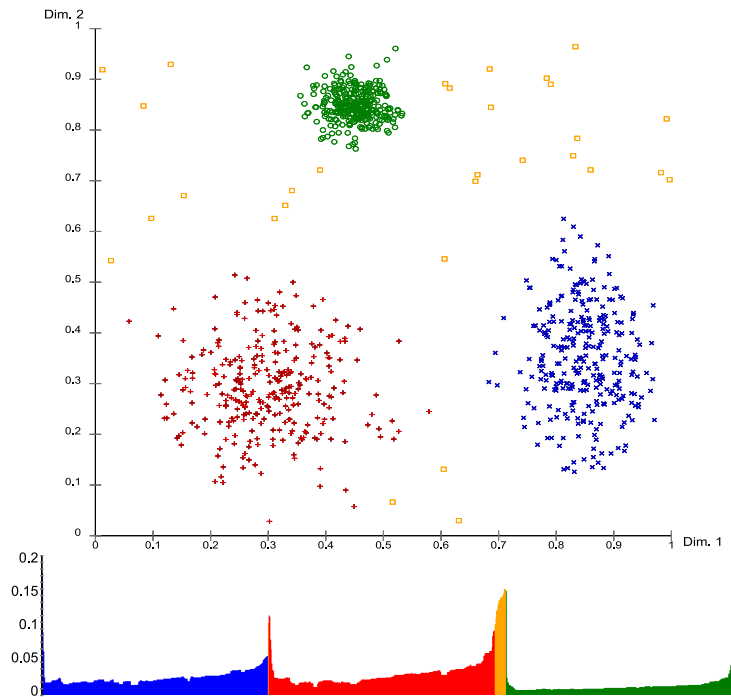
Figure 6: Reachability plot for a sample 2D data set.

A discussion of hierarchical methods, different linkage-methods, and their relation to k-means is provided in [19]. A short overview of algorithmic paradigms for nonparametric clustering is given in [30].

The shared nearest neighbor (SNN) clustering [13] uses a similarity measure based on the number of shared neighbors of two objects instead of a traditional distance measure like Euclidean distance. The algorithm proposed in [13] can be considered as an instance of the GDBSCAN framework where the neighborhood predicate is based on SNN similarity. Using shared nearest neighbor similarity is a possibility to account for different density levels.

DECODE [32] assumes the data is generated by different point processes and tries to find clusters as connected regions of points whose distances to their m-th nearest neighbor (="core-distance" in OPTICS) are "similar". The method determines bins into which the m-nearest neighbor distances are classified, using a reversible jump Markov Chain Monte Carlo strategy; m-nearest neighbor distances that fall into the same bin are then considered similar.

11

Recent Directions in Density-Based Clustering

A large number of specializations and extensions for different applications and data types have been proposed. In the following, two examples of such specializations are sketched: clustering high dimensional data and semi-supervised clustering.

For applications where the data points are high dimensional (sometimes, 10-20 attributes are already regarded as high-dimensional data), specialized approaches for density-based clustering have been proposed to overcome the so-called *curse of dimensionality*, i.e., a number of difficulties arising with increasing dimensionality of a data set (many difficulties usually become apparent between 10 and 20 dimensions). A detailed discussion on the effects of high dimensional data and the curse of dimensionality in the context of clustering can be found in [21, 24].

The authors of [13] claim shared nearest neighbor (SNN) clustering to be more stable in higher dimensions than standard distance measures like L_p norms. This claim has been experimentally evaluated only recently [21]. A different approach to solve the problems of high dimensional data spaces is to search for clusters in subspaces of the original data space. The density-based approach has been applied successfully to search for clusters in axis-parallel subspaces (e.g. [23, 6, 5]) and arbitrarily oriented subspaces (e.g. [7, 2, 1]). Essentially, these approaches can be seen as special instances of the GDBSCAN framework using different neighborhood predicates.

Semi-supervised density-based clustering approaches try to exploit information about cluster membership, which may be available in form of class labels for a small subset of objects. It is based on the intuition that class membership information could be used to derive information about the density parameters that characterize different clusters, when assuming that the labels are *consistent* with the density based clustering structure of the data in the sense that objects with different labels have to be in different density based clusters — this excludes clusters with *mixed* labels, but does not exclude having more than one cluster in which objects have the same label. Two recent proposals are [8, 26].

Conclusion

Density-based clustering has been applied successfully for cluster analysis in many different contexts. In general, density-based clustering aims at identifying clusters as areas of high point density that are separated by areas of low point density and, thus, can be arbitrarily shaped in the data space. Several similar formalizations have been proposed for density-based clusters in different communities. Typically, these approaches are based on the common idea of defining clusters as connected dense components. The approaches usually only differ in the definition of density and connectedness. Also different algorithms have been proposed in different contexts and communities for computing density-based clusters. While most of the original approaches in statistics did not focus on efficiency issues, successive algorithms in the data mining and the database communities were designed for large databases. In addition, the

problem of finding suitable density-thresholds is an inherently difficult problem, alleviation has been proposed by hierarchical methods. Finally, several extensions and adaptations of the basic notion of density-based clustering have been discussed recently.

Notes

The data set of topographical information on Auckland's Maunga Whau Volcano (Mt. Eden), based on a topographic map by Ross Ihaka, (Figure 2) and the Iris data (Figure 3) come with R [33]. Most Figures have been plotted with R (we acknowledge the assistence of Marisa Thoma) or with ELKI [3].

References

[1] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek. On exploring complex relationships of correlation clusters. In *Proceedings of the 19th International Conference on Scientific and Statistical Database Management (SSDBM), Banff, Canada*, 2007.

[2] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek. Robust, complete, and efficient correlation clustering. In *Proceedings of the 7th SIAM International Conference on Data Mining (SDM), Minneapolis, MN*, 2007.

[3] E. Achtert, H.-P. Kriegel, L. Reichert, E. Schubert, R. Wojdanowski, and A. Zimek. Visual evaluation of outlier detection models. In *Proceedings of the 15th International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan*, 2010.

[4] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Philadelphia, PA*, 1999.

[5] I. Assent, R. Krieger, E. Müller, and T. Seidl. DUSC: dimensionality unbiased subspace clustering. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), Omaha, NE*, 2007.

[6] C. Böhm, K. Kailing, H.-P. Kriegel, and P. Kröger. Density connected clustering with local subspace preferences. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM), Brighton, UK*, 2004.

[7] C. Böhm, K. Kailing, P. Kröger, and A. Zimek. Computing clusters of correlation connected objects. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Paris, France*, 2004.

[8] C. Böhm and C. Plant. HISSCLU: a hierarchical density-based method for semi-supervised clustering. In *Proceedings of the 11th International Conference on Extending Database Technology (EDBT), Nantes, France*, 2008.

[9] J. W. Carmichael, J. A. George, and R. S. Julius. Finding natural clusters. *Systematic Zoology*, 17(2):144–150, 1968.

[10] A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Computational Statistics and Data Analysis*, 36:441–459, 2001.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–31, 1977.

[12] L. P. Devroye and T. J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *The Annals of Statistics*, 5(3):536–540, 1977.

[13] L. Ertöz, M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM), San Francisco, CA*, 2003.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR*, 1996.

[15] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[16] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(3760):279–284, 1967.

[17] J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

[18] J. A. Hartigan. *Clustering Algorithms*. John Wiley&Sons, 1975.

[19] J. A. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2(1):63–76, 1985.

[20] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the 4th ACM International Conference on Knowledge Discovery and Data Mining (KDD), New York City, NY*, 1998.

[21] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management (SSDBM), Heidelberg, Germany*, 2010.

[22] N. Jardine and R. Sibson. The construction of hierarchic and non-hierarchic classifications. *The Computer Journal*, 11(2):177–184, 1968.

[23] K. Kailing, H.-P. Kriegel, and P. Kröger. Density-connected subspace clustering for high-dimensional data. In *Proceedings of the 4th SIAM International Conference on Data Mining (SDM), Lake Buena Vista, FL*, 2004.

[24] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1–58, 2009.

[25] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies. 1. hierarchical systems. *The Computer Journal*, 9(4):373–380, 1967.

[26] L. Lelis and J. Sander. Semi-supervised density-based clustering. In *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM), Miami, FL*, 2009.

[27] D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

[28] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematics, Statistics, and Probabilistics*, volume 1, pages 281–297, 1967.

[29] D. S. Moore and J. W. Yackel. Consistency properties of nearest neighbor density function estimates. *The Annals of Statistics*, 5(1):143–154, 1977.

[30] F. Murtagh. A survey of algorithms for contiguity-constrained clustering and related problems. *The Computer Journal*, 28(1):82–88, 1985.

[31] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[32] T. Pei, A. Jasra, D. J. Hand, A.-X. Zhu, and C. Zhou. DECODE: a new method for discovering clusters of different densities in spatial data. *Data Mining and Knowledge Discovery*, 18(3):337–369, 2009.

[33] R project. http://www.r-project.org/.

[34] D. M. Raup. Geometric analysis of shell coiling: General problems. *Journal of Paleontology*, 40(5):1178–1190, 1966.

[35] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[36] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998.

[37] J. Sander, X. Qin, Z. Lu, N. Niu, and A. Kovarsky. Automatic extraction of clusters from hierarchical clustering representations. In *Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Seoul, Korea*, pages 75–87, 2003.

[38] R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

[39] P. H. A. Sneath. The application of computers to taxonomy. *Journal of General Microbiology*, 17:201–226, 1957.

[40] W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20(1):25–47, 2003.

[41] W. T. Williams and J. M. Lambert. Multivariate methods in plant ecology: V. similarity analyses and information-analysis. *Journal of Ecology*, 54(2):427–445, 1966.

[42] D. Wishart. Mode analysis: A generalization of nearest neighbor which reduces chaining effects. In *Numerical Taxonomy*, 1969.

[43] M. A. Wong and T. Lane. A kth nearest neighbour clustering procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 45(3):362–368, 1983.

Cross-References