An Unsupervised Clustering Method using the Entropy Minimization

Gintautas Palubinskas
Deutsches Zentrum für Luft- und Raumfahrt (DLR) e.V.
Deutsches Fernerkundungsdatenzentrum Oberpfaffenhofen
82234 Wessling, Germany
Gintautas.Palubinskas@dlr.de

Xavier Descombes
I.N.R.I.A.
2004 route des Lucioles, BP 93
06902 Sophia Antipolis, Cedex, France
Xavier.Descombes@sophia.inria.fr

Frithjof Kruggel
Max Planck Institute of Cognitive Neuroscience
Inselstraße 22-26
04103 Leipzig, Germany
kruggel@cns.mpg.de

Abstract

We address the problem of unsupervised *clustering* using a Bayesian framework. *The entropy is considered to define a prior and enables us to overcome the* problem *of defining a* priori *the number of clusters and an initialization of their centers. A deterministic algorithm derived from the standard k-means algorithm is proposed and* compared *with simulated annealing algorithms. The robustness of the proposed method is shown on a magnetic* resonance *(MR) images database containing 65 volumetric (3D) images.*

1. Introduction

Unsupervised clustering methods such as popular ones: k-means, fuzzy c-means and the maximum likelihood with expectation maximization require an initialization of the number of clusters and of the cluster centers. Various measures [1, 3, 7] were proposed to find out the number of clusters automatically in a dataset. All of them are based on the statistical characteristics of clusters (variance, a priori probabilities and the difference of cluster centers) and are data dependent. Some criteria issued from the information theory have been proposed. The Minimum description length criterion evaluates a compromise between the likelihood of the classification and the complexity of the model [5]. The prior concerning the complexity of the model is not really adapted to image modeling and the theoretical values for the hyper-parameter do not provide satisfactory results. Empirical values for the hyper-parameter are also data dependent.

Herein we propose to embed the clustering problem into a Bayesian framework to automatically detect the number of clusters. The prior of the proposed model is derived from the entropy. Some automatic thresholding methods have been proposed using entropy either by maximizing the information between two clusters derived from Renyi's entropy [6] or by minimizing the cross entropy [8]. We consider the clustering problem where we have to reduce the complexity of the grey level description. We therefore minimize the entropy associated with the clustering histogram. To this prior a Gaussian likelihood term is added.

2. The Information model

Denote an image by X = {xi, ... $x_j, \ldots x_N\}$ where the subscripts j refer to coordinates of the lattice L and the x_j to as the grey values. A clustering is defined by a partitioning the grey level set C = $\{C_i, i = 1, \ldots, k\}$. A partition of the image corresponds to a cluster of the image (classification) defined by the region $S_i = \{j \in L | x_j \in C_i\}$. Therefore, a clustering can be obtained either by partitioning the grey level set or by partitioning the image itself.

Usual clustering algorithms search for the partition which minimizes a distance between the data and the classification. However, the number of clusters is fixed in the optimization function and is therefore required. In this paper, we propose to add a prior on the minimized function to include the number of clusters as a variable in the minimized function and to estimate it.

Denote by Y = $\{y_1, \ldots . y_N\}$ the classified image. A

1051-4651/98 $10.00 © 1998 IEEE

classified image is obtained by maximizing:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \propto P(X|Y)P(Y), \quad (1)$$

where $P(X|Y)$ is the likelihood and $P(Y)$ is the prior model. We assume that the data are independent conditionally to Y:

$$P(X|Y) = \exp\left(\sum_{j \in L} \ln p(x_j|y_j)\right). \quad (2)$$

In the proposed approach we consider that the number of clusters is unknown. Therefore we first consider one cluster for each grey level. To reduce this number of clusters we have to sharpen the histogram associated with the clustering. We propose to minimize the entropy of the classified image histogram. The maximum of the entropy is achieved for uniform images and the entropy decreases as the number of levels with probability 0 increases. Therefore, when adding the likelihood term to an entropy prior we reach a compromise between the likelihood of the classification and the simplicity of the description (a few number of clusters). We define the prior as an exponentially shaped probability:

$$P(Y) \propto \exp\left(\alpha_E \sum_{i=1}^{K} p_i \ln p_i\right), \quad (3)$$

where $p_i = \#S_i/N$ is the prior probability of cluster i, K is the number of grey levels (a priori clusters).

The I-model (Information model) [2] is defined by the posterior probability:

$$P(Y|X) \propto \exp(-U), \quad (4)$$

$$U = -\sum_{j \in L} \ln p(x_j|y_j) - \alpha_E \sum_{i=1}^{K} p_i \ln p_i, \quad (5)$$

where α_E refers to as the hyper-parameter.

3. K-means algorithm

We generalize the k-means algorithm to minimize the energy associated with the I-model. We assume that the x_j are Gaussian distributed with mean values $y_i, i = 1, ..., k$ and constant variation for clusters:

$$p(x_j|y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_j - y_i)^2}{2\sigma^2}\right). \quad (6)$$

From equations (5,6) we have:

$$U = \sum_{i=1}^{k} \sum_{j \in S_i} \left[\frac{(x_j - y_i)^2}{2\sigma^2} - \frac{\alpha_E}{N} \ln p_i\right]. \quad (7)$$

When $\alpha_E = 0, U$ is the energy of k-means (KM) clustering algorithm. So the standard KM algorithm is a particular case of the k-means (KME) which uses the entropy and minimizes the energy from equation (7).

The KME clustering algorithm allows us to avoid the initialization of the number of clusters and cluster centers. We start the iterative procedure with the number of clusters equal to the number of grey values in an image and with the cluster means equal to corresponding grey values. Initial values for a priori probabilities of clusters are computed from the image histogram. During iterations some clusters vanish (p_i = 0) due to the entropy term. The final number of clusters depends on the α_E value. For small values of α_E we get a lot of clusters whereas for large values we get very few clusters. We propose a heuristic estimate for α_E by assuming an equilibrium between the entropy and the likelihood in equation (7): $\alpha_E = -AN/2 \ln p_i$, where p_i = l/M, M is the expected number of clusters. So the α_E is normalized to the data. The proportionality constant A is used to control the number of the clusters. For A = 1, a reasonable clustering can be achieved for most of the images. Smaller value of A result in a finer clustering, larger value - coarser clustering. This iterative clustering procedure is very fast as it works on the image histogram.

4. Results: KME versus KM

Visual evaluation of the results of clustering was performed on 65 high-resolution volumetric MR brain datasets. One slice of the sample image is presented in Figure l(top). This image was clustered with KM into 3 clusters, with KME (A = 1.5), and with KM into 6 clusters. When comparing with the hand segmentation, KM with 3 clusters underestimates white matter (WM) because of the intensity inhomogeneities in an image (left). KME produces 6 clusters: 2 for cerebrospinal fluid, 1 for grey matter and 3 for WM (right). So it respects intensity inhomogeneities in an 3D image. A similar result is achieved with KM and a known number of clusters equal to 6. Tests on the 65 images showed that KME is a robust procedure as we use the same value for A = 1.5 whereas the number of clusters found varies from 4 to 7.

5. Reaching the MAP criterion

We propose two versions of the simulated annealing (SA) to optimize the energy equation (5) either in the grey level or in the image lattice space. The simulated annealing is a widely spread algorithm used to minimize cost functions (or energies) when deterministic algorithms (Gradient descent, Conjugate gradient,...) fail because of local minima. The convergence of this stochastic algorithm to the MAP criterion is proven in case of Markov Random Fields in [4].

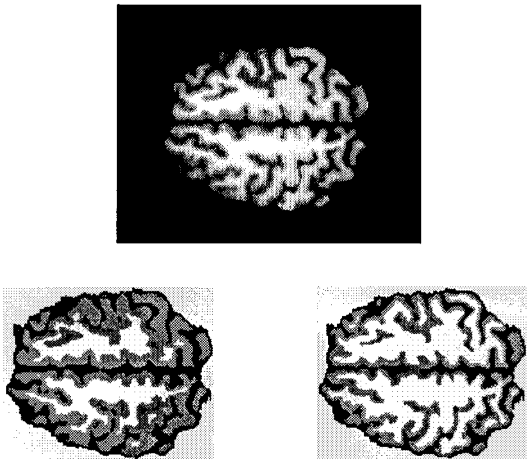

Figure 1. Sample MR image: T_1 slice (top), KM with 3 clusters (left), KME with A = 1.5 (right).

6. Results: SA versus KME

The results are summarized in table 1 for a sample image from our data base. The first remark concerning the energy is that the SA performing on the grey level space leads to worse results than the KME algorithm - despite of the theoretical result concerning the convergence of the SA to the global minimum of the energy. The SA on the grey level space tends to provide too many clusters. A visual inspection shows that the extra-clusters contain few pixels.

Using a SA on the lattice space allows us to use a pixel-wise updating scheme. The obtained results are very close to those obtained with the KME algorithm and the run dependency is negligible. However, the results are not better when comparing the energy than those obtained with the deterministic KME algorithm and the required CPU time is much greater. Nevertheless, this algorithm is still interesting as it allows us to incorporate other priors such as Markov Random Fields (MRFs).

7. Conclusion

We have compared a deterministic algorithm derived from the k-means algorithm (KME) and two simulated annealing defined on the grey level space and on the lattice space. Despite the theoretical properties of the simulated annealing, we obtain better results with the KME algorithm than with the SA defined on the grey level space. The SA defined on the lattice space provides results very close to those obtained with the KME algorithm but requires more CPU time. However, this SA algorithm provides a general framework for adding other priors such as MRFs to get regularizing properties. This last point is currently under study. Our investigations concern the optimization of the energy composed on a global prior (the entropy term) and a local prior defined by the MRF.

Algorithm	Energy	# clusters
KME	1.37	6
SA histogram 1	1.47	11
SA histogram 2	1.48	14
SA histogram 3	1.40	9
SA lattice 1	1.38	5
SA lattice 2	1.37	5
SA lattice 3	1.39	6

Table 1. Comparison of KME and the two SA on sample MR image.

References

[1] S.-T. Bow. Pattern Recognition and Image *Preprocessing*. Marcel Dekker, New York, 1992.

[2] X. Descombes and F. Kruggel. A markov pixon information approach for low level image description. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997. (under review).

[3] I. Gath and A. Geva. Unsupervised optimal fuzzy clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 11(7):773–781, 1989.

[4] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE* Trans. on Pattern Analysis and Machine Intelligence, 6(6):721–741, 1984.

[5] Z. Liang, R. Jaszczak, and R. Coleman. Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing. *IEEE* Trans. on Nuclear Science, 39:1126-1133, 1992.

[6] P. Sahoo, C. Wilkins, and J. Yeager. Threshold selection using Renyi's entropy. Pattern Recognition, 30(1):71–84, 1997.

[7] X. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE* Trans. on *Pattern* Analysis and *Machine Intelligence,* 13(8):841–847, 1991.

[8] Y. Zimmer, R. Tepper, and S. Akselrod. A two-dimensional extension of minimum cross entropy thresholding for the segmentation of ultrasound images. *Ultrasound* in *Med.* and *Biol.*,22(9):1183–1190, 1996.