Clustering Principles and Empirical Risk Approximation

Joachim M. Buhmann

Rheinische Friedrich-Wilhelms-Universität, Institut f. Informatik III, 53117 Bonn
(e-mail: jb@informatik.uni-bonn.de)

Abstract. Data Clustering is one of the fundamental techniques in pattern recognition to extract structure from data. We discuss a maximum entropy approach to clustering for different clustering cost functions and relate it to the estimation principle of *Empirical Risk Approximation*. Large deviation techniques from statistical learning theory provide guarantees for the stability of clustering solutions.

1 Introduction

Intelligent data analysis extracts symbolic information and relations between objects from quantitative or qualitative data. A prominent class of methods are clustering or grouping principles which are designed to discover and extract structures hidden in data sets (Jain and Dubes (1988)). The parameters which represent the clusters are estimated on the basis of quality criteria or cost functions. Clustering as a fundamental pattern recognition problem can be characterized by the following four design steps: (i) *Data Representation*, (ii) *Cluster Modeling*, (iii) *Cluster Optimization/Estimation* , (iv) *Cluster Validation*. It is important to note that the data representation predetermines what kind of cluster structures can be discovered in the data. Vectorial data, proximity or similarity data and histogram data are three examples of a wide variety of data types which are analyzed in the clustering literature.

Most clustering algorithms are based on the quality criteria of compactness or connectedness. Compactness stresses the partition of objects into subsets such that the mutual similarity is maximized. Connectedness refers to the graph-theoretical approach to clustering where a pair of objects are assigned to the same cluster if they can be connected by a chain of mediating objects with high similarity. We will give precise meaning to these two criteria in Sect 2. The search for cluster structure in data and its validation is discussed in Sects.3 and 4, respectively.

2 Modeling of Cluster Structure

Various data types have been introduced in the pattern recognition literature. Mathematically, a datum is defined as a relation between a design space $\mathfrak{O}$ and a measurement space $\mathfrak{F}$. The pair $(\mathbf{o}, \mathbf{x}) \in \mathfrak{O} \times \mathfrak{F}$ of an object configuration

$\mathbf{o} \in \mathfrak{O}$ and a measurement $\mathbf{x} \in \mathfrak{F}$ can represent a functional dependency $\mathbf{x} : \mathbf{o} \mapsto \mathbf{x}(\mathbf{o})$ between objects and measurements or a stochastic dependency $\mathbf{P}\{\mathbf{x}|\mathbf{o}\}$. This categorization yields the following data types which are most common in data analysis problems:

Vectorial data characterize an object $\mathbf{o}$ by a number of attributes which are combined to a d-dimensional feature vector $\mathbf{x}(\mathbf{o}) \in \mathfrak{F} \subset \mathbb{R}^d$.

Distributional data of an object $\mathbf{o}$ are described by an empirical probability distribution or histogram $\mathbf{P}\{\mathbf{x}|\mathbf{o}\}$ over features $\mathbf{x} \in \mathfrak{F}$.

Proximity data are characterized by pairwise comparisons between objects according to a proximity measure, e.g., $\mathbf{x}(\mathbf{o}_i) := \{\mathcal{D}(\mathbf{o}_i, \mathbf{o}_j) \in \mathbb{R} : 1 \leq j \leq n\}$.

Various polyadic data types like co-occurrence data (word bigrams in linguistics, consumer behavior data in economics, ...) or even more complex data types (trigrams) are occasionally considered in the empirical sciences.

The goal of data clustering is to determine a partitioning of object space $\mathfrak{O}$ into subsets $\mathfrak{G}_\alpha := \{\mathbf{o} \in \mathfrak{O} : m(\mathbf{o}) = \alpha\},\ 1 \leq \alpha \leq k$. Mathematically, assignments of objects to clusters are represented by an assignment function $m : \mathfrak{O} \to \{1, 2, \ldots, k\},\ \mathbf{o} \mapsto m(\mathbf{o})$. The space of all clustering solutions is the set of all assignment functions $\mathfrak{M} = \{m : \mathfrak{O} \to \{1, 2, \ldots, k\}\}$. The quality of these partitions are evaluated by an appropriate homogeneity measure for the respective data type. The most commonly used clustering costs are invariant under permutations of the cluster indices. Hierarchical and topological clustering methods impose additional structure on the partitions.

Central Clustering or Vector Quantization: Clustering objects which are represented as vectorial data amounts to partition the feature space $\mathfrak{F} \subset \mathbb{R}^d$. A set of n objects is represented by a set of data points $\mathfrak{X} := \{\mathbf{x}_i \in \mathfrak{F} : 1 \leq i \leq n\}$. The set of objects is partitioned into clusters in such a way that the average distance of data points to their cluster centers $\mathcal{Y} = \{\mathbf{y}_\nu \in \mathfrak{F} : 1 \leq \nu \leq k\}$ is minimized. The representation of data $\mathbf{x}_i$ by the centroid $\mathbf{y}_{m(i)}$ induces distortion/quantization costs $\mathcal{D}_{i,m(i)}$ due to information loss. The functional form of $\mathcal{D}_{i,m(i)}$ depends on the weighting of data distortions, in which quadratic costs $\mathcal{D}_{i,m(i)} = \left|\mathbf{x}_i - \mathbf{y}_{m(i)}\right|^2$ and k means $\mathbf{y}_\alpha = \sum_{\mathbf{o}_i \in \mathfrak{G}_\alpha} \mathbf{x}_i / |\mathfrak{G}_\alpha|$ are the most common choices. More general distortion measures like l_p-norms are occasionally considered. The cost function for k-means clustering is defined as

$$\mathcal{H}^{\mathrm{cc}}(m, \mathcal{Y}; \mathfrak{X}) = \sum_{i \leq n} \mathcal{D}_{i,m(i)} = \sum_{i \leq n} \left|\mathbf{x}_i - \mathbf{y}_{m(i)}\right|^2 . \tag{1}$$

The size k of the cluster set, i.e., the complexity of the clustering solution, has to be determined a priori or, as discussed by Buhmann and Kühnel (1993), by a problem-dependent complexity measure. A minimum of the cost function (1) can be found by varying the assignments $m(i)$ which effectively is a search in a discrete space with exponentially many states. The optimization procedure implicitly yields the cluster means $\{\mathbf{y}_\nu\}$ by estimating optimized assignments $\{m(i)\}$.

Distributional Clustering: Distributional data represent the co-occurrence of objects and features by histograms (Tishby et al. (1999)). Denote by $\mathfrak{O} \times \mathfrak{F}$ the data space, i.e., the product space of objects $\mathbf{o}_i \in \mathfrak{O}, 1 \le i \le n$ and features $\mathbf{x}_j \in \mathfrak{F}, 1 \le j \le f$. In information retrieval, objects might be documents and features might be keywords. The $\mathbf{o}_i \in \mathfrak{O}$ are characterized by the set of n observations $\mathfrak{Z} = \big\{(\mathbf{o}_{i(r)}, \mathbf{x}_{j(r)}) : 1 \le r \le l\big\} \subseteqq (\mathfrak{O} \times \mathfrak{F})^l$. The sufficient statistics of how often the object–feature pair $(\mathbf{o}_i, \mathbf{x}_j)$ occurs in $\mathfrak{Z}$ is measured by the frequencies $\{n_{ij}$: number of observations $(\mathbf{o}_i, \mathbf{x}_j)/l\}$.

These distributional data can be generated/explained by a mixture of data sources: (i) select an object $\mathbf{o}_i \in \mathfrak{O}$ with probability n_i; (ii) choose the cluster ν according to the cluster membership of $\nu = m(i)$; (iii) select $\mathbf{x}_j \in \mathfrak{F}$ according to the class–conditional distribution $q_{j|\nu}$. The different values ν of the data assignments denote the mixture components. The negative log-likelihood of the data yields the clustering cost function

$$\mathcal{H}^{\mathrm{hc}}(m, q; \mathfrak{Z}) = \sum_{i \le n} \mathcal{D}_{i,m(i)} = -\sum_{i \le n} \sum_{j \le f} n_{ij} \log q_{j|m(i)} \qquad (2)$$

The negative logarithm of the cluster probabilities $q_{j|m(i)}$ weighted with the frequency n_{ij} is the loss $\mathcal{D}_{i,m(i)}$ of observing $(\mathbf{o}_i, \mathbf{x}_j)$. Tishby et al. (1999) have suggested an insightful information theoretic interpretation of 2 which stresses the importance of context information.

Pairwise Clustering: Clustering non-metric data which are characterized by proximity information and not by explicit Euclidean coordinates can be formulated as a graph optimization problem. Given is a graph $(\mathcal{V}, \mathcal{E})$ with weights $\mathfrak{D} := \{\mathcal{D}_{ij}\}$ on the edges (i, j). The vertices denote the objects to be grouped and the edge weights encode dissimilarity information. Compact clusters are represented by a partition of the vertex set with small dissimilarities between all objects which belong to the same cluster. To simplify the notation, the subset of edges with both vertices in cluster α is denoted by $\mathcal{E}_\alpha = \{(i, j) \in \mathcal{E} : \mathbf{o}_i, \mathbf{o}_j \in \mathfrak{G}_\alpha\}$. A meaningful cost function for pairwise clustering which primarily avoids grouping dissimilar objects into one cluster is defined by

$$\mathcal{H}^{\mathrm{pc}}(m; \mathfrak{D}) = \sum_{i \le n} \mathcal{D}_{i,m(i)} = \sum_{i \le n} \frac{\left|\mathfrak{G}_{m(i)}\right|}{\left|\mathcal{E}_{m(i)}\right|} \sum_{j:(i,j) \in \mathcal{E}_{m(i)}} \mathcal{D}_{ij}. \qquad (3)$$

Preferred clusters according to this cost function are those subsets of objects with minimal average intra cluster dissimilarities, weighted by the cluster size $|\mathfrak{G}_\nu|$. This cost function has the remarkable and for applications extremely valuable invariance that the assignments do not change if all dissimilarities are shifted by the same offset $\mathcal{D}_0$, i.e., $\mathcal{D}_{ij} \to \mathcal{D}_{ij} + \mathcal{D}_0$. $\mathcal{H}^{\mathrm{pc}}(m; \mathfrak{D})$ is identical to the k-means clustering criterion for Euclidean distances $\mathcal{D}_{ij} = |\mathbf{x}_i - \mathbf{x}_j|^2$.

Path-based Clustering: A connection to graph theoretic clustering methods is provided by a concept of path-based clustering. This idea replaces direct dissimilarity measurements $\mathcal{D}_{ij}$ by an effective dissimilarity D^{eff}_{ij} between two objects $\mathbf{o}_i$ and $\mathbf{o}_j$ which reflects the degree of smoothness in the transition from $\mathbf{o}_i$ to $\mathbf{o}_j$. The effective dissimilarity D^{eff}_{ij} is defined as the maximal inter object distance on the minimal connecting path:

$$\mathcal{D}^{\text{eff}}_{ij}(m, \mathcal{D}) = \min_{\mathbf{p}\in\mathcal{P}_{ij}(m)} \left\{ \max_{h\in\{1,\dots,|\mathbf{p}|-1\}} \left\{ \mathcal{D}_{\mathbf{p}[h]\mathbf{p}[h+1]} \right\} \right\}, \tag{4}$$

where $\mathbf{p}$ denotes a path from $\mathbf{o}_i$ to $\mathbf{o}_j$ and $\mathcal{P}_{ij}(m)$ is the set of all paths from $\mathbf{o}_i$ to $\mathbf{o}_j$ with all vertices in cluster $\mathfrak{G}_{m(i)}$. If both objects belong to different clusters, $\mathcal{P}_{ij}(m)$ is the empty set and the effective dissimilarity is not defined.

3 Optimization

The clustering cost functions can be minimized in principle by various deterministic or stochastic methods from combinatorial and continuous optimization. The class of stochastic Markov Chain Monte Carlo optimization algorithms with *Simulated Annealing* as a prominent technique plays an eminent role in pattern recognition. The variables of the optimization problem, e.g., the assignments in clustering, are treated as random variables of a stochastic (Markovian) process. Robust clustering methods are derived from the maximum entropy principle by Rose et al. (1990) which states that assignments are distributed according to the Gibbs distribution

$$\mathbf{P}(m; \mathfrak{D}) = \exp\bigl(-(\mathcal{H}(m; \mathfrak{D}) - \mathcal{F})/T\bigr), \tag{5}$$

$$\mathcal{F} = -T \log \sum_{m\in\mathfrak{M}} \exp\left(-\mathcal{H}(m; \mathfrak{D})/T\right) . \tag{6}$$

The "computational temperature" T serves as a Lagrange parameter to control the expected costs. The free energy $\mathcal{F}$ in Eq. (6) normalizes the Boltzmann factor $\exp(-\mathcal{H}(m; \mathfrak{D})/T)$. For clustering vectorial data according to k-means clustering the cost function $\mathcal{H}^{\text{cc}}$ is linear in the assignments and, therefore, yields a factorized Gibbs distribution

$$\mathbf{P}(m; \mathfrak{D}) = \prod_{i\le n} \mathbf{P}_{i,m(i)} \quad \text{with} \quad \mathbf{P}_{i,\nu} := \frac{\exp(-\mathcal{D}_{i,\nu}/T)}{\sum_{\mu\le k} \exp(-\mathcal{D}_{i,\mu}/T)} \quad 1 \le \nu \le k. \tag{7}$$

$\mathbf{P}_{i,\nu}$ denote expectation values of assignments. The centroids have to maximize the entropy of the Gibbs distribution which yields the centroid constraint

$$0 = \sum_{i\le n} \mathbf{P}_{i,\nu} \frac{\partial}{\partial \mathbf{y}_\nu} \mathcal{D}_{i,\nu}, \quad \forall \nu \in \{1, \dots, k\} . \tag{8}$$

The Gibbs distribution (7) can also be interpreted as the complete data likelihood for mixture models with parameters Υ. Basically, the Gibbs distribution of the k clusters describes a mixture model with equal priors for each component and equal, isotropic covariances. The assignments $m(i)$ and their expectations $\mathbf{P}_{i,\nu}$ correspond to the unobservable variables in mixture models and the component densities, respectively. Algorithmically, the centroids and the expected assignments are estimated in an iterative fashion by solving the centroid equation (8) for fixed expected assignments and, subsequently, inserting the centroids in (7) (Rose et al. (1990), Buhmann and Kühnel (1993)).

The temperature parameter T controls the uncertainty in the clustering problem, i.e., in the limit $T \to 0$ the solution of (7) corresponds to hard clustering with Boolean assignments $\mathbf{P}_{i,\nu} \in \{0, 1\}$ of a data vector $\mathbf{x}_i$ to the closest cluster center $\mathbf{y}_\nu$. Large temperature represents the fuzzy limit with partial assignments of data vectors to several clusters $(0 \leq \mathbf{P}_{i,\nu} \leq 1)$.

3.1 Mean Fields for Pairwise Clustering

Minimization of the quadratic cost function (3) turns out to be algorithmically complicated due to pairwise, potentially conflicting correlations between assignments. The deterministic annealing technique, which produces robust reestimation equations for central clustering in the maximum entropy framework, is not directly applicable to pairwise clustering since there is no analytical technique known to capture correlations between assignments $m(i)$ and $m(j)$ in an exact way. Meanfield Annealing, however, approximates the intractable Gibbs distribution by the best factorial distribution. The influence of the random variables $m(j)$, $j \neq i$ on $m(i)$ is treated by a mean field which measures the average feedback on $m(i)$. Mathematically, the approximation problem to calculate the Gibbs distribution is replaced by a minimization of the Kullback-Leibler divergence between the approximating factorial distribution and the Gibbs distribution (Hofmann and Buhmann (1998)). A maximum entropy estimate of the mean fields $h_{i\nu}$ yields the transcendental equations

$$\mathbf{P}_{i\nu} = \frac{\exp(-h_{i\nu}/T)}{\sum_{\mu \leq k} \exp(-h_{i\mu}/T)}, \tag{9}$$

$$h_{i\nu} = \frac{1}{n\pi_\nu} \sum_{(i,j)\in\mathcal{E}} \mathbf{P}_{j\nu} \left(\mathcal{D}_{ij} - \frac{1}{2n\pi_\nu} \sum_{(j,r)\in\mathcal{E}} \mathbf{P}_{r\nu} \mathcal{D}_{jr} \right). \tag{10}$$

The variables $h_{i\nu}$ depend on the given distance matrix $\mathcal{D}_{ik}$, the average assignment variables $\{\mathbf{P}_{i\nu}\}$ and the cluster weights $\pi_\nu := \sum_{i \leq n} \mathbf{P}_{i\nu}$. Equation (10) suggests an algorithm for learning the optimized cluster assignments which resembles the EM algorithm: In the E-step, the assignments $\{\mathbf{P}_{i\nu}\}$ are estimates for given $\{h_{i\nu}\}$. In the M-step the $\{h_{i\nu}\}$ are reestimated on the basis of new assignment estimates. This iterative algorithm converges to a local minimum of the Kullback Leibler divergence between the factorial ansatz and the correct Gibbs distribution which can be interpreted as consistent assignments for the pairwise data.

4 Empirical Risk Approximation and Validation

An indispensable property of data clustering solutions is their stability to sampling noise. A clustering solution with low costs on a *training* instance should yield comparably low costs on a second *test* instance. This robustness requirement limits e.g. the number of clusters which can be reliably inferred from the data. If too many clusters are supposed to be estimated then instance noise will strongly influence the values of the cluster parameters.

The field of *Statistical Learning Theory* addresses robustness questions and model complexity issues in the context of supervised learning, in particular for classification and regression. The same tradeoff between the complexity of the hypothesis class and the size of the data set limits the inference precision in data clustering. A theoretical analysis has to estimate the probability of large deviations between solutions found on two different sample sets.

The space of clustering solutions is composed of the product space $\mathfrak{M} \times \mathfrak{T}$ of assignments and continuous parameters, e.g., the means $\mathfrak{T} = \mathcal{Y}$ in k-means clustering or the conditional probabilities $\mathfrak{T} = \{q_{j|\nu}\}$ in distributional clustering. Assume that the cluster solution is quantified by the costs $\mathcal{H}(m; \mathfrak{D})$ and that we have two data sets $\mathfrak{D}_1, \mathfrak{D}_2$ drawn from the same probability distribution $(\mathcal{H}_{1,2}(m) := \mathcal{H}(m; \mathfrak{D}_{1,2})\ )$. The optimal cluster assignments w.r.t. the two data sets are denoted by m_1, m_2. It is assumed that an optimization algorithm is able to sample randomly from the set of approximating solutions $m_\gamma \in \mathfrak{L}_\gamma := \{m : \mathcal{H}_1(m) - \mathcal{H}_1(m_1) \leq \gamma\}$. A robustness criterion

$$\Delta\mathcal{H}_2(m_\gamma) := \mathcal{H}_2(m_\gamma) - \mathcal{H}_2(m_2) \tag{11}$$

which estimates the quality of an approximating training solution m_γ on a test instance should be upper bounded in probability by large deviation arguments, i.e.,

$$\mathbf{P}\left\{\Delta\mathcal{H}_2(m_\gamma) > \epsilon\right\} \leq \delta. \tag{12}$$

Notice that m_γ depends on the selection algorithms and is a random variable for stochastic sampling from the approximation set $\mathfrak{L}_\gamma$.

Statistical learning theory relates this deviation to the complexity of the solution space $\mathfrak{M}$. Since the space of all data partitionings is too large to yield meaningful bounds we coarsen this space by a minimal γ-cover $\mathfrak{M}_\gamma$, i.e., every function $m \in \mathfrak{M}$ is at most γ distant from the closest function in $\mathfrak{M}_\gamma$. Distances between two functions are measured by the l_1-distance, i.e., $d(m, \hat{m}) := \sum_{i=1}^{n} |\mathcal{D}^{(2)}_{i,m(i)} - \mathcal{D}^{(2)}_{i,\hat{m}(i)}|/n$.

The robustness measure $\varDelta\mathcal{H}_2(m_\gamma)$ can be bound by Vapnik and Chervonenkis type inequalities (1971).

$$\begin{aligned}\varDelta\mathcal{H}_2(m_\gamma) &\le \mathcal{H}_2(m_\gamma) - \inf_{m\in\mathfrak{M}_\gamma} \mathcal{H}_2(m) + \gamma \\ &\le \mathcal{H}_2(m_\gamma) - \mathcal{H}_1(m_\gamma) + \sup_{m\in\mathfrak{M}_\gamma} |\mathcal{H}_1(m) - \mathcal{H}_2(m)| + 2\gamma \\ &\le 2 \sup_{m\in\mathfrak{M}_\gamma\cup\mathfrak{L}_\gamma} |\mathcal{H}_1(m) - \mathcal{H}_2(m)| + 2\gamma. \qquad (13)\end{aligned}$$

We now have to find bounds on the probability of large deviations of training costs from test costs $|\mathcal{H}_1(m) - \mathcal{H}_2(m)|$. This probability can be bound by Bernstein's inequality (van der Vaart and Wellner (1996)) which is sensitive to the scale of cost contributions from single objects.

The expected risk of the empirical minimizer exceeds the global minimum of the expected risk by 2ϵ with a probability bounded by Bernstein's inequality ($\mathfrak{M}'_\gamma := \mathfrak{M}_\gamma \cup \mathfrak{L}_\gamma$)

$$\begin{aligned}\mathbf{P}\left\{\varDelta\mathcal{H}_2(m_\gamma) > 2\epsilon\right\} &\le \mathbf{P}\left\{\sup_{m\in\mathfrak{M}'_\gamma} |\mathcal{H}_1(m) - \mathcal{H}_2(m)| \ge \epsilon - \gamma\right\} \\ &\le \sum_{m\in\mathfrak{M}'_\gamma} \mathbf{P}\left\{|\mathcal{H}_1(m) - \mathcal{H}_2(m)| \ge \epsilon - \gamma\right\} \\ &\le 2\left|\mathfrak{M}'_\gamma\right| \sup_{m\in\mathfrak{M}'_\gamma} \exp\left(-\frac{n\,(\epsilon-\gamma)^2}{2\sigma^2 + \tau\sigma\,(\epsilon-\gamma)}\right) =: \delta. \qquad (14)\end{aligned}$$

The parameters σ, τ which determine the Bernstein inequality are dependent on the specific solution $m \in \mathfrak{M}_\gamma$. The complexity of the considered γ-cover $|\mathfrak{M}_\gamma|$ of the hypothesis class $\mathfrak{M}$ has to be small enough to guarantee with high confidence small ϵ–deviations.

This large deviation inequality weighs two competing effects in the learning problem, i. e., the probability of a large deviation exponentially decreases with growing sample size n, whereas a large deviation becomes increasingly likely with growing cardinality of the γ–cover of the hypothesis class. According to eq. (14) the sample complexity $n_0\,(\gamma, \epsilon, \delta)$ is defined by

$$\log\left|\mathfrak{M}'_\gamma\right| - \sup_{m\in\mathfrak{M}'_\gamma} \frac{n_0\,(\epsilon-\gamma)^2}{2\sigma^2 + \tau\sigma\,(\epsilon-\gamma)} + \log\frac{2}{\delta} = 0. \qquad (15)$$

Equation 15 relates the precision ϵ and the coarsening of the hypothesis class γ to the sample size n_0 with $\epsilon^{\mathrm{opt}} := \min_\gamma \epsilon(\gamma, n_0, \delta)$ and $\gamma^{\mathrm{opt}} := \arg\min_\gamma \epsilon(\gamma, n_0, \delta)$. With probability $1-\delta$ the deviation of the training costs $\mathcal{H}(m_\gamma; \mathfrak{D}_1)$ from the test costs $\mathcal{H}(m_\gamma; \mathfrak{D}_2)$ is bounded by $(\epsilon^{\mathrm{opt}} - \gamma^{\mathrm{opt}})$. Averaging over a function sphere with radius γ^{opt} around the minimizer of the training instance yields a hypothesis corresponding to a statistically significant structure in the data. The key task in the following remains to calculate an upper bound for the cardinality $\left|\mathfrak{M}'_\gamma\right|$ of the γ–cover which can be achieved by Markov Chain Monte Carlo methods (see Sinclair (1993)).

5 Discussion

Data clustering as one of the most fundamental information processing procedure to extract symbolic information from sub-symbolic data follows the four design steps of Pattern Recognition: (i) data representation, (ii) structure definition, (iii) structure optimization and (iv) structure validation. The structure definition for clusters emphasizes homogeneity or connectivity for the different data representations, e.g., vectorial, distributional and dissimilarity data. A natural choice to optimize the cluster parameters are stochastic optimization algorithms with their theoretically supported robustness to noise. Large deviation techniques from statistical learning theory and empirical process theory allow us to understand this insensitivity and to address the model selection problem in clustering. The strict separation of the four design steps greatly facilitates the search for application adapted clustering principles and provides a basis for rational algorithm design in data analysis.

References

BUHMANN, J. and KÜHNEL, H. (1993): Vector quantization with complexity costs. *IEEE Transactions on Information Theory*, *39*, 1133–1145.

HOFMANN, T. and BUHMANN, J. (1997): Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 1–14.

JAIN, A. K. and DUBES, R. C. (1988): *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ 07632.

ROSE, K., GUREWITZ, E., and FOX, G. (1990): A deterministic annealing approach to clustering. *Pattern Recognition Letters*, *11*, 589–594.

SINCLAIR, A. (1993): *Algorithms for Random Generation and Counting.* Birkhäuser, Boston, Basel, Berlin.

TISHBY, N., PEREIRA, F., and BIALEK, W. (1999): The information bottleneck method. In *Proceedings of the 37-th Allerton Conference on Communication, Control and Computing*, pages 368–377. IEEE Computer Society Press.

VAN DER VAART, A. W. and WELLNER, J. A. (1996): *Weak Convergence and Empirical Processes.* Springer-Verlag, New York, Berlin, Heidelberg.

VAPNIK, V. N. and CHERVONENKIS, A. Y. (1971): On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, *16*, 264–280.