ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Recent developments in clustering algorithms

C. Bouveyron[1], B. Hammer [2], and T. Villmann[3]

1- University of Paris 1 - SAMM Laboratory, France
2- University of Bielefeld - CITEC centre of excellence, Germany
3- University of Applied Sciences - CI Group, Mittweida

Abstract. In this paper, we give a short review of recent developments in clustering. We shortly summarize important clustering paradigms before addressing important topics including metric adaptation in clustering, dealing with non-Euclidean data or large data sets, clustering evaluation, and learning theoretical foundations.

1 Introduction

Data clustering constitutes one of the most fundamental problems tackled in data mining as demonstrated by numerous algorithms, applications, and theoretical investigations covered in review articles or textbooks such as [43, 55, 38, 101, 98, 49]. Applications can be found in virtually every possible area such as bioinformatics, economics, robotics, or text and web mining. With electronic data sets becoming larger and larger, the need of clustering algorithms as a first step to make large data sets accessible is even constantly increasing.

The aim of clustering is often informally characterized as the task to decompose a given data set into subsets such that data are as similar as possible within clusters, while different clusters are separated. Apart from this informal description, however, there does not exist a single formalism, algorithm, or evaluation measure for clustering which is accepted as universally appropriate. The main reason behind this observation is given by the fact that clustering per se constitutes an ill-posed problem: the notion of what is a valid cluster and, in consequence, what is an appropriate clustering algorithm to detect such clusters depends on the application at hand. As such, eventually, clustering has to be designed for and evaluated in a given specific setting [40]. There cannot exist a universally suited algorithm or paradigm since the notion of a valid cluster can change from one application to the other and a cluster might be valid in one setting while it only accumulates noise in another.

Nevertheless, there exist many popular paradigms how to formalize clustering and how to derive a method thereof with a wide range of successful applications, reaching from popular K-means clustering to model based approaches or graph theoretic methods. In addition, many different evaluation measures of clustering results exist which are vital to quantitatively compare clustering results and to automatically optimize meta-parameters such as the number of clusters, see e.g. [4, 69]. More and more clustering techniques are dedicated to challenges which arise in modern data sets such as very high dimensional data, non-Euclidean settings, or very large data sets. In addition, quite interesting theoretical results have recently been developed which shed some light on the process of clustering from a statistical learning perspective, formalizing principles such as the generalization ability and axiomatic characterizations of clustering techniques.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

2 Clustering techniques

Clustering deals with the problem, given a set of n data points $\{x_1, \ldots, x_n\}$, to divide this set into K homogeneous groups. Often, the number of clusters K is thereby fixed by the user a priori. The definition of what means 'homogeneous' constitutes a key issue to turn this informal description into a concrete algorithm. One of the most popular clustering algorithm, K-means clustering, relies on the assumption that data are given by Euclidean vectors in $\mathbb{R}^p$. $d(x_i, x_j)$ refers to the dissimilarity given by the squared Euclidean metric in this vector space. For K-means, the clusters are represented by prototypes $w_1, \ldots, w_K$ in $\mathbb{R}^p$. Every prototype defines its receptive field $R(w_i) = \{x_j \mid \forall k\ d(w_i, x_j) \leq d(w_k, x_j)\}$. Clustering aims at a minimization of the quantization error

$$E_{\mathrm{QE}} = \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{n} 1_{x_j \in R(w_i)} d(w_i, x_j)\,.$$

There exist basically two different ways to optimize this cost function: an online approach, usually referred to as vector quantization; here a stochastic gradient descent is used, i.e. given a random data point, the respective closest prototype is adapted into the direction of the data point. Batch clustering leads to the classical K-means algorithm where, in turn, data are assigned to the closest prototypes and prototypes are relocated to the centre of gravity of their receptive field. Both algorithms converge to local optima of the cost function, provided the learning rate of online vector quantization is chosen appropriately. Often, K-means converges very fast, but it can take an exponential number of steps in the worst case even if only two-dimensional data are dealt with [89]. Convergence of K-means can also be proved if a general L_p metric is used instead of the standard Euclidean distance [83]. Note that clustering constitutes an NP hard problem if referring e.g. to the quantization error even in the plane, such that convergence to local optima is likely unavoidable [64]. Nevertheless, various heuristics of how to better initialize prototypes have been proposed [52].

Based on the basic quantization error, many extensions have been proposed. This concerns fuzzy variants of the algorithm, allowing a fuzzy membership degree of data points to the clusters [76]. Apart from a more general notion of a cluster, fuzzy K-means is less sensitive to prototype initialization. Alternative extensions take into account neighborhood relations of the prototypes, such as proposed in Neural Gas [66]. Besides better insensitivity to initialization, neural gas allows to extract further information corresponding to the topographic structure of the underlying data set. The self-organizing map (SOM) [55] integrates a priorly fixed neighborhood structure, a regular low-dimensional lattice, in a similar way. This way, additional functionality in the form of a data visualization is offered, but, due to topological restrictions, usually more prototypes are needed. Therefore, SOM is usually used as a first step only and it is combined with a merging of prototypes afterwards [90]. As a side effect, the final number of clusters is determined by the model itself. A similar effect is reached by growing techniques which adapt the necessary number K of clusters while training, this way also improving the insensitivity to prototype initialization [60].

K-means clustering usually severely suffers from initialization sensitivity which is partly overcome by continuous extensions to neighborhood incorporation or

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

fuzzy labeling. An alternative is offered by techniques which do not change the cost function itself, rather they use continuous optimization techniques to optimize the discrete quantization error. A fundamental approach has been proposed in [80] where the principle of deterministic annealing is transferred to the range of clustering. An alternative way is taken in [31]: Affinity propagation treats the quantization error as a factor graph for which an approximate solution can be obtained by means of the max-sum algorithm. As side effect of this formalization is the observation that the number of clusters need not be specified a priori, rather it is determined by the self-similarities of data points which correspond to direct costs for the found clusters. An even more advanced approach based on deep insights into statistical physics is offered by super paramagnetic clustering as proposed in [10]. Given a clustering objective, metaheuristics such as genetic algorithms or swarm intelligence can be used as a powerful but possible time consuming alternative to turn these costs into a clustering algorithm [68].

A fundamentally different cost function forms the base of spectral techniques. Here, data points are identified with vertices in a graph; possibly non-Euclidean pairwise similarities (rather than dissimilarities) $s(x_i, x_j)$ give rise to weighted edges in the graph. A clustering corresponds to a decomposition of the data $\{x_1, \ldots, x_n\}$ into clusters $I_1, \ldots, I_k$ such that as few vertices as possible are cut by means of this decomposition. The so-called graph cut measures the costs

$$E_{\mathrm{CUT}} = \sum_{i=1}^{K} \sum_{x_j \in I_i, x_k \notin I_i} s(x_j, x_k)\,.$$

It turns out that this simple cut is usually not a desirable objective, because it favors highly unbalanced clusters. Therefore, the so-called normalized or ratio cut are taken which normalize every summand by its size or its accumulated vertex degree, respectively. Similar to the optimization of the quantization error, optimization of the normalized or ratio cut is NP-hard, such that approximations are used. Spectral clustering is based on the observation that the cut can equivalently be formulated as an optimization problem of an algebraic form induced by the graph Laplacian over the set of vectors with entries 0 and 1. A continuous relaxation of this problem can be solved by the eigenvectors of the graph Laplacian. To turn this relaxation into a crisp clustering assignment, simple k-means is usually used to decompose the points induced by the eigenvector components. An overview of spectral clustering as well as its convergence properties can be found in [91, 93]. Often, the similarities s of the cost function are based on local neighborhood graphs; consistency of this construction can be proved under certain conditions [65]. Clustering techniques such as [61] rely on the same objective as spectral clustering but they compute the eigenvectors by means of different techniques e.g. based on a von Mises iteration.

An alternative principled formalization of clustering is offered by generative or model-based clustering, which has been widely studied by [30, 70], for example. In its simplest form, data are modeled as a mixture of Gaussians. Assuming that $\{x_1, ..., x_n\}$ are independent identically distributed realizations of a random vector, the mixture model density is given as

$$p(x) = \sum_{k=1}^{K} \pi_k f(x; \theta_k),$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

where $f(.;\theta_k)$ is the multivariate Gaussian density $\phi(.;\mu_k,\Sigma_k)$ parametrized by a mean vector μ_k and a covariance matrix Σ_k for the kth component and π_k is the class prior. This objective can be optimized by a standard expectation maximization (EM) algorithm. Once the model parameters are estimated, the maximum a posteriori (MAP) rule provides the partition of the data into K groups. Interestingly, in the limit of small bandwidth, the classical K-means algorithm is recovered this way. Extensions of this model change the probabilities according to the given setting, e.g. substituting Gaussians by binomial distributions for a latent trade model or integrating neighborhood cooperation.

Besides clustering models based on cost function, a variety of popular iterative clustering schemes exists. DBSCAN [82] constitutes a very efficient algorithm which is particularly suited for large data sets and priory unknown cluster shapes,. It relies on an iterative enlargement of clusters based on an approximation of the underlying data density. The approach [38] formalizes cluster costs in an information theoretic way using the Renyi entropy, and iteratively optimizes cluster assignments based on this notion. Popular hierarchical clustering techniques such as linkage methods arrive at a hierarchy of clusters by an iterative greedy combinations of clusters, starting at the given points [43]. Under strict assumptions on the metric (such as an additive or ultrametric), they provide the correct result, but they are very sensitive to noise in the general setting.

Note that quite a variety of standard clustering algorithms has been extended to yield hierarchical results, including spectral clustering or affinity propagation [37, 21]. A hierarchical graph clustering scheme is also proposed in this volume [25]. Further variations on the basic clustering objective aim at a more advanced scheme, such as co-clustering, i.e. the simultaneous clustering of objects and object features [22], outlier detection which is essentially a one-class clustering problem [15], or clustering with simultaneous dimensionality reduction [26].

3 Metric adaptation and variable selection

All clustering techniques severely depend on the given metric used to compare pairs of data. In many settings, the dissimilarity measure is given by the standard Euclidean distance, which encounters problems in particular for very high dimensional data. In consequence, many approaches address the metric used to compare the data and try to adapt it such that the resulting clustering structure is more pronounced. Recently, several authors have been interested to simultaneously cluster data and reduce their dimensionality by selecting relevant variables for the clustering task. The clustering task aims therefore to group the data on a subset of relevant features, resulting in an improved quality and interpretability. A recent overview about different feature selection techniques and a connection to supervised feature selection can be found in [62].

For model based clustering, variable selection can be tackled within a Bayesian framework [78, 67]. Maugis et al. [67], consider three kinds of subsets of variables: relevant variables, irrelevant variables which can be explained by a linear regression from relevant variables and finally irrelevant variables which are useless for the clustering. The models in competition are selected using the BIC criterion. In the Gaussian mixture model context, Pan and Shen [77] introduced a penalty term in the log-likelihood function in order to yield sparsity in the features and to select relevant variables. Witten and Tibshirani [96] proposed a general non-

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

probabilistic framework for the variable selection problem which is based on a general penalized criterion and governs variable selection and clustering.

However, such approaches have two limitations: they remove variables which are possibly discriminative for the clustering and they are time-consuming. Extensions simply adapt the metric to a more appropriate form, whereby a more general view can be taken than the deletion of input variables. Particularly useful alternatives are offered by a weighted Euclidean metric or a general quadratic form. Thereby, metric parameters are adapted automatically based on the given setting at hand. Relevance learning, i.e. the adaptation of diagonal terms according to the relevance of the considered dimensions, has been introduced in [46]. In the contribution [36] in this volume, relevance parameters are adapted according to auxiliary supervised information, resulting in a powerful data representation scheme adapted to the given semi-supervised setting.

The adaptation of full matrices has been mathematically investigated in [3], among others, in the context of K-means and extensions. It turns out that local principal components are detected this way. Thus, the result resembles the result of subspace clustering techniques which explicitly project the given data locally onto meaningful low dimensional subspaces. In the context of model-based clustering, early strategies [81] are based on the factor analysis model which assumes that the latent space is related to the observation space through a linear relationship. This model was recently extended in [71] and yields in particular the well known mixture of probabilistic principal component analyzers [87]. Recent work [13, 72] proposed two families of parsimonious and regularized Gaussian models which partially encompass previous approaches. These techniques are very efficient for high-dimensional data. Despite this fact, these probabilistic methods mainly aim at clustering, neglecting visualization or interpretability.

4 General metrics and non-Euclidean settings

Besides adaptive Euclidean metrics to enhance the representation capacity of clustering algorithms in Euclidean space, more and more algorithms deal with more general non-Euclidean data such as discrete sequences or tree structures, time series data, or complex data for which a dedicated dissimilarity measure is appropriate. Some techniques extend given clustering formalisms to dedicated data formats, such as extensions to time series data by means of recursive processing [42], algorithms for chains of data [88], or algorithms for categorical data [47]. We refer for instance to [95] for a survey on time series clustering.

Another very important data format concerns functional data which display an infinite dimensionality in the ideal setting. Non-parametric approaches to functional clustering, as for instance [27, 85], lead to powerful clustering algorithms. In contrast, model-based clustering techniques for functional data have interesting interpretability properties. Unlike in the case of finite dimensional data vectors, model-based methods for clustering functional data are not directly available since the notion of probability density function generally does not exist for such data [20]. Consequently, the use of model-based clustering methods on functional data consists usually in first transforming the infinite dimensional problem into a finite dimensional one and then in using a model-based clustering method designed for finite dimensional data. The representation of functions in a finite space can be carried out by FPCA [79], projection on a basis of natural

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

cubic splines [94] or using ARMA or GARCH [32] models. Recently, two new approaches [50, 14] allow the interaction between the discretization and the clustering steps by introducing a stochastic model for the basis coefficients. Another method [48] of this type is proposed in this volume which approximates the density of functional random variables using the functional principal components.

A very general interface to complex data is offered by the notion of similarity or dissimilarity. Techniques such as affinity propagation, spectral clustering, or agglomerative methods can directly deal with such settings, whereas prototype based methods or model based approaches encounter difficulties, unless prototype positions are restricted to the given data points as proposed e.g. in [18]. A very general approach which implicitly refers to Euclidean data is offered by kernelization, as proposed e.g. in [99, 12]. More general dissimilarities can be treated by means of an implicit embedding of data into pseudo-Euclidean space, where many standard techniques such as K-means, fuzzy clustering, and neural gas can be applied. See e.g. [44, 41] for such algorithms and [41] for a mathematical treatment of the convergence properties in such spaces. Often, instead of the quantization error, its relational dual is considered, which is equivalent to the quantization error if prototypes are located in cluster centers, as shown in [41]:

$$E_{\mathrm{dual}} = \frac{1}{4} \sum_{i=1}^{K} \frac{1}{|I_k|} \sum_{x_j \in I_k, x_{j'} \in I_k} d(x_j, x_{j'})$$

I_k refers to the decomposition into clusters, and d to the given dissimilarity of data points. Similar to the standard quantization error, its dual can be optimized based on deterministic annealing techniques [45]. Since there does not yet exist a universally accepted notion of a probability measure for pseudo-Euclidean space, however, it is not clear in which sense these algorithms yield meaningful clusters.

A very interesting model for clustering relational data together with convergence results has been proposed in [11]. The clustering method is capable of recovering block structures in a relational data set provided blocks are connected with low probability, but within blocks connection probabilities are higher. This constitutes an interesting step towards an exact probabilistic modeling of what it means that clustering relational data converges to a true underlying clustering.

5 Large data sets

Due to its wide usage, optimization techniques to make K-means faster are common such as dedicated data structures which allow a fast neighborhood determination [51]. With data sets becoming larger and larger, there is a need for instantaneous or streaming algorithms which can deal with large data sets in at most linear time and constant memory. For classical K-means, a couple of efficient streaming methods have been proposed often accompanied by formal guarantees [84, 39]. Alternatives with linear or even sub-linear effort can rely on geometrical concepts known as core methods [5] or subsampling [6]. In addition, many iterative methods have been directly designed for very large data sets such as CLARANS, STING, or BIRCH [74, 2, 39].

The problem of large data sets is particularly pronounced if relational data or kernel methods are dealt with, since the similarity or dissimilarity matrix is already of quadratic size. Approximations are required to make the methods

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

feasible for large data sets. Popular techniques can be found around two approximation schemes: the Nyström approximation for kernel methods approximates the dissimilarities by a low rank matrix; it has been introduced into spectral clustering, for example [29]. Efficient decomposition techniques such as patch processing have been proposed in [41]. In this volume, novel heuristics to speed up relational clustering schemes are introduced in [17].

Often, parallelization schemes offer further speed-up, such as discussed in [41] for patch processing. Cloud computing offers new opportunities as discussed in [28]. However, a naive parallelization is not always successful, as demonstrated in this session in the contribution [24] for the classical vector quantization.

6 Evaluation of clustering

As mentioned above, clustering is an ill-posed task. Hence an optimal cluster partition as well as an optimum cluster number per se is not well defined without further constraints. Nevertheless, there exist several approaches to evaluate and compare the quality of a given cluster solution reflecting commonly accepted aspects of clustering like separation and compactness of the clusters. Famous measures are the Davies-Bouldin-index or the Xie-Beni-measure both relating the within-cluster variances (compactness) to the inter-cluster variances (separation) [19, 97]. Similar approaches which emphasize aspects like overlapping clusters or different cluster shapes are presented in [23, 53, 58, 75].

Another group of evaluation measures is based on information theoretic concepts. These approaches mainly judge the validity of a given cluster solutions rather than a comparison of clusterings. Approaches relate to different kinds of partition entropies most of them derived from Bezdek's original proposal [8]. The underlying assumption is that an information optimum coding of data constitutes a basic principle of clustering. Several measures are applicable to fuzzy clustering [9, 97]. An extension which takes into account a dual step of topographic vector quantization and subsequent clustering has been proposed in [86]. A fuzzy counterpart of this idea is presented in [33] in this volume.

Another aspect of cluster evaluation is the determination of the appropriate number K of clusters. Frequently, the above indices are also used to determine K assuming that a good choice leads to a better evaluation measure. Alternative ways to determine K inlcude the eigengap heuristics or spectral clustering [63] or kernel based clustering [35]. In affinity propagation, the number of clusters is controlled by self-similarities [31]. An estimator of the correct number K based on a stability analysis of the cluster solutions is presented in [92].

A third class of evaluation measures relates to indices designed for cluster comparison. Popular examples are variants of the Jaccard index, originally developed for the comparison of classifications but nowadays extended for clustering comparisons [100]. Fuzzy variants based on t-norms are presented in [34]. Other measures are based on extensions of Mallows distances [102] or they are derived from lattice approaches [73]. Yet, these approaches can only reflect several aspects of cluster solutions. The final decision about an optimum task specific choice eventually relies on the user.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

7 Learning theory of clustering and an axiomatic view

Starting with the popular approach of Kleinberg [54], there has been effort to treat clustering algorithms in an axiomatic way and to characterize techniques based on their axiomatic properties. Essentially, Kleinberg formalizes the following three axioms of clustering: scale invariance of the clustering algorithm, richness that means its ability to reach all possible clusterings, and consistency which refers to an invariance of results if within cluster distances are reduced and between cluster distances are enlarged. Interestingly, it can be shown that these three axioms cannot be fulfilled simultaneously. Recently, these axioms have been further refined into interesting properties of clustering algorithms in [7, 1]. Here, different properties of algorithms are specified including invariance properties such as scale invariance, isomorphism invariance, or order variance; consistency properties; and range properties such as the capability of reaching every desired clustering. In the contributions [7, 1, 16] it is analyzed in how far popular clustering algorithms including single and complete linkage or K-means clustering possess these properties. As a result, a catalogue of important characteristics of popular clustering algorithms is compiled which can help to select an appropriate clustering scheme in a given setting.

Unlike for supervised machine learning, the question in how far clustering techniques allow valid learning and generalization based on a finite set of given data has long been an open question. A major problem consists in the fact that it is not priorly clear what is the meaning of the generalization error of a clustering and in how far does the notion of convergence provided the number of points gets larger and larger make sense: since clustering refers to a discrete composition of a set of data, there is no generic notion of convergence due to a lack of a common vector space for different size data sets and clustering decompositions in the general setting.

The situation is much clearer if a clustering algorithm which naturally generalizes to an embedding vector space is considered. The classical quantization error, for example, can directly be extended to a continuous error

$$E_{\mathrm{QE}} = \frac{1}{2} \sum_{i=1}^{K} \int d(w_i, x) 1_{x \in R(w_i)} P(x) dx$$

where P is an underlying probability distribution of the given data space. It has been shown in [6] that, provided data are i.i.d. a small empirical quantization error on a given training set also implies a small quantization error for the underlying data distribution in this setting, hence the generalization ability of K-means clustering can be guaranteed in this sense. These arguments can be extended to clustering schemes provided that the clustering scheme can be characterized by a local compression scheme (such as a prototype for K-means), and that the expectation of the considered loss function can be computed by means of a simple function which depends on this compressor only (such as the quantization error for K-means). For clusterings characterized by general sets, these properties are usually not fulfilled.

Alternative notions of how to formalize the generalization ability and consistency of clustering schemes have been proposed for different scenarios. Consistency of spectral clustering relies on the observation that the normalized graph Laplacian converges to an operator on the space of continuous functions. This

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

implies a convergence of the corresponding eigenvectors and eigenvalues, such that a clustering induced by these eigenvectors has high probability to retrieve the original clusters in the data space. For nearest neighbor based approaches, consistency results have been investigated in [57] concerning hierarchical tree structures derived from this information. This approach also proposes a provably correct algorithm to prune such trees such that the result corresponds to the cluster tree as induced by the level sets of the underlying probability distribution. These theoretical developments constitute very promising steps towards a learning theory of unsupervised clustering.

8 Conclusions

We presented a short review about modern trends in clustering including challenging topics like non-Euclidean data, kernel clustering and metric adaptation. Obviously, this short review can not be complete, referring to the cited literature for further reading.

References

[1] M. Ackerman, S. Ben-David, and D. Loker. Towards property-based classification of clustering paradigms. In Lafferty et al. [59], pages 10–18.
[2] C. C. Aggarwal. A framework for clustering massive-domain data streams. In *ICDE*, pages 102–113, 2009.
[3] B. Arnonkijpanich, A. Hasenfuss, and B. Hammer. Local matrix adaptation in topographic neural maps. *Neurocomputing*, 74(4):522–539, 2011.
[4] P. Awasthi and R. B. Zadeh. Supervised clustering. In Lafferty et al. [59], pages 91–99.
[5] M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *STOC*, pages 250–257, 2002.
[6] S. Ben-david. A framework for statistical clustering with a constant time approximation algorithms for k-median clustering. In *In COLT*, pages 415–426. Springer, 2004.
[7] S. Ben-David and M. Ackerman. Measures of clustering quality: A working set of axioms for clustering. In Koller et al. [56], pages 121–128.
[8] J. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3:58–73, 1974.
[9] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York, 1981.
[10] M. Blatt, S. Wiseman, and E. Domany. Superparamagnetic clustering of data. *Phys Rev Lett*, 76(18):3251–3254, 1996.
[11] R. Boulet. Disjoint unions of complete graphs characterized by their laplacian spectrum. *The Electronic journal of Linear Algebra*, 18:773–783, 2009.
[12] R. Boulet, B. Jouve, F. Rossi, and N. Villa. Batch kernel som and related laplacian methods for social network analysis. *Neurocomputing*, 71(7-9):1257–1273, 2008.
[13] C. Bouveyron, S. Girard, and C. Schmid. High-Dimensional Data Clustering. *Computational Statistics and Data Analysis*, 52(1):502–519, 2007.
[14] C. Bouveyron and J. Jacques. Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5(4):281–300, 2011.
[15] F. Camastra and A. Verri. A novel kernel method for clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):801–805, 2005.
[16] G. Carlsson and F. Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *J. Mach. Learn. Res.*, 11:1425–1470, August 2010.
[17] B. Conan-Guez and F. Rossi. Dissimilarity clustering by hierarchical multi-level refinement. In *ESANN'12*, 2012.
[18] M. Cottrell, B. Hammer, A. Hasenfuss, and T. Villmann. Batch and median neural gas. *Neural Networks*, 19:762–771, 2006.
[19] D. Davies and D. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.
[20] A. Delaigle and P. Hall. Defining pobability density for a distribution of random functions. *The Annals of Statistics*, 38:1171–1193, 2010.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

[21] I. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, 2007.
[22] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pages 89–98, 2003.
[23] J. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):1974, 95–104.
[24] M. Durut, B. Patra, and F. Rossi. A discussion on parallelization schemes for stochastic vector quantization algorithms. In *ESANN'12*, 2012.
[25] M. K. el Mahrsi and F. Rossi. Modularity-based clustering for network-constrained trajectories. In *ESANN'12*, 2012.
[26] E. Elhamifar and R. Vidal. Sparse manifold clustering and embedding. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 55–63. 2011.
[27] F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006.
[28] R. L. Ferreira Cordeiro, C. Traina, Junior, A. J. Machado Traina, J. López, U. Kang, and C. Faloutsos. Clustering very large multi-dimensional datasets with mapreduce. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 690–698, New York, NY, USA, 2011. ACM.
[29] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.
[30] C. Fraley and A. Raftery. Model-based clustering, discriminant analysis and density estimation. *J. Amer. Statist. Assoc.*, 97:611–631, 2002.
[31] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.
[32] S. Frühwirth-Schnatter and S. Kaufmann. Model-based clustering of multiple time series. *Journal of Business and Economic Statistics*, 26:78–89, 2008.
[33] T. Geweniger, M. Kaestner, M. Lamge, and T. Villmann. Modified conn-index for the evaluation of fuzzy clusterings. In *ESANN'12*, 2012.
[34] T. Geweniger, D. Zühlke, B. Hammer, and T. Villmann. Median fuzzy c-means for clustering dissimilarity data. *Neurocomputing*, 73(7–9):1109–1116, 2010.
[35] M. Girolami. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, 13(3):780–784, 2002.
[36] A. Gisbrecht, D. Sovilj, B. Hammer, and A. Lendasse. Relevance learning for time series inspection. In *ESANN'12*, 2012.
[37] I. Givoni, C. Chung, and B. Frey. Hierarchical affinity propagation. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 238–246, Corvallis, Oregon, 2011. AUAI Press.
[38] E. Gokcay and J. Principe. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):158–171, 2002.
[39] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):515–528, 2003.
[40] I. Guyon, U. V. Luxburg, and R. C. Williamson. Clustering: Science or art. In *NIPS 2009 Workshop on Clustering Theory*, 2009.
[41] B. Hammer and A. Hasenfuss. Topographic mapping of large dissimilarity datasets. *Neural Computation*, 22(9):2229–2284, 2010.
[42] B. Hammer, A. Micheli, A. Sperduti, and M. Strickert. A general framework for unsupervised processing of structured data. *Neurocomputing*, 57:3–35, 2004.
[43] J. A. Hartigan. *Clustering algorithms*. Wiley New York, 1975.
[44] R. Hathaway and J. Bezdek. Nerf c-means: Non-euclidean relational fuzzy clustering. *Pattern Recognition*, 27(3):429–437, 1994.
[45] T. Hofmann and J. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):1–14, 1997. cited By (since 1996) 179.
[46] J. Huang, M. Ng, H. Rong, and Z. Li. Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668, 2005.
[47] Z. Huang and M. Ng. A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, 7(4):446–452, 1999.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

[48] J. Jacques and C. Preda. Curves clustering with approximation of the density of functional random variables. In *ESANN*, 2012.
[49] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.
[50] G. James and C. Sugar. Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.*, 98(462):397–408, 2003.
[51] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. An efficient k-means clustering algorithms: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892, 2002.
[52] S. Khan and A. Ahmad. Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, 25(11):1293–1302, 2004.
[53] D.-W. Kim, K. Lee, and D. Lee. On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition*, 37(10):2009–2025, 2004.
[54] J. M. Kleinberg. An impossibility theorem for clustering. In *NIPS*, pages 446–453, 2002.
[55] T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001.
[56] D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors. *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*. Curran Associates, Inc., 2009.
[57] S. Kpotufe and U. von Luxburg. Pruning nearest neighbor cluster trees. In L. Getoor and T. Scheffer, editors, *ICML*, pages 225–232. Omnipress, 2011.
[58] S. Kwon. Cluster validity index for fuzzy clustering. *Electronics Letters*, 34(22):2176 – 2177, 1998.
[59] J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*. Curran Associates, Inc., 2010.
[60] A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2002.
[61] F. Lin and W. W. Cohen. Power iteration clustering. In J. Fürnkranz and T. Joachims, editors, *ICML*, pages 655–662. Omnipress, 2010.
[62] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, 2005.
[63] U. V. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
[64] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is np-hard. In *Lecture Notes in Computer Science 5431*, pages 274–285. Springer, 2009.
[65] M. Maier, U. von Luxburg, and M. Hein. Influence of graph construction on graph-based clustering measures. In Koller et al. [56], pages 1025–1032.
[66] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, 1993.
[67] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection for Clustering with Gaussian Mixture Models. *Biometrics*, 65(3):701–709, 2009.
[68] U. Maulik and S. Bandyopadhyay. Genetic algorithm-based clustering technique. *Pattern Recognition*, 33(9):1455–1465, 2000.
[69] U. Maulik and S. Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.
[70] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Interscience, New York, 2000.
[71] G. McLachlan, D. Peel, and R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, (41):379–388, 2003.
[72] P. McNicholas and B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008.
[73] M. Meila. Comparing clusterings: an axiomatic view. In L. D. Raedt and S. Wrobel, editors, *ICML*, volume 119 of *ACM International Conference Proceeding Series*, pages 577–584. ACM, 2005.
[74] R. Ng and J. Han. Clarans: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14(5):1003–1016, 2002.
[75] M. Pakhira, S. Bandyopadhyay, and U. Maulik. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(3):487–501, 2004.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

[76] N. Pal, K. Pal, J. Keller, and J. Bezdek. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 13(4):517–530, 2005.
[77] W. Pan and X. Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8:1145–1164, 2007.
[78] A. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.
[79] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.
[80] K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9):589–594, 1990.
[81] D. Rubin and D. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.
[82] 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*, 2:169–194, 1998.
[83] S. Z. Selim and M. Ismail. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(1):81–87, 1984.
[84] M. Shindler, A. Wong, and A. W. Meyerson. Fast and accurate k-means for large datasets. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2375–2383. 2011.
[85] T. Tarpey and K. Kinateder. Clustering functional data. *J. Classification*, 20(1):93–114, 2003.
[86] K. Tasdemir and E. Merényi. A validity index for prototype-based clustering of data sets with complex cluster structures. *IEEE Transactions on Systems, Man, and Cybernetics*, 41(4):1039 – –1053, 2011.
[87] E. Tipping and C. Bishop. Mixtures of Probabilistic Principal Component Analysers. *Neural Computation*, 11(2):443–482, 1999.
[88] A. Ukkonen. Clustering algorithms for chains. *Journal of Machine Learning Research*, 12:1389–1423, 2011.
[89] A. Vattani. k-means requires exponentially many iterations even in the plane. *Discrete and Computational Geometry*, 45(4):596–616, 2011.
[90] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, 2000.
[91] U. von Luxburg. A tutorial on spectral clustering. Technical Report 149, Max Planck Institute for Biological Cybernetics, Aug. 2006.
[92] U. von Luxburg. Clustering Stability: An Overview. *Foundations and Trends in Machine Learning*, 2(3):235–274, July 2010.
[93] U. von Luxburg, O. Bousquet, and M. Belkin. Limits of spectral clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 857–864. MIT Press, Cambridge, MA, 2005.
[94] G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.
[95] T. Warren Liao. Clustering of time series data – a survey. *Pattern Recognition*, 38:1857–1874, 2005.
[96] D. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 150(490):713–726, 2010.
[97] X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991.
[98] R. Xu and D. Wunsch II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.
[99] H. Yin. On the equivalence between kernel self-organising maps and self-organising mixture density networks. *Neural Networks*, 19(6-7):780–784, 2006.
[100] K. Zalik. Validity index for clusters of different sizes and densities. *Pattern Recognition Letters*, 32:221–2234, 2011.
[101] S. Zhong and J. Ghosh. A unified framework for model-based clustering. *JMLR*, 4:1001–1037, 2003.
[102] D. Zhou, J. Li, and H. Zha. A new mallows distance based metric for comparing clusterings. In *Proceedings of the 22nd international conference on Machine learning*, pages 1028–1035. ACM, 2005.