Inferring Hierarchical Clustering Structures by Deterministic Annealing

Thomas Hofmann & Joachim M. Buhmann *
Rheinische Friedrich–Wilhelms–Universität
Institut für Informatik III, Römerstraße 164, D-53117 Bonn, Germany
email:{th,jb}@cs.uni-bonn.de, http://www-dbv.cs.uni-bonn.de

Abstract

The unsupervised detection of hierarchical structures is a major topic in unsupervised learning and one of the key questions in data analysis and representation. We propose a novel algorithm for the problem of learning decision trees for data clustering and related problems. In contrast to many other methods based on successive tree growing and pruning, we propose an objective function for tree evaluation and we derive a non–greedy technique for tree growing. Applying the principles of maximum entropy and minimum cross entropy, a deterministic annealing algorithm is derived in a meanfield approximation. This technique allows us to canonically superimpose tree structures and to fit parameters to averaged or 'fuzzified' trees.

Introduction

Clustering is one of the fundamental problems in exploratory data analysis. Data clustering problems occur in pattern recognition, statistics, unsupervised learning, neural networks, data mining, machine learning and many other scientific fields. The wide range of applications is explained by the fact that clustering procedures are important tools for an automated or interactive detection of structure in data sets. Especially for large data sets grouping data and extracting typical prototypes is important for a compact representation and is a precondition for further symbolic processing stages. In the context of data clustering the detection of hierarchical structures is an essential goal of data analysis. In this paper we consider binary trees with stochastic transition nodes, (Breiman *et al.* 1984) applied to vector–valued data.

We will formulate data clustering as a stochastic optimization problem to be addressed in the *maximum entropy framework.* Maximum entropy methods have been introduced as a stochastic optimization method, called *simulated annealing* in a seminal paper of Kirkpatrick et al. (Kirkpatrick, Gelatt, & Vecchi 1983). To overcome the computational burden of Monte Carlo sampling, efficient *deterministic annealing* variants have been derived for a number of important optimization problems (Yuille 1990; Kosowsky & Yuille 1994; Buhmann & Hofmann 1994; Gold & Rangarajan 1996), including unconstrained clustering and vector quantization. (Rose, Gurewitz, & Fox 1990; Buhmann & Kühnel 1993). Maximum entropy methods have recently been successfully applied to the case of tree–structured vector quantization in (Miller & Rose 1994; 1996). Similar methods have also been used in the context of regression (Jordan & Jacobs 1994) and for unsupervised learning problems (Dayan, Hinton, & Zemel 1995). The key idea in simulated and deterministic annealing is to reformulate a given combinatorial optimization problem as a stochastic optimization problem. A temperature parameter T is introduced to control the amplitude of the induced noise. In the zero temperature limit, $T \to 0$, the combinatorial problem is recovered, while for high temperatures the objective function is smoothened. Tracking solutions from high temperatures thus helps us to avoid unfavorable local minima.

The major novelty of our approach is an *explicit* treatment of the topology of binary trees in the maximum entropy framework, which results in a systematic and well–founded 'fuzzification' of binary tree topologies. At a finite computational temperature different trees are superimposed resulting in an average tree structure. An average tree is not a single tree but a tree mixture. The proposed algorithm optimizes the tree topology *jointly* with all other relevant parameters, e.g. data assignments to clusters and decision node parameters.

Unconstrained Data Clustering

We restrict our attention to the case of real–valued data vectors $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^d : 1 \leq i \leq N\}$, and a corresponding set of prototypes $\mathcal{Y} = \{\mathbf{y}_\nu \in \mathbb{R}^d : 1 \leq \nu \leq K\}$, $K \ll N$, $\mathbf{y}_\nu$ representing a group G_ν. To describe the mapping of data vectors to prototypes we introduce an indicator function representation by Boolean assignment matrices $M \in \{0,1\}^{N \times K}$ obeying the con-

*Supported by the Federal Ministry for Science and Technology (BMBF) under grant # 01 M 3021 A/4. The LANDSAT data have been provided by D. Klaus.

straints $\sum_{\nu=1}^{K} M_{i\nu} = 1$, for all i. The objective function for unconstrained data clustering is usually stated as (Duda & Hart 1973)

$$\mathcal{H}(M,\mathcal{Y}|\mathcal{X}) = \sum_{i=1}^{N}\sum_{\nu=1}^{K} M_{i\nu}\, \mathcal{D}(\mathbf{x}_i,\mathbf{y}_\nu), \qquad (1)$$

where $\mathcal{D}$ is a problem dependent distortion measure, e.g. $\mathcal{D}(\mathbf{x}_i,\mathbf{y}_\nu) = \|\mathbf{x}_i - \mathbf{y}_\nu\|^2$. Applying the principle of maximum entropy, Boolean assignments are replaced by assignment probabilities $\langle M_{i\nu}\rangle$, maximizing the entropy $S = -\sum_{i=1}^{N}\sum_{\nu=1}^{K}\langle M_{i\nu}\rangle \log\langle M_{i\nu}\rangle$ subject to fixed expected costs $\langle\mathcal{H}\rangle$. For a given set of prototypes the assignment probability of vector $\mathbf{x}_i$ to group G_ν is the Gibbs distribution

$$\langle M_{i\nu}\rangle = \frac{\exp\left[-\mathcal{D}(\mathbf{x}_i,\mathbf{y}_\nu)/T_M\right]}{\sum_{\mu=1}^{K}\exp\left[-\mathcal{D}(\mathbf{x}_i,\mathbf{y}_\mu)/T_M\right]}, \qquad (2)$$

where T_M is the computational temperature. Minimization of the expected costs with respect to the prototype vectors results in an additional set of centroid equations,

$$\mathbf{y}_\nu = \sum_{i=1}^{N}\langle M_{i\nu}\rangle \mathbf{x}_i \,/\, \sum_{i=1}^{N}\langle M_{i\nu}\rangle, \qquad (3)$$

for the case of squared Euclidean distances. Eqs. (2) and (3) can be solved efficiently by an EM algorithms (Dempster, Laird, & Rubin 1977). In a more general situation additional prior assignment probabilities $\pi_{i\nu}$ are given. Applying the principle of minimum cross entropy this results in modified, 'tilted' assignment probabilities,

$$\langle M_{i\nu}\rangle^\pi = \frac{\pi_{i\nu}\exp\left[-\mathcal{D}(\mathbf{x}_i,\mathbf{y}_\nu)/T_M\right]}{\sum_{\mu=1}^{K}\pi_{i\mu}\exp\left[-\mathcal{D}(\mathbf{x}_i,\mathbf{y}_\mu)/T_M\right]}, \qquad (4)$$

which minimize the cross entropy to the prior for fixed costs $\langle\mathcal{H}\rangle$ ((Miller & Rose 1994)). For uniform priors, we recover Eq. (2) as expected. Tilted assignments will be used in the following section to model the influence of the cluster hierarchy on data assignments.

Decision Trees for Data Clustering

In this paper we consider stochastic binary decision trees with a given number of K leaves, representing the data clusters. We denote nodes of the tree by n_α, $0 \le \alpha \le 2K-2$. Associated with each inner node n_α are two test vectors $\mathbf{y}_\alpha^l, \mathbf{y}_\alpha^r \in \mathbb{R}^d$ and a control parameter $\lambda_\alpha \in \mathbb{R}^+$. The test vectors determine transition probabilities $p_{i\alpha}^l(\mathbf{x})$ and $p_{i\alpha}^r(\mathbf{x})$ for a given vector $\mathbf{x}$ according to the formula

$$p_\alpha^{l/r}(\mathbf{x}) = \frac{\exp\left[-\lambda_\alpha\mathcal{D}(\mathbf{x},\mathbf{y}_\alpha^{l/r})\right]}{\exp\left[-\lambda_\alpha\mathcal{D}(\mathbf{x},\mathbf{y}_\alpha^l)\right] + \exp\left[-\lambda_\alpha\mathcal{D}(\mathbf{x},\mathbf{y}_\alpha^r)\right]}. \qquad (5)$$

$p_\alpha^l(\mathbf{x})$ and $p_\alpha^r(\mathbf{x})$ are the probability for vector $\mathbf{x}$ to continue its path with the left and right successor of n_α, respectively. λ_α controls the stochasticity of the transition, hard decision boundaries are obtained for $\lambda_\alpha \to \infty$. The path probability $\pi_\gamma(\mathbf{x})$ of a data vector $\mathbf{x}$ from the root to a node n_γ is given by the product of all transition probabilities at inner nodes on that path. In the limit of all $\lambda_\alpha \to \infty$ the tree defines a unique partitioning of the data space. Following (Miller & Rose 1996) we optimize the tree in order to minimize the deviation of the decision tree data partitioning from an unconstrained clustering solution with assignment probabilities $\{\langle M_{i\nu}\rangle\}$. As a suitable measure of divergence between probabilities the cross-entropy or *Kullback–Leibler divergence* is employed,

$$\mathcal{I}\left(\{\langle M_{i\nu}\rangle\}\|\{\pi_{i\nu}\}\right) = \sum_{i=1}^{N}\sum_{\nu=1}^{K}\langle M_{i\nu}\rangle\log\frac{\langle M_{i\nu}\rangle}{\pi_{i\nu}}, \qquad (6)$$

where $\pi_{i\nu} = \pi_{\nu+K-2}(\mathbf{x}_i)$. The binary tree is optimized such that the leaf probabilities $\pi_{i\nu}$ approximate as closely as possible the target probabilities. Conversely, for a given tree the prototype vectors $\mathcal{Y}$ are selected to minimize the expected distortion $\mathcal{H}(\mathcal{Y},\{\pi_{i\nu}\}) = \sum_{i=1}^{N}\sum_{\nu=1}^{K}\langle M_{i\nu}\rangle^\pi\ \mathcal{D}(\mathbf{x}_i,\mathbf{y}_\nu)$, where $\langle M_{i\nu}\rangle^\pi$ is the tilted distribution from Eq. (4). The path probabilities obtained from the tree take the role of a prior to impose structural constraints on the selection of prototypes.

Since our goal is to explicitly optimize the tree topology, we introduce an adjacency matrix representation for binary trees. Let $U^l, U^r \in \{0,1\}^{(K-1)\times(2K-1)}$ encode the successor relation between nodes in the tree. $U_{\alpha\gamma}^{l/r} = 1$ denotes that n_γ is the immediate left/right successor of inner node n_α. To avoid directed cycles we use the node numbering as a total order, where successing nodes are required to have a higher index. Furthermore every inner node has exactly one left and one right successor and all nodes except the root n_0 are required to have a unique predecessor. The path probabilities $\pi_\gamma(\mathbf{x})$ are related to the adjacency matrices by the formula,

$$\pi_\gamma(\mathbf{x}) = \sum_{\alpha=0}^{\gamma-1}\pi_\alpha(\mathbf{x})\left[U_{\alpha\gamma}^l\, p_\alpha^l(\mathbf{x}) + U_{\alpha\gamma}^r\, p_\alpha^r(\mathbf{x})\right], \qquad (7)$$

with $\pi_0(\mathbf{x}) = 1$. Path probabilities are efficiently calculated by sequentially propagating the probabilities from the root to the leaf nodes. This results in a well-defined optimization problem with a single objective function for the tree topology encoded by U^l, U^r and all involved continuous decision node parameters.

Optimizing the Tree Topology

The problem of finding an optimal decision tree is computationally difficult for two reasons: (i) the number of binary trees grows exponentially with the number of leaves; (ii) evaluating the quality of a single topology requires to fit all continuous parameters for test vectors and prototypes. The maximum entropy method offers a stochastic framework which renders an average over tree topologies feasible. Parameters are fitted not to a single tree, but to a weighted superposition of structures which converges only in the zero temperature limit towards a uniquely determined topology. This results in a 'fuzzification' of structures at finite temperatures, which is gradually eliminated in an annealing process.

Consider an extension of the probabilistic partitioning model, such that not only the transitions are stochastic, but also the successors of n_α are randomly drawn from the set of nodes $\{n_\gamma, \gamma > \alpha\}$. This means the connection between n_α and n_γ, encoded by $U^l_{\alpha\gamma}, U^r_{\alpha\gamma}$ is a random variable, with expectations $q^l_{\alpha\gamma} = \langle U^l_{\alpha\gamma}\rangle$ and $q^r_{\alpha\gamma} = \langle U^r_{\alpha\gamma}\rangle$, respectively. The probabilities have to be chosen such that $\sum_{\gamma>\alpha} q^l_{\alpha\gamma} = \sum_{\gamma>\alpha} q^r_{\alpha\gamma} = 1$ in order to obtain a correct normalization. A class of probabilities which is of special interest in this context are *fair* probability distributions. A fair probability distribution possesses the additional property that every node except the root has the same average number of predecessor, i.e. $\sum_{\alpha=0}^{\gamma-1} \left(q^l_{\alpha\gamma} + q^r_{\alpha\gamma}\right) = 1$, for all $\gamma > 0$. Fair probability distribution have the advantage, that the constraints on U^l and U^r are at least fulfilled in the average. In the extended model we can calculate path probabilities for $\mathbf{x}$ simply by replacing the Boolean variables in Eq. (7) by their probabilities.

Applying the maximum entropy principle to the objective function in Eq. (6), we assign the Gibbs probabilities $P(U^l, U^r) = \frac{1}{\mathcal{Z}} \exp\left[-\mathcal{I}(U^l, U^r)/T_U\right]$ to every tree topology U^l, U^r. $\mathcal{Z}$ is a normalization constant and T_U a temperature (or Lagrange) parameter. Ideally, we would like to average tree topologies according to the Gibbs distribution, without performing a tedious Monte Carlo sampling of trees. A standard approximation technique to analytically calculate Gibbs averages is the *meanfield approximation.* In the meanfield approximation we restrict the set of admissible probability distributions to distributions Q which are *factorial* and *fair.* Within this restricted set we want to pick a Q^* which maximizes the entropy for fixed expected costs or equivalently minimizes the cross entropy to the true Gibbs distribution $\mathcal{I}\,(Q\|P)$.

Omitting the technical details, the link probabilities $q^{l/r}_{\alpha\gamma}$ of Q^* are 0 for $\alpha \geq \gamma$ and are otherwise given by

$$q^{l/r}_{\alpha\gamma} = \frac{\exp\left[-\left(h^{l/r}_{\alpha\gamma} + \rho_\gamma\right)\right]}{\sum_{\bar{\gamma}>\alpha} \exp\left[-\left(h^{l/r}_{\alpha\bar{\gamma}} + \rho_{\bar{\gamma}}\right)\right]}, \; h^{l/r}_{\alpha\gamma} = \frac{\partial \mathcal{I}}{\partial q^{l/r}_{\alpha\gamma}}. \quad (8)$$

The above cross entropy minimization problem has been reduced to the problem of finding values for the Lagrange parameters ρ_γ, such that Q is fair. Standard methods from combinatorial optimization, developed in the context of matching problems, can be applied to find solutions for Eq. (8) if all $h^{l/r}_{\alpha\gamma}$ are kept fixed. In our simulations we used an iterative procedure known as Sinkhorn's algorithm (Sinkhorn 1964; Kosowsky & Yuille 1994). To give the basic idea, the Lagrange parameter ρ_γ can be interpreted as the 'price' of linking n_γ to another node n_α. These prices have to be simultaneously adjusted, such that every node has in the average exactly one predecessors. To arrive at a final solution we recalculate the derivatives

$$\frac{\partial \mathcal{I}}{\partial q^{l/r}_{\alpha\gamma}} = -\sum_{i=1}^{N}\sum_{\nu=1}^{K} \frac{\langle M_{i\nu}\rangle}{\pi_{i\nu}} \frac{\partial \pi_{i\nu}}{\partial q^{l/r}_{\alpha\gamma}} \quad (9)$$

and insert into Eq. (8), until a stationary state is reached. This is similar to the application of Sinkhorn's algorithm for graph matching problems (Gold & Rangarajan 1996).

Fitting Continuous Tree Parameters

The continuous decision node parameters are chosen in order to minimize $\mathcal{I}$. Applying the chain rule in calculating derivatives yields the final formula

$$\frac{\partial \mathcal{I}}{\partial \mathbf{y}^{l/r}_\alpha} = -2\lambda_\alpha \left(\mathbf{x}_i - \mathbf{y}^{l/r}_\alpha\right) \Big[s^{l/r}_\alpha(\mathbf{x}_i) - p^{l/r}_\alpha(\mathbf{x}_i) \left(s^l_\alpha(\mathbf{x}_i) + s^r_\alpha(\mathbf{x}_i)\right)\Big], \quad (10)$$

where $s^{l/r}_\alpha(\mathbf{x}_i)$ denotes up-propagated unconstrained leaf probabilities. The test vectors can be optimized by gradient methods, e.g. steepest descent or conjugate gradient techniques. The derivation of similar equations for the control parameters λ_α is straightforward.

The optimization of prototype vectors $\mathbf{y}_\nu$ proceeds according to the centroid condition in Eq. (3), with the unconstrained assignment probabilities replaced by the 'tilted' probabilities of Eq. (4). The only remaining variables are the temperature parameters T_M and T_U, which are iteratively decreased according to an appropriate annealing schedule.

Tree Clustering Algorithm (TCA)

```
INITIALIZATION
  choose $\mathbf{y}_\nu, \mathbf{y}^{l/r}_\alpha, \lambda_\alpha$ randomly
  chose $\langle M_{i\nu}\rangle, \langle U^{l/r}_{\alpha\gamma}\rangle \in (0,1)$ randomly;
  temperature $T_M \leftarrow T_0, T_U \leftarrow cT_M$;
WHILE $T_M > T_{\mathrm{FINAL}}$
  REPEAT
    estimate tilted assignm. $\{\langle M^\pi_{i\nu}\rangle\}$, Eq.(4)
    update prototypes $\{\mathbf{y}_\nu\}$ with tilted assignm.
    calc. unconstr. assignm. $\{\langle M_{i\nu}\rangle\}$, Eq.(2)
    adapt $\{\mathbf{y}^{l/r}_\alpha\}$ and $\{\lambda_\alpha\}$ by gradient descent
    apply Sinkhorn's algorithm to calc. $\{q^{l/r}_{\alpha\gamma}\}$
  UNTIL all $\{\mathbf{y}_\nu\}, \{\mathbf{y}^{l/r}_\alpha\}, \{\lambda_\alpha\}$ are stationary
  $T_M \leftarrow T_M/2$; $T_U \leftarrow cT_M$;
```

Results

The tree clustering algorithm can in principle be applied to any set of vector–valued data. As a test example we chose synthetic two–dimensional data and

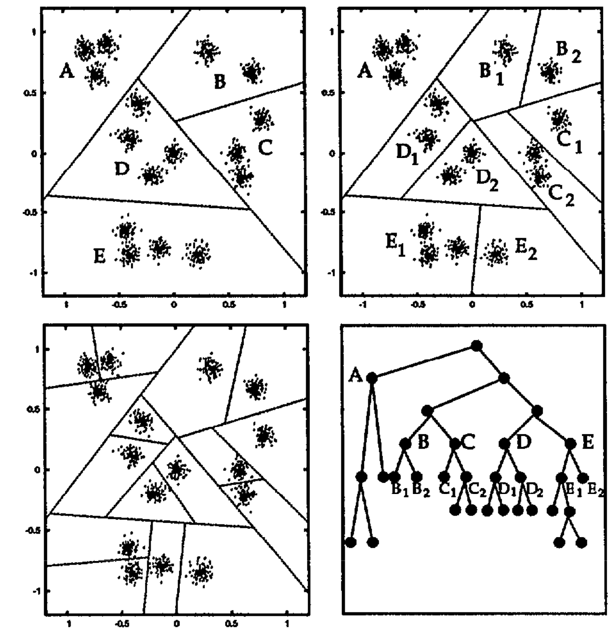

Figure 1: Hierarchical clustering of artificial two-dimensional data from 16 isotropic Gaussian modes. Displayed are partitionings with $K = 4, 9$ and $K = 16$ clusters.

real world data from multispectral LANDSAT images with seven spectral channels. The results on the synthetic data at different levels are shown in Fig. 1, together with a representation of the final tree topology. The obtained hierarchical data partitioning retrieves the structure of the generating source. Fig. 2 shows the hierarchical split of the LANDSAT data. Regions which correspond to particular clusters are grey-scale coded and can be identified with the help of external information. The split between meadows and other areas occurs first in our example. Further down in the hierarchy, a separation of forest from urban areas can be observed. The hierarchy is stable for runs with different initial conditions.

References

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees.* Belmont, California: Wadsworth Intern. Group.

Buhmann, J., and Hofmann, T. 1994. A maximum entropy approach to pairwise data clustering. In *Proc. Intern. Conf. on Pattern Recognition, Jerusalem*, volume II, 207–212. IEEE Computer Society Press.

Buhmann, J., and Kühnel, H. 1993. Vector quantization with complexity costs. *IEEE Transactions on Information Theory* 39(4):1133–1145.

Dayan, P.; Hinton, G.; and Zemel, R. 1995. The Helmholtz machine. *Neural Computation* 7(5):889–904.

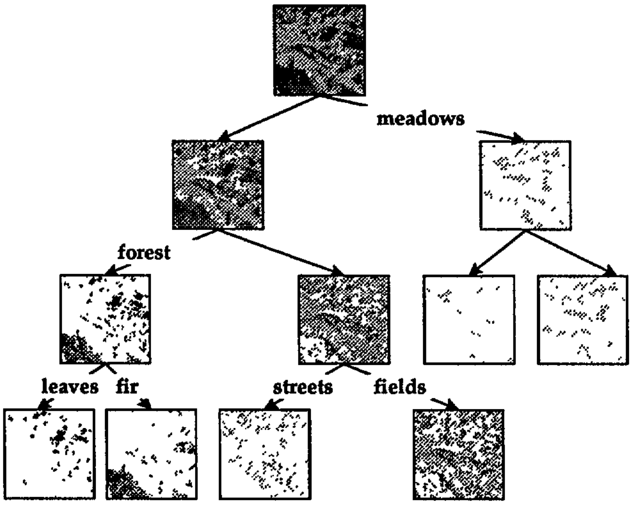

Figure 2: Hierarchical clustering of a LANDSAT multispectral image with 7 channels. Hierarchical clustering reveals the underlying relationship between regions. Only the top levels of the tree are displayed.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statist. Soc. Ser. B* 39:1–38.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis.* New York: Wiley.

Gold, S., and Rangarajan, A. 1996. A graduated assignment algorithm for graph matching. *IEEE PAMI.* in press.

Jordan, M., and Jacobs, R. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6(2):181–214.

Kirkpatrick, S.; Gelatt, C.; and Vecchi, M. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Kosowsky, J., and Yuille, A. 1994. The invisble hand algorithm: Solving the assignment problem with statistical physics. *Neural Computation* 7(3):477–490.

Miller, D., and Rose, K. 1994. A non-greedy approach to tree-structured clustering. *Pattern Recognition Letters* 15(7):683–690.

Miller, D., and Rose, K. 1996. Hierarchical, unsupervised learning with growing via phase transitions. to appear.

Rose, K.; Gurewitz, E.; and Fox, G. 1990. Statistical mechanics and phase transition in clustering. *Phy. Rev. Lett.* 65:945–948.

Sinkhorn, R. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.* 35:876–879.

Yuille, A. 1990. Generalized deformable models, statistical physics, and matching problems. *Neural Computation* 2(1):1–24.