

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/3533391>

Unsupervised and supervised data clustering with competitive neural networks

Conference Paper · July 1992

DOI: 10.1109/IJCNN.1992.227220 · Source: IEEE Xplore

CITATIONS

33

READS

113

2 authors, including:



Joachim M Buhmann

ETH Zurich

467 PUBLICATIONS 22,241 CITATIONS

SEE PROFILE

Unsupervised and Supervised Data Clustering with Competitive Neural Networks

Joachim Buhmann

Lawrence Livermore National Laboratory
Computational Physics Division
P.O.Box 808, L-270
Livermore, CA 94550, USA
jb@s1.gov

Hans Kühnel

Technische Universität München
Physik Department, T35, Boltzmannstraße
D-8046 Garching, Fed. Rep. Germany
Hans.Kuehnel@physik.tu-muenchen.de

Data clustering addresses the optimization problem of how to represent a set of data points by a reduced set of prototypes. We propose a cost function for unsupervised and supervised data clustering which comprises distortion costs, complexity costs and supervision costs. A maximum entropy estimation of the clustering cost function yields an optimal number of clusters, their positions and their cluster probabilities. A three-layer neural network with a *winner-take-all* connectivity in the clustering layer implements the proposed clustering algorithm. Complexity optimized clustering establishes a unifying framework for different clustering methods like K -means clustering, fuzzy clustering, entropy constrained vector quantization or topological feature maps and competitive neural networks.

Objective Functions for Data Clustering

Natural and artificial information processing systems in vision, speech recognition and telecommunication rely on data compression and data recoding techniques either to process large amounts of data efficiently or to discard noise and to reveal the underlying structure in a data set [1, 2, 3, 4, 5]. Mainly two different approaches to data clustering are discussed in the literature: (I) The data clustering task can be formulated as a search for a model of the data source. Bayesian inference is the method of choice for that. (II) Optimizing an objective function for data clustering defines a non-parametric strategy to partition a data set and, thereby, to reduce its information content. Following the second philosophy in this paper, we discuss objective functions for unsupervised and supervised data clustering and the respective competitive neural networks which implement these clustering algorithms. The *joint optimization* of distortion costs and complexity costs determines an optimal number of cluster centers. An additional supervision cost term is introduced in the case of supervised clustering to guide the cluster training according to a priori class information and to avoid data partitions which violate class boundaries.

The problem to represent a data set $\{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{R}^d; i = 1, \dots, N\}$ by a reduced set of data prototypes $\{\mathbf{y}_\alpha | \mathbf{y}_\alpha \in \mathbb{R}^d; \alpha = 1, \dots, K\}$ based on a distortion cost function $\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha)$ arises in a large variety of different data processing applications. The primary objective is to determine an appropriate number K of clusters ($K \ll N$) and to find their centers $\mathbf{y}_\alpha$ and their cluster probabilities p_α . The assignments $\{M_{i\alpha} | \alpha = 1, \dots, K; i = 1, \dots, N\}$, $M_{i\alpha} \in \{0, 1\}$ of data point $\mathbf{x}_i$ to cluster α are chosen such that a distortion error $\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha)$ between a data point $\mathbf{x}_i$ and a cluster center $\mathbf{y}_\alpha$ is minimized and that the resulting complexity of the cluster set is limited. $M_{i\alpha^*} = 1$ denotes that data point $\mathbf{x}_i$ is uniquely assigned to cluster α^* which implies the uniqueness constraint $\sum_\alpha M_{i\alpha} = 1$. The most frequently used distortion measure is the Euclidean distance or a power r of it $\mathcal{D}_{i\alpha} \equiv \|\mathbf{x}_i - \mathbf{y}_\alpha\|^r$. Rate distortion theory (see [6]) specifies the optimal choice of $\mathbf{y}_\alpha$, i.e., it has to be the centroid $\mathbf{y}_\alpha$ of a cluster α , $\mathbf{y}_\alpha$ being defined by $\sum_i M_{i\alpha} \frac{\partial}{\partial \mathbf{y}_\alpha} \mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha) = 0$. The

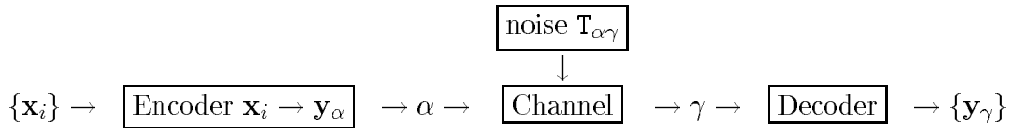
complexity $\mathcal{C}_\alpha$ of cluster α depends on the specific information processing application at hand, in particular, we assume that $\mathcal{C}_\alpha$ is only a function of the cluster probability $p_\alpha = \sum_{i=1}^N M_{i\alpha}/N$. In the case of supervised clustering we add supervision costs which penalize clusters with data points from more than class. The supervision costs $\mathcal{S}_\alpha(\{p_{\gamma\alpha}\})$ depend on the joint probabilities $p_{\gamma\alpha} = \frac{1}{N} \sum_i \Gamma_{i\gamma} M_{i\alpha}$, $\gamma = 1, \dots, G$ that a data point $\mathbf{x}_i$ out of class γ ($\Gamma_{i\gamma} = 1$) is assigned to cluster α . The proposed clustering cost function

$$\mathcal{E}(\{M_{i\alpha}\}) = \sum_{i=1}^N \sum_{\alpha=1}^K M_{i\alpha} \left(\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha) + \lambda \mathcal{C}_\alpha(p_\alpha) + \sigma \mathcal{S}_\alpha(\{p_{\gamma\alpha}\}) \right). \quad (1)$$

compromises between distortion costs, complexity costs and supervision costs, thereby determining an optimal number K of clusters. λ and σ are weighting parameters required to adjust the complexity cost and the supervision costs to the scale of the distortion costs. The cost function (1) depends only on the assignment variables $\{M_{i\alpha} | \alpha = 1, \dots, K; i = 1, \dots, N\}$ which implicitly define the variables $p_\alpha, p_{\gamma\alpha}, \mathbf{y}_\alpha$. $\mathcal{E}(\{M_{i\alpha}\})$ has to be optimized in an iterative fashion: (i) vary the assignment variables $M_{i\alpha}$ for a fixed number K of clusters such that the costs $\mathcal{E}(\{M_{i\alpha}\})$ decrease; (ii) increment the number of clusters $K \rightarrow K + 1$ and optimize $M_{i\alpha}$ again. Configurations with degenerate clusters, i.e., $\mathbf{y}_\alpha \equiv \mathbf{y}_\beta$ for $\alpha \neq \beta$ have to be rejected since they cause an overestimation of the “true” complexity of a cluster set.

An application dependent complexity measure could be the mass storage space of a storage medium, the processing hardware in electronics, the number of neurons in brains or the channel bandwidth in communication systems. Complexity costs which penalize sparsely populated clusters, i.e., $\mathcal{C}_\alpha = 1/p_\alpha^s, s = 1, 2, \dots$, favor equal cluster probabilities, thereby emphasizing the hardware aspect of a clustering solution. The special case $s = 1$ with constant total complexity costs NK corresponds to K -means clustering. A complexity measure which provides a lower bound on encoding costs in the context of data compression and bandwidth limited data transmission is the Shannon entropy of a cluster set $\langle \mathcal{C} \rangle = -\sum_\alpha p_\alpha \log p_\alpha$ ($\mathcal{C}_\alpha = -\log p_\alpha$) [3]. $\langle \mathcal{C} \rangle$ measures the minimum average number of bits necessary to uniquely encode a given cluster set.

The choice of an appropriate distortion measure $\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha)$ is another degree of freedom in the design of a clustering algorithm. The normally used squared Euclidean distance $\mathcal{D}_{i\alpha} = \|\mathbf{x}_i - \mathbf{y}_\alpha\|^2$ leaves $\mathcal{E}(\{M_{i\alpha}\})$ invariant under permutations of the cluster indices α . We will now establish a connection to a class of topological vector quantization algorithms known as self-organizing feature maps [4, 7]. Luttrell [8] discovered in his studies of hierarchical vector quantizers that a fictitious noise process in the communication channel between sender and receiver can break the permutation symmetry and favors a topological arrangement of the cluster centers $\mathbf{y}_\alpha$.



At the encoding stage data point $\mathbf{x}_i$ is assigned to cluster α . The index α , however, gets corrupted by channel noise and arrives as index γ at the receiver’s side of the communication channel. Let us denote the transition probability from index α to index γ by $T_{\alpha\gamma}$, $\sum_{\gamma=1}^K T_{\alpha\gamma} = 1$. The receiver, consequently, reconstructs the “wrong” cluster center $\mathbf{y}_\gamma$ as a representation of $\mathbf{x}_i$ instead of the “correct” reference vector $\mathbf{y}_\alpha$. Therefore, the transitions $T_{\alpha\gamma}$ cause additional distortions which have to be taken into account when we estimate the most likely assignment variables. The average clustering costs of the data set $\{\mathbf{x}_i\}$ are

$$\mathcal{H} = \sum_i \sum_\alpha M_{i\alpha} \left(\langle \mathcal{D}_{i\alpha} \rangle + \lambda \mathcal{C}(p_\alpha) + \sigma \mathcal{S}_\alpha(\{p_{\gamma\alpha}\}) \right). \quad (2)$$

The average distortion error $\langle\langle \mathcal{D}_{i\alpha} \rangle\rangle \equiv \sum_{\gamma} \mathbf{T}_{\alpha\gamma} \mathcal{D}_{i\gamma}(\mathbf{x}_i, \mathbf{y}_{\gamma})$ between data point $\mathbf{x}_i$ and cluster center $\mathbf{y}_{\alpha}$ quantifies the intrinsic quantization errors as well as the noise induced errors. The centroids $\mathbf{y}_{\alpha}$ are defined by $\sum_i \sum_{\gamma} M_{i\gamma} \mathbf{T}_{\gamma\alpha} \frac{\partial}{\partial \mathbf{y}_{\alpha}} \mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_{\alpha}) = 0$. The noise process with its transition probabilities $\mathbf{T}_{\alpha\gamma}$ induces a topology onto the set of clusters, e.g. a tridiagonal transition matrix $\mathbf{T}_{\alpha\alpha} = 1 - \eta$; $\mathbf{T}_{\alpha, \alpha \pm 1} = \eta/2$; $\mathbf{T}_{\gamma\alpha} = 0 \forall |\alpha - \gamma| > 1$ (see Fig. 2(c)) defines a linear chain with nearest neighbor transitions. Note, however, that complexity optimized clustering determines the length of the chain or, for a more general noise distribution, an optimal size of the cluster set. This stopping criterion for adding new nodes generalizes self-organizing topological maps [4, 7] and removes arbitrariness in the design of these networks. Topology preserving clustering reduces to non-topological clustering if we set $\mathbf{T}_{\alpha\gamma} = 1$ for $\alpha = \gamma$ and $\mathbf{T}_{\alpha\gamma} = 0$ otherwise.

The cost term $\mathcal{S}_{\alpha}(\{p_{\gamma\alpha}\})$ in Eqs. (1,2) allows us to supervise the training of a cluster set. Many applications provide labeled training data, each data point being assigned to a class $\gamma = 1, \dots, G$. The cluster set after training is then used to classify unlabeled test data. The supervision cost term $\mathcal{S}_{\alpha}$ is supposed to assure that clusters do not overlap with class boundaries, i.e., data belonging to different classes should not be assigned to the same cluster. If this constraint is fulfilled $\mathcal{S}_{\alpha}$ should vanish. $\mathcal{S}_{\alpha}$ depends on the conditional probability $p_{\gamma|\alpha} \equiv p_{\gamma\alpha}/p_{\alpha}$, $p_{\gamma\alpha} = \sum_i \Gamma_{i\gamma} M_{i\alpha}/N$ being the joint probability of a data point out of class γ being assigned to cluster α . $\mathcal{S}_{\alpha}$ should be positive and it should vanish if $p_{\gamma|\alpha}$ approached certainty or vanishes. A supervision cost term which suffices these conditions is the “conditional class entropy” $\mathcal{S}_{\alpha} = -\sum_{\gamma} p_{\gamma|\alpha} \log(p_{\gamma|\alpha})$. As discussed in [9] the “conditional class entropy” as a penalty term favors a partitioning of the input space with clusters of unique class adherence. The simplest functional form of $\mathcal{S}_{\alpha}$ being polynomial in $p_{\gamma\alpha}$ is $\mathcal{S}_{\alpha} = \sum_{\gamma} p_{\gamma|\alpha}(1 - p_{\gamma|\alpha})$.

Maximum Entropy Estimation of Cluster Centers $\mathbf{y}_{\alpha}$ and Joint Probabilities $p_{\gamma\alpha}$

Different combinations of complexity terms, topology constraints and class information for supervised clustering define a variety of clustering algorithms which are relevant in different information processing contexts. To derive robust, preferably parallel algorithms for these data clustering cases, we study the optimization problem (2) in a probabilistic framework (see Rose et al. [2]). The most likely distribution of assignment variables $M_{i\alpha}$ can be determined by the maximum entropy principle [10], and equals the Gibbs distribution $P(\{M\}) = \exp(-\beta \mathcal{E})/\mathcal{Z}$ with the partition function $\mathcal{Z}$ as normalization. The “computational temperature” $T = 1/\beta$, which play the role of a Lagrange parameter for the average clustering costs, controls the randomness in the assignment process. Statistical physics (see e.g. [11]) states that maximizing the entropy at a fixed temperature is equivalent to minimizing the free energy

$$\mathcal{F} = -\sum_{\alpha} p_{\gamma\alpha} \hat{p}_{\gamma\alpha} - \frac{1}{\beta N} \sum_i \log \left(\sum_{\alpha} \exp[-\beta(\langle\langle \mathcal{D}_{i\alpha} \rangle\rangle + \lambda \mathcal{C}_{\alpha} + \sigma \mathcal{S}_{\alpha} + \sum_{\gamma} \Gamma_{i\gamma} \hat{p}_{\gamma\alpha})] \right). \quad (3)$$

with respect to the variables $p_{\gamma\alpha}, \mathbf{y}_{\alpha}$. The term $\sum_{\gamma} \Gamma_{i\gamma} \hat{p}_{\gamma\alpha} \equiv \lambda p_{\alpha} \frac{\partial \mathcal{C}_{\alpha}}{\partial p_{\alpha}} + \sigma p_{\alpha} \sum_{\gamma} \Gamma_{i\gamma} \frac{\partial \mathcal{S}_{\alpha}}{\partial p_{\gamma\alpha}}$ enforces the constraint $p_{\gamma\alpha} = \sum_{i=1}^N \Gamma_{i\gamma} M_{i\alpha}/N$. The resulting re-estimation equations for the expected cluster probabilities and the expected centroid positions are

$$p_{\gamma\alpha} = \frac{1}{N} \sum_i \Gamma_{i\gamma} \langle M_{i\alpha} \rangle, \quad (4)$$

$$0 = \frac{1}{N} \sum_i \sum_{\gamma} \mathbf{T}_{\gamma\alpha} \langle M_{i\gamma} \rangle \frac{\partial}{\partial \mathbf{y}_{\alpha}} \mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_{\alpha}), \quad (5)$$

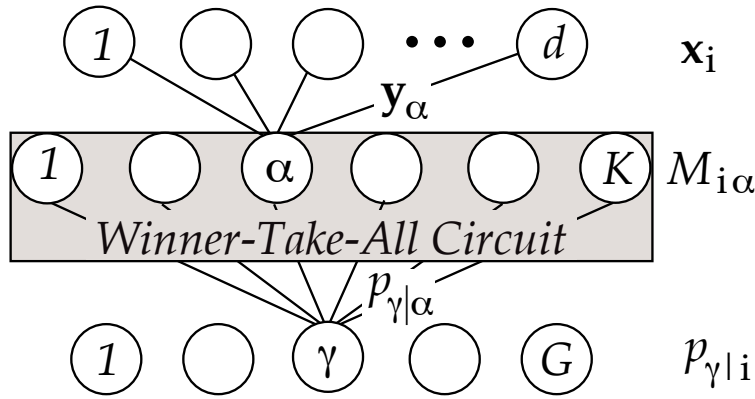


Figure 1: Architecture of a three layer competitive neural network for data clustering with d neurons in the input layer, K neurons in the clustering layer with activity $\langle M_{i\alpha} \rangle$ and G neurons in the classification layer with activity $p_{\gamma|i}$.

$$\langle M_{i\alpha} \rangle = \frac{\exp \left[-\beta (\langle \mathcal{D}_{i\alpha} \rangle + \lambda \mathcal{C}_\alpha + \sigma \mathcal{S}_\alpha + \sum_\gamma \hat{p}_{\gamma\alpha} \Gamma_{i\gamma}) \right]}{\sum_{\nu=1}^K \exp \left[-\beta (\langle \mathcal{D}_{i\nu} \rangle + \lambda \mathcal{C}_\nu + \sigma \mathcal{S}_\nu + \sum_\gamma \hat{p}_{\gamma\nu} \Gamma_{i\gamma}) \right]}. \quad (6)$$

The expectation value $\langle M_{i\alpha} \rangle$ of the assignment variable $M_{i\alpha}$ can be interpreted as the conditional probability that data point $\mathbf{x}_i$ is assigned to cluster α . The temperature T controls the “fuzziness” of this assignment.

The global minimum of the free energy (3) with respect to $p_{\gamma\alpha}, \mathbf{y}_\alpha$ determines the maximum entropy solution of the cost function (1). Note that the K^N dimensional optimization problem (1) or (2) has been reduced to a $K(d + G)$ dimensional minimization of $\mathcal{F}$. To find the optimal parameters $p_{\gamma\alpha}, \mathbf{y}_\alpha$ and the number of clusters K which minimize the free energy we start with one cluster located at the centroid of the data distribution, split that cluster and reestimate $p_{\gamma\alpha}, \mathbf{y}_\alpha$ using Eqs. (4,5). The new configuration is accepted as an improved solution if the free energy (3) has been decreased. This splitting and reestimation loop is continued until we fail to find a new configuration with lower free energy. An iterative splitting strategy is required because the costfunction exhibits numerous local minima, especially since we compare solutions with different numbers of clusters. Cluster splitting in the topological case should be performed such that a new cluster does not violate the topology constraint severely and gives rise to a topological defect. The resulting configuration is stable under single cluster splitting, but we are not assured to find the global minimum.

The clustering algorithm is implemented by a competitive three-layer neural network with an architecture depicted in Fig. 1. The classification layer (third layer) consists of linear units γ which calculate the conditional probability $p_{\gamma|i}$ that data point $\mathbf{x}_i$ is assigned to class γ . This layer is absent in unsupervised clustering. The synaptic vector $\mathbf{y}_\alpha$ defines a localized receptive field of the clustering unit α which is similar to radial basis function units. The connections between the clustering unit α and the classification unit γ is given by the conditional probability $p_{\gamma|\alpha}$. The synaptic weights $y_{\alpha i}, p_{\gamma|\alpha}$ are calculated from Eqs. (4,5) or from an iterative *online* learning rule described in the next section. No time-consuming backpropagation of errors is necessary. The activity in the classification layer determines the conditional probability $p_{\gamma|i}$ that data point i is classified as γ .

A non-topological clustering result ($T = 0$) for the logarithmic complexity ($\mathcal{C}_\alpha = -\log p_\alpha$) measure is shown in Fig. 2(b). At high complexity costs ($\lambda = 2.5$ [12]) the algorithm places four clusters at the centers of the Gaussians. In the limit of small complexity costs (b) the best clustering solution densely covers the data distribution. The specific choice of logarithmic complexity costs causes an almost homogeneous density of cluster centers, a phenomenon which is explained by the vanishing average complexity costs $\langle \mathcal{C}_\alpha \rangle = -p_\alpha \log p_\alpha$ of very sparsely occupied clusters. In the clustering

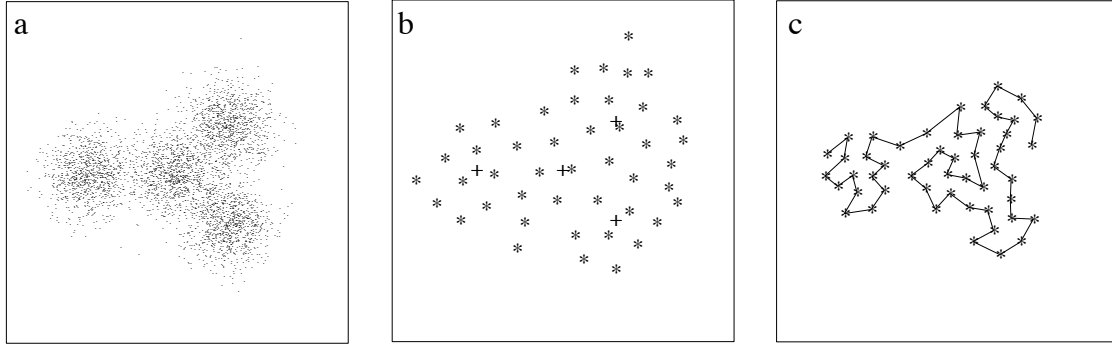


Figure 2: A data distribution (4000 data points) (a), generated by four normally distributed sources is clustered with the complexity measure $\mathcal{C}_\alpha = -\ln p_\alpha$ and $\lambda = 0.4$ (b). The plus signs (+) denote the centers of the Gaussians and stars (*) denote cluster centers. Figure (c) shows a topology preserving clustering solution with complexity $\mathcal{C}_\alpha = 1/p_\alpha$. The short average distances between consecutive clusters demonstrate that the total distortion error has been minimized, including the distortions induced by external noise ($\eta = 0.05$).

experiment depicted in Fig. 2(c) we have added channel noise which triggers nearest neighbor transitions between clusters with probability $\eta = 0.05$. The short links between neighboring clusters demonstrate that the distortions due to external noise have also been optimized.

Online Clustering

The described clustering procedures are *batch* algorithms since they require all data points $\{\mathbf{x}_i\}$ to be available at the same time, a very restrictive constraint for applications dealing with a large data volume. *Online* clustering algorithms process a data stream sequentially by updating the synaptic weights $p_{\gamma|\alpha}$, $\mathbf{y}_\alpha$ iteratively. The optimal update step for $p_{\gamma|\alpha}$, $\mathbf{y}_\alpha$, after $N - 1$ data points have been processed, is

$$\Delta p_{\gamma|\alpha} \equiv p_{\gamma|\alpha}^{(N)} - p_{\gamma|\alpha}^{(N-1)} = \frac{\langle M_{N,\alpha} \rangle}{(N-1)p_\alpha^{(N-1)}} \left(\Gamma_{N\gamma} - p_{\gamma|\alpha}^{(N-1)} \right), \quad (7)$$

$$\Delta \mathbf{y}_\alpha \equiv \mathbf{y}_\alpha^{(N)} - \mathbf{y}_\alpha^{(N-1)} = - \frac{\sum_\gamma \langle M_{N,\gamma} \rangle \mathbf{T}_{\gamma\alpha} \frac{\partial}{\partial \mathbf{y}_\alpha} \mathcal{D}_{N,\alpha}(\mathbf{y}_\alpha^{(N-1)}, \mathbf{x}_N)}{\sum_{i=1}^N \sum_\gamma \langle M_{i,\gamma} \rangle \mathbf{T}_{\gamma\alpha} \frac{\partial^2}{\partial \mathbf{y}_\alpha^2} \mathcal{D}_{i,\alpha}(\mathbf{y}_\alpha^{(N-1)}, \mathbf{x}_i)} \quad (8)$$

The upper indices (N) and $(N-1)$ denote the estimates of $p_{\gamma|\alpha}$, $\mathbf{y}_\alpha$ after $\mathbf{x}_N$ and $\mathbf{x}_{N-1}$ have been processed, respectively. To derive Eqs. (7,8) we have expanded Eqs. (4,5) up to linear terms in $\Delta p_{\gamma\alpha}$, $\Delta \mathbf{y}_\alpha$, keeping $\langle M_{N,\alpha} \rangle$ fixed. All terms proportional to derivatives of $\langle M_{N,\alpha} \rangle$ vanish in the hard clustering limit ($T \rightarrow 0$).

For squared Euclidean distortion costs ($\mathcal{D}_{i\alpha} = \|\mathbf{x}_i - \mathbf{y}_\alpha\|^2$), Eq. (8) reduces to the form

$$\Delta \mathbf{y}_\alpha = \frac{1}{N \sum_\gamma \mathbf{T}_{\gamma\alpha} p_\gamma^N} \sum_\gamma \mathbf{T}_{\gamma\alpha} \langle M_{N,\gamma} \rangle (\mathbf{x}_N - \mathbf{y}_\alpha^{(N-1)}) \quad (9)$$

which is similar to learning in competitive neural networks. $\mathbf{y}_\alpha^{(N-1)}$ is moved towards the most recent data point $\mathbf{x}_N$ proportionally to the mismatch $(\mathbf{x}_N - \mathbf{y}_\alpha^{(N-1)})$ and proportional to $\mathbf{x}_N$'s effective

membership $\sum_{\gamma} \mathbf{T}_{\gamma\alpha} \langle M_{N,\gamma} \rangle$ in cluster α . In addition to that, the update formula (9) weights any change in $\mathbf{y}_{\alpha}^{(N)}$ by the effective number of data points $N \sum_{\gamma} \mathbf{T}_{\gamma\alpha} p_{\gamma}^{(N)}$ which have already been assigned to cluster α . This history dependent learning rate generalizes the algorithm for self-organizing topological maps which postulates the same learning rate for all units [4, 7].

Online optimization of the number of clusters K relies on a heuristics for cluster creation: A data point $\mathbf{x}_i$ initializes a new cluster $K+1$ with $\mathbf{y}_{K+1} = \mathbf{x}_i$ if the clustering costs by assigning $\mathbf{x}_i$ to an already existing cluster exceeds the complexity and supervision costs of the new cluster $K+1$, i.e., $\lambda \mathcal{C}_{K+1} < \min_{\alpha} (\mathcal{D}_{i\alpha} + \lambda \mathcal{C}_{\alpha} + \sigma \mathcal{S}_{\alpha})$. Results of supervised clustering experiments employing this heuristics are discussed in [12].

Conclusion

Complexity optimized clustering is a maximum entropy approach to unsupervised and supervised data clustering which determines the optimal number of clusters as a compromise between distortion errors and complexity of a cluster set. The complexity term turns out to be as important for the design of a cluster set as the distortion measure. We have derived a batch and an online version of the proposed clustering algorithm which maps onto a three layer competitive neural network with *winner-take-all* connectivity in the second layer. The proposed framework for data clustering unifies traditional clustering techniques like K -means clustering, entropy constraint clustering or fuzzy clustering with neural network approaches as topological vector quantizers. The network size and the cluster parameters are determined by a problem adapted complexity function which removes considerable arbitrariness present in other non-parametric clustering methods.

JB has been supported by the German Federal Ministry of Science and Technology (ITR-8800-H1) and by AFOSR (88-0274) (PI: C. von der Malsburg) while working at the University of Southern California, Center for Neural Engineering. HK is a recipient of a graduate fellowship, Technische Universität München.

- [1] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communications COM*, vol. 28, pp. 84–95, 1980.
- [2] K. Rose, E. Gurewitz, and G. Fox, "Statistical mechanics and phase transitions in clustering," *Physical Review Letters*, vol. 65, no. 8, pp. 945–948, 1990.
- [3] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. on Acoustics, Speech and Signal Processing ASSP*, vol. 37, pp. 31–42, 1989.
- [4] T. Kohonen, *Self-organization and Associative Memory*. Berlin: Springer, 1984.
- [5] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," *Cognitive Science*, vol. 9, pp. 75–112, 1985.
- [6] R. M. Gray, "Vector quantization," *IEEE Acoustics, Speech and Signal Processing Magazine*, pp. 4–29, April 1984.
- [7] H. Ritter, T. Martinetz, and K. Schulten, *Neural Computation and Self-organizing Maps*. New York: Addison Wesley, 1992.
- [8] S. P. Luttrell, "Hierarchical vector quantisation," *IEE Proceedings*, vol. 136, pp. 405–413, 1989.
- [9] M. Bichsel and P. Seitz, "Minimum class entropy: A maximum information approach to layered networks," *Neural Networks*, vol. 2, pp. 133–141, 1989.
- [10] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, pp. 620–630, 1957.
- [11] D. Amit, *Modelling Brain Function*. Cambridge: Cambridge University Press, 1989.
- [12] J. Buhmann and H. Kühnel, "Complexity optimized data clustering by competitive neural networks." Preprint, 1992.