

ESCM Algorithm: Application to multimodal image segmentation

Preetham Reddy Armoor

Department of Mathematics
University of Applied Sciences, Mittweida

11 January 2017

Outline

- 1 Introduction
- 2 Preliminaries
- 3 ESCM
- 4 Example
- 5 Conclusion

C-Means (CM) is clustering method based on numerical approximation to the maximum likelihood technique for the estimation of probability mixtures parameters.

It is subject to the problem of trapping in local optima of its objective function.

Evolutionary Strategy to the search for the global minimum of the C-Means objective function .

ESCM algorithm is applied to the clustering step of an interactive system for the segmentation of multimodal medical volumes.

To aggregate voxels with similar properties in the different diagnostic imaging volumes, clustering is performed in a multidimensional space where each independent dimension is a particular volumetric image. Sets of voxels with similar intensity values can be defined within the whole multimodal medical volume. These sets of voxels can then be used to delineate regions of interest, that is to make a segmentation of the multimodal volumetric image

Parametric Learning Approach to Clustering

Let $X = \{X_k | X_k \in R^d, k = 1, \dots, n\}$ be a set of unlabeled random sampled vectors $X_k = (x_{1k}, \dots, x_{dk})$ or training set,
 $Y = \{Y_j | Y_j \in R^d, j = 1, \dots, c\}$ be the set of centers of clusters (or classes) ω_j .

Parametric Learning Approach to Clustering

Following a parametric learning approach, we make the following assumptions:

- the samples come from a known number of c classes $\omega_j, j \in 1, \dots, c$.
- the a priori probabilities $P(\omega_j)$ (i.e. the probability of drawing patterns of class ω_j from X) are known.
- the form of class-conditional probabilities densities $p(x|\omega_j, \Theta_j)$ (i.e. the probability density of sample X_k inside class ω_j) are known, while the vectors of parameters Θ_j are unknown.

Note that the third assumption reduces the clustering problem to the problem of estimation of the vectors Θ_j (parametric learning).

Parametric Learning Approach to Clustering

- we assume that samples are obtained by selecting a class ω_j and then selecting a pattern x according to the probability law $p(x|\omega_j, \Theta_j)$

$$p(x|\Theta) = \sum_{j=1}^c p(x|\omega_j, \Theta_j)p(\omega_j)$$

where $\Theta = (\Theta_1, \dots, \Theta_c)$

$p(x|\omega_j, \Theta_j)$ are called the component densities

$p(\omega_j)$ are called the mixing parameters

Parametric statistics method for estimating the parameter vector Θ is based on maximum likelihood

It assumes that the parameter vector Θ is fixed but unknown

Parametric Learning Approach to Clustering

- The likelihood of the training set X is the joint density

$$p(X|\Theta) = \prod_{k=1}^n p(\mathbf{x}_k|\Theta)$$

Then the maximum likelihood estimate $\hat{\Theta}$ is that value of Θ that maximizes the likelihood of the observed training set X

Parametric Learning Approach to Clustering

- If $p(X|\Theta)$ is a differentiable function of Θ , maximizing the logarithm of the likelihood, we can obtain the following conditions for the maximum-likelihood estimate $\hat{\Theta}_j$:

$$\sum_{k=1}^n p(\omega_j | \mathbf{x}_k, \hat{\Theta}) \nabla_{\hat{\Theta}_j} \log(p(\mathbf{x}_k | \omega_j, \hat{\Theta}_j)) = 0 \forall j$$

if the a priori class probabilities $P(\omega_j)$ are also unknown, the clustering problem can be faced as the constrained maximization of the likelihood $p(X|\Theta)$ over Θ and $P(\omega_j)$ subject to the constraints:
 $P(\omega_j) \geq 0$ and $\sum_{j=1}^c p(\omega_j) = 1$

Parametric Learning Approach to Clustering

- If $p(X|\Theta)$ is differentiable and the a priori probabilities estimate $\hat{P}(\omega_j) \neq 0$ for any j , then $\hat{P}(\omega_j)$ and $\hat{\Theta}_j$ must satisfy:

$$\hat{P}(\omega_j) = \frac{1}{n} \sum_{k=1}^n \hat{P}(\omega_j | \mathbf{x}_k, \hat{\Theta})$$

and

$$\sum_{k=1}^n \hat{p}(\omega_j | \mathbf{x}_k, \hat{\Theta}) \nabla_{\hat{\Theta}_j} \log(p(\mathbf{x}_k | \omega_j, \hat{\Theta}_j)) = 0$$

where

$$\hat{p}(\omega_j | \mathbf{x}_k, \hat{\Theta}) = \frac{p(\mathbf{x}_k | \omega_j, \hat{\Theta}_j) \hat{P}(\omega_j)}{\sum_{h=1}^c p(\mathbf{x}_k | \omega_h, \hat{\Theta}_h) \hat{P}(\omega_h)}$$

Parametric Learning Approach to Clustering

- Let we assume now that the component densities are multivariate normal, i.e.:

$$p(\mathbf{x}_k | \omega_i, \hat{\Theta}_j) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \mathbf{y}_j)^t \Sigma_j^{-1} (\mathbf{x}_k - \mathbf{y}_j)\right]$$

Parametric Learning Approach to Clustering

- The local-maximum-likelihood estimate for $P(\omega_j)$ is the same while

$$\hat{\mathbf{y}}_j = \frac{\sum_{k=1}^n \hat{p}(\omega_j | \mathbf{x}_k, \hat{\Theta}_j) \mathbf{x}_k}{\sum_{k=1}^n \hat{p}(\omega_j | \mathbf{x}_k, \hat{\Theta}_j)}$$

$$\hat{\Sigma}_j = \frac{\sum_{k=1}^n \hat{p}(\omega_j | \mathbf{x}_k, \hat{\Theta}_j) (\mathbf{x}_k - \hat{\mathbf{Y}}_j)^t (\mathbf{x}_k - \hat{\mathbf{Y}}_j)}{\sum_{k=1}^n \hat{p}(\omega_j | \mathbf{x}_k, \hat{\Theta}_j)}$$

$$\hat{p}(\omega_j | \mathbf{x}_k, \hat{\Theta}_j) = \frac{|\hat{\Sigma}_j|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x}_k - \hat{\mathbf{Y}}_j)^t \hat{\Sigma}_j^{-1} (\mathbf{x}_k - \hat{\mathbf{Y}}_j)] \hat{p}(\omega_j)}{\sum_{h=1}^c |\hat{\Sigma}_h|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x}_k - \hat{\mathbf{Y}}_h)^t \hat{\Sigma}_h^{-1} (\mathbf{x}_k - \hat{\mathbf{Y}}_h)] \hat{p}(\omega_h)}$$

Parametric Learning Approach to Clustering

An efficient implementation of the previous procedure is based on the following approximation

$$\hat{p}(\omega_j | \mathbf{x}_k, \hat{\Theta}_j) = \begin{cases} 1 & \text{if } D_j(\mathbf{x}_k) = \min_{1 \leq j \leq C} D_j(\mathbf{x}_k) \\ 0 & \text{otherwise} \end{cases}$$

where $D_j(\mathbf{x}_i)$ is a *local cost function* or *distortion* measure and in many cases can be assumed as the scaled *Mahalanobis distance* $M_j(\mathbf{x}_k)$,

$$M_j^2(\mathbf{x}_k) \equiv \left| \sum_j \right|^{\frac{1}{d}} (\mathbf{x}_k - \hat{\mathbf{Y}}_j)^t \sum_j^{-1} (\mathbf{x}_k - \hat{\mathbf{Y}}_j)$$

It is worth noting that the usage of the Mahalanobis distance still involves a heavy computational overhead.

Parametric Learning Approach to Clustering

while maximizes the likelihood of the training set, minimizes at the same time a global error function J_w defined as the expectation of the squared local cost function:

$$J_w \equiv \langle D^2 \rangle = \sum_{k=1}^n \sum_{j=1}^c u_{jk} D_j^2(\mathbf{x}_k)$$

where $u_{jk} \equiv P(\omega_j | \mathbf{x}_k)$ or, in general, a membership value of pattern $\mathbf{x}_k (k = \{1, \dots, n\})$ to cluster $\omega_j (j = \{1, \dots, c\})$. some intrinsic problems. In particular, it is subject to the problem of trapping in local minima of J_w

implementation of the CM using the Euclidean distance

- assign the number of clusters and the tolerance ϵ_1 for the stop criterion
- initialize the centers of clusters
- do until any center changes less than ϵ_1
- assign the samples to the clusters with smaller Euclidean distance
- recalculate the centers
- end do

Evolutionary Computation methods for continuous parameter optimization problems founded on the model of organic evolution

During each generation (iteration of the ES algorithm) a population of individuals (potential solutions) is evolved to produce new solutions. Only the highest-fit solutions survive to become parents for the next generation.

In biological terms, the genetic encoding for an individual is called genotype. New genotypes are created from existing ones by modifying the genetic material. The interaction of a genotype with its environment induces an observed response called phenotype.

Individuals in the population are composed by object variables and strategy parameters. In basic ES, an individual is represented as a vector $\mathbf{a} = (x_1, \dots, x_n, \sigma_1, \dots, \sigma_n) \in R^{2n}$ consisting of n object variables and their corresponding n standard deviations for individual mutations.

There are two variants of an ES.

- The multi-membered ES plus strategies (denoted as $(\mu + \lambda)$ -ES) In $(\mu + \lambda)$ -ES μ parents create $\lambda \geq 1$ offspring individuals by means of recombination and mutation. The μ best parents and offspring are selected to form the next population.
- multi-membered ES comma strategies (denoted as (μ, λ) -ES) In (μ, λ) -ES, with $\lambda > \mu \geq 1$ the μ best individuals are selected from offspring only.

Recombination (or crossover) in ES is performed on individuals of the population

- discrete recombination: the components of two parents are selected at random from either the first or the second parent to form an offspring individual.
- intermediate recombination: offspring components are somewhere between the corresponding components of the parents.
- global and discrete recombination: one parent is selected and fixed and for each component a second parent is selected anew from the population to determine the component values using discrete recombination.
- global and intermediate recombination: one parent is selected and fixed and for each component a second parent is selected anew from the population to determine the component values using intermediate recombination.
- no recombination.

For mutations each x_j is mutated by adding an individual, $(0, \sigma_j)$ -normally distributed random number. The σ_j themselves are also subject to mutation and recombination a complete mutation step $m(a) = a'$ is obtained by the following equations:

$$S = \exp(N(0, \tau))$$

$$\sigma_j' = \sigma_j \cdot \exp(N_j(0, \tau')). S$$

$$x_j' = x_j + N_j(0, \sigma_j')$$

Mutation is performed on the σ_j by multiplication with two log-normally distributed factors, one individual factor, sampled for each σ_j ($\tau' = 1/\sqrt{2\sqrt{n}}$), and one common factor $s(\tau = 1/\sqrt{2n})$, sampled once per individual. This way, a scaling of mutations along the coordinate axes can be learned by the algorithm itself, without an exogenous control of the σ_j .

Selection based on the rank of fitness. It is called also an extinctive selection, as $\lambda - \mu$ worst individuals are definitively excluded from contribution offspring to the next generation.

$(\mu + \lambda)$ -ES is elitist and therefore, while performance is monotonously improved, the implemented search is local and unable to deal with changing environment.

(μ, λ) -ES enables the search algorithm to escape from local optima, to follow a moving optimum, to deal with noisy objective function and to self adapt strategy parameters effectively.

The ratio $\frac{\mu}{\lambda}$ is named the degree of extinctiveness and is linked to the probability to locate the global optimum. If it is large there is a high convergence reliability, whereas if it is small there is a high convergence velocity.

Evolution Strategy based C-Means (ESCM) algorithm

- ① assign μ , λ the number of clusters, and the threshold ϵ_2
- ② initialize the population
- ③ evaluate J_w for each individual
- ④ do until $\frac{\Delta J_w^{best}}{J_w^{best}}$ is greater than ϵ_2
- ⑤ count1=0
 - (a) while count1 less than μ
 - ① count1++
 - ② select by rank two individuals for mating
 - ③ order consistently the centers of clusters in both selected individuals using algorithm RI

Evolution Strategy based C-Means (ESCM) algorithm

- ④ crossover object variables (discrete recombination)
- ⑤ crossover strategy parameters (intermediate recombination)
- ⑥ mutate individual
 - (b) end do
 - (c) evaluate J_w for each individual
 - (d) select the μ fittest individuals for next population
- ⑥ end do

- 1 compile the matrix of distances M among centers of clusters of the two individuals
- 2 $\text{count2}=0$
- 3 while count2 less than c
- 4 $\text{count2}++$;
- 5 find the minimal item of the matrix
- 6 assign the same index to both centers of clusters in the two individuals
- 7 delete the corresponding row and column in the matrix of distances M
- 8 end do

if we want to reduce the interference of big blobs to the localization of the centers of small clusters, it is straightforward to change in the algorithm J_w with the following scaled global error function J_s :

$$J_s \equiv \sum_{j=1}^c \frac{1}{C_j} \sum_{k=1}^n u_{jk} D_j^2(\mathbf{x}_k)$$

where C_j is the cardinality of cluster ω_j

segmentation of MMV

Multimodal volumes can be derived from sets of such different diagnostic volumes by spatial coregistration of volumes in order to fully correlate complementary information about the same patient.

The extraction of such volumes or other entities of interest from imaging data is named segmentation

A supervised approach has two major drawbacks:

- it is very time-consuming (especially for large volumes), as it requires the labeling of prototypical samples needed for applying the generalization process. Even if the number of clusters is predefined, a careful manual labeling of voxels in the training set belonging with certainty to the different clusters is not trivial, especially when it concerns multimodal data sets
- heavy biases may be introduced by physicians unskilled or fatigued due to the large inter-user and intra-user variability generally observed when manual labeling is performed.

clustering based inference approach to MMV segmentation

Unsupervised methods may fully exploit the implicit multidimensional structure of data and make clustering of the feature space independent from the users definition of training regions due to their self-organizing approach.

multimodal volume may be defined by the spatial registration of a set of d different imaging volumes. As a consequence, its voxels are associated with an array of d values, each representing the intensity of a single modality in a voxel. the d different intensity values related to the voxel in such multimodal volume can be viewed as the coordinates of the voxel within a d -dimensional feature space .

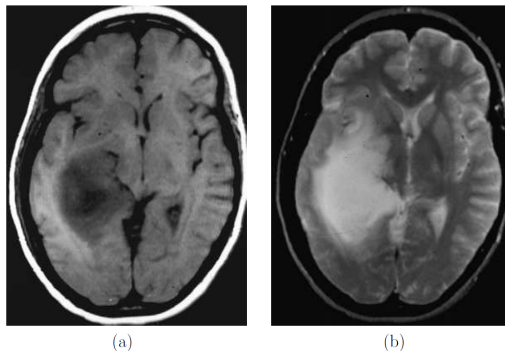


Figure: T1-weighted (a) and T2-weighted (b) MRI images of a patient with glioblastoma multiforme in the right temporal lobe.

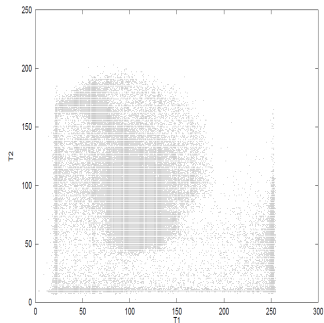


Figure: Feature space (T2 versus T1) obtained from the MRI images

implementation of clustering algorithms used in the experimental analysis.

- The CM uses $c = 7$, $\epsilon_1 = .01$, centers of clusters are initialized at random, and convergence is noticed in 10-15 fast iterations.
- For the ESCM using J_w according to the $\frac{\mu}{\lambda} = 1/7$ rule we selected $\mu = 10$ and $\lambda = 70$. Moreover, we initialized $c = 7$, $\epsilon_2 = .005$, and the centers of clusters at random. We implemented the selection by rank using a linear probability distribution with negative slope, while the intermediate recombination is implemented as the average of components of parents.

- The implementation of ESCM using J_s is identical to the previous one, with the obvious exception of the objective function. A typical plot of J_s^{best} Using $\frac{\Delta J_s^{best}}{J_s^{best}} \leq \epsilon_2$ as the stop condition, the ESCM ends in 15 iteration.

Results

the results of the CM algorithm can be seen. CM almost correctly defines scalp and white matter. Nevertheless it produces mistakes in classification of gray matter and edema in the left side of brain, and especially is not able to separate tumor, necrosis and CSF

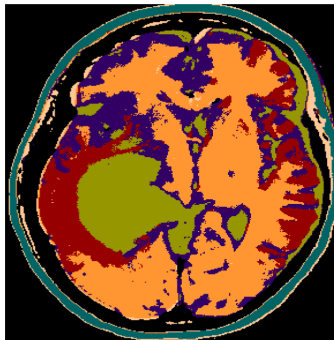


Figure: Segmentation obtained by the CM algorithm with 7 clusters.

Results

ESCM with the standard cost function J_w results to be largely more stable than CM with respect to the positions of centroids and to the extension of clusters in the feature space.

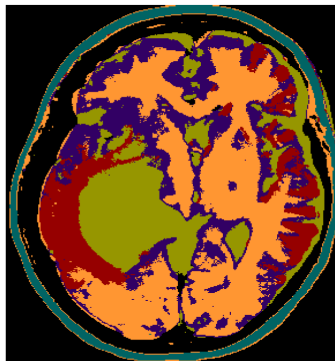


Figure: Segmentation obtained by the ESCM algorithm using J_w and with 7 clusters.

Results

ESCM with J_s dramatically improve. we can notice that, in comparison with CM, and with the basic version of ESCM, it correctly distinguishes between tumor and CSF, and within the tumor region is able to find the necrosis region. Correct definition of scalp and white matter and misclassification in the left side of the brain remains as from CM.

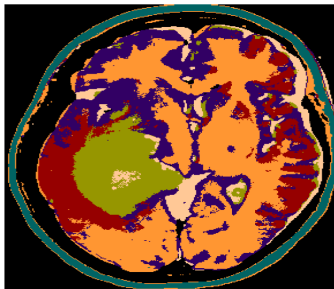




Figure: Segmentation obtained by the ESCM algorithm using J_s and with 7 clusters.

Conclusion

- The C-Means (CM) is an efficient approximation of the maximum likelihood procedure for estimating the centers of clusters, shows some intrinsic problems. In particular, it is subject to the problem of trapping in local minima of its objective function J_w
- The ESCM is based on a (μ, λ) -ES strategy where the object variables of genotypes are the centers of clusters. The implementation of the (μ, λ) -ES strategy is quite standard, but before mixing object variables of parents using discrete recombination crossover, they are re-indexed, in such a way centers with same index are likely to correspond to the same cluster with the straightforward change of J_w with the scaled global error function J_s it is possible to reduce the interference of big blobs to the localization of the centers of small clusters.

-  Francesco Masulli, Anna Maria Massone and Andrea Schenone.
Evolution Strategy for the C-Means Algorithm: Application to multimodal image segmentation.
February 24, 2013.
-  R.O. Duda and P.E. Hart
Pattern Classification and Scene Analysis.
Wiley, New York, 1973.