

# A probabilistic model for classifying segmented images

Liang Wu, Predrag Neskovic and Leon Cooper

*Department of Physics, Institute for Brain and Neural Systems  
Brown University, Providence, RI 02912*

## Abstract

*In this work we introduce a probabilistic model for classifying segmented images. The proposed classifier is very general and it can deal both with images that were segmented with deterministic algorithms, such as the k-means algorithm, and with probabilistic clustering approaches, such as the Hidden Markov Random Field (HMRF) algorithm. Similarly, our model can be used on either binary images or on images that contain multiple clustering labels as well as on images with any cluster boundaries (sharp, fuzzy or irregular). We tested our classifier on real fMRI images and showed that it outperforms the region-based Maximum Likelihood k-means classifier. Furthermore, we showed that higher classification rates are obtained when the images are segmented using a probabilistic HMRF algorithm compared to deterministic k-means method.*

## 1. Introduction

Classification and segmentation are among the most commonly used methods for analyzing images. They are especially useful for medical applications, for analyzing and decoding fMRI images [3], and for understanding the functional properties of the human brain [4]. However, these two methods are usually used independently despite the fact that they could potentially benefit each other. For example, voxel activations are extremely noisy and utilizing information from the neighboring voxels, e.g. using clusters instead of voxels, could help deal with noise. However, classifying segmented images presents its own challenges: how should one evaluate similarity between two different segmentations? Which cluster from one segmentation corresponds to which cluster from another segmentation, and how should one compare the shapes of different regions?

Probably the simplest solution is to consider only two cluster labels, e.g. a background label and a region

of interest (ROI) label, and then focus only on the ROIs and compare their shapes across different images [5]. Although this approach can be useful in some clinical applications [5], the analysis is limited to only ROIs and its associated binary assignment (active versus non-active) is not sufficient. For example, in many situations the same region of the brain might be involved in several functional activities at the same time and therefore the same voxel should have multiple labels (belong to several clusters) and this assignment should be probabilistic.

In this work, we address the previous problems using a Bayesian approach and introduce a Categorical Distribution-based (CD) classifier. Instead of focusing only on specific ROI, our model can classify segmented fMRI images of the whole brain. The proposed classifier is voxel-based (uses cluster assignment information from each voxel) as opposed to region-based (e.g. representing clusters with density functions) and it can therefore easily deal with any kind of region boundaries (e.g. sharp, fuzzy or irregular). The algorithm is very general in that it can utilize both deterministic and probabilistic voxel to clusters assignments, and it can also deal with clusters with multiple labels.

To segment images, we implemented a deterministic k-means algorithm and a probabilistic Hidden Markov Random Field (HMRF) finite mixture model [7]. The advantage of the HMRF model is that it imposes spatial constraints on the neighboring voxels which is biologically realistic assumption since neighboring voxels tend to have similar activations. In this work, we build a HMRF Dirichlet process mixture model and derive a collapsed Variational Bayesian (VB) approach [1] to integrate out the mixture weights.

We tested our model on real fMRI images and demonstrated that our classifier significantly outperforms the Maximum Likelihood k-means (ML-KMeans) classifier. Furthermore, we showed that higher classification rates are obtained when the images are segmented using a probabilistic HMRF approach compared to deterministic k-means method.

## 2 Categorical distribution-based classifier

The input to the classifier is a segmented image, which can be obtained using a number of clustering algorithms such as the k-means or the HMRF algorithm, and the objective is to classify the image into one of several classes. Each voxel of the segmented image is associated with a sequence of numbers that provide an assignment of the voxel to each of the  $K$  clusters. We call this sequence of numbers a clustering distribution vector (CDV). The assignment can be either fixed, in which case a voxel is assigned to only one cluster, or probabilistic, in which case the sequence represents probabilities of assigning a given voxel to each of  $K$  clusters. For example, if the CDV is obtained after maximizing a posterior or from other deterministic algorithms such as the expectation-maximization algorithm, then the assignment matrix  $v_{ik}$  is the indicator matrix and for some  $k^*$ ,  $v_{ik^*} = 1$ . However, if the CDV is obtained using a probabilistic clustering approach, such as the HMRF finite mixture model, the assignment for each voxel is  $v_i = \{p(c_i = 1), \dots, p(c_i = K)\}$ ,  $\sum_k v_{ik} = 1$ .

We will assume that there exists a true underlying distribution that assigns a voxel to each of the  $K$  clusters and that this distribution is class specific. For example, the probability of assigning the  $i$ -th voxel to the  $k$ -th cluster when the voxel belongs to the segmented image obtained from the class  $y$  is denoted as  $\eta_{ik}^y$ . For simplicity, we will omit the subscript  $y$ . If as a result of the deterministic clustering procedure the  $i$ -th voxel is assigned to  $k^*$ th cluster,  $v_{ik^*} = 1$ , then the distribution of  $\mathbf{v}$  is given as  $p(\mathbf{v}_i|\boldsymbol{\eta}_i) = p(v_{ik^*} = 1|\boldsymbol{\eta}_i) = \eta_{ik^*}$ , which can be viewed as a generalization of the Bernoulli distribution to more than two outcomes (or the categorical distribution). When the cluster assignment consist of probabilities instead integers, we generalize the distribution as  $p(\mathbf{v}_i|\boldsymbol{\eta}_i) \sim \prod_{k=1}^K \eta_{ik}^{v_{ik}}$ , which reduces to  $\eta_{ik^*}$  in a deterministic case  $v_{ik^*} = 1$ .

In principle, the parameter  $\boldsymbol{\eta}$  can be estimated from the training data and then used to calculate the class conditional likelihood function (CCLF) of observing a specific segmentation. If we denote the segmentation of a test image as  $U = \{\mathbf{u}_1, \dots, \mathbf{u}_N\}$ , where  $N$  is the number of voxels in the image, then the likelihood that a particular cluster assignment is generated by the distribution defined by the parameter of the class  $y$  is  $p(U|\boldsymbol{\eta}^y)$ . However, this likelihood can also be calculated without explicitly estimating the parameter  $\boldsymbol{\eta}$ . In this work we use Bayesian approach and integrate over the parameter. If we denote the training examples of a given class  $y$  as  $V^j = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}^j$ , then the class conditional likelihood can be written as  $p(U|V^1, \dots, V^M)$ , where  $M$  is the number of training

examples. More specifically, the likelihood function  $p(U|V^1, \dots, V^M)$  is given by,

$$\begin{aligned} p(U|V^1, \dots, V^M) &= \prod_{i=1}^N p(\mathbf{u}_i|\mathbf{v}_i^1, \dots, \mathbf{v}_i^M) \\ &= \prod_{i=1}^N \int d\boldsymbol{\eta}_i p(\mathbf{u}_i|\boldsymbol{\eta}_i) p(\boldsymbol{\eta}_i|\mathbf{v}_i^1, \dots, \mathbf{v}_i^M) \end{aligned} \quad (1)$$

The unknown parameters  $\boldsymbol{\eta}_i$  of the  $i$ th voxel can be estimated from the  $M$  training samples,

$$p(\boldsymbol{\eta}_i|\mathbf{v}_i^1, \dots, \mathbf{v}_i^M) = \frac{p(\mathbf{v}_i^1, \dots, \mathbf{v}_i^M|\boldsymbol{\eta}_i)p(\boldsymbol{\eta}_i)}{\int d\boldsymbol{\eta}_i p(\mathbf{v}_i^1, \dots, \mathbf{v}_i^M|\boldsymbol{\eta}_i)p(\boldsymbol{\eta}_i)} \quad (2)$$

where  $p(\mathbf{v}_i^1, \dots, \mathbf{v}_i^M|\boldsymbol{\eta}_i) = \prod_{j=1}^M p(\mathbf{v}_i^j|\boldsymbol{\eta}_i)$ . We choose for the prior  $p(\boldsymbol{\eta}_i)$  to be Dirichlet distribution

$$p(\boldsymbol{\eta}_i) = \frac{\Gamma(K\lambda)}{\Gamma(\lambda)^K} \prod_k \eta_{ik}^{\lambda-1} \quad (3)$$

and therefore our posterior will also have the form of Dirichlet distribution which will allow the exact calculation of the integral. Note that in calculating the integral over  $\boldsymbol{\eta}_i$ , one has to include the constraint that  $\eta_{ik} \geq 0$ ,  $\sum_k \eta_{ik} = 1$ . Knowing that,

$$\int d\boldsymbol{\eta}_i \prod_{k=1}^K \eta_{ik}^{\lambda_k-1} = \frac{\prod_{k=1}^K \Gamma(\lambda_k)}{\Gamma(\sum_{k=1}^K \lambda_k)} \quad (4)$$

Eq (1) can be integrated as,

$$\begin{aligned} p(U|V^1, \dots, V^M) &= \prod_{i=1}^N \frac{\prod_{k=1}^K \Gamma(s_{ik} + u_{ik} + \lambda) \Gamma(M + K\lambda)}{\prod_{k=1}^K \Gamma(s_{ik} + \lambda) \Gamma(M + K\lambda + 1)} \end{aligned} \quad (5)$$

where  $s_{ik} = \sum_{j=1}^M v_{ik}^j$ . If one uses a deterministic clustering, such as the k-means, then the quantity  $s_{ik}$  represents the number of times the  $i$ -th voxel has been assigned to the  $k$ -th cluster across all the training images.

### 2.1 The correspondence problem

The result of a clustering algorithm is a segmented image where all the voxels from one region (or cluster) have the same label. However, the labeling of each region is essentially random and therefore even the clusters representing two exact segmentations can have different labels. We assume that all the images from one class will have similar segmentations and we want to find an assignment between clusters of two images that reflects this property. The problem of which cluster

from one image should be assigned to which cluster of another image is known as the correspondence problem. It is a combinatorial optimization problem and it can be solved using the Hungarian algorithm in polynomial time [2]. The algorithm models an assignment problem as a  $n \times m$  cost matrix, where each element represents the cost of assigning the  $k$ th cluster in one image to the  $j$ th cluster in a different image. The algorithm performs minimization on the elements of the cost matrix.

We define the distance between the  $k$ th cluster from image 1 and the  $l$ th cluster from image 2 as

$$d_{k,l} = \sum_i |v_{ik}^1 - v_{il}^2|. \quad (6)$$

The objective is to find a permutation of the clusters from the second image that produces the highest overlap among the clusters of the two images. This is equivalent to minimizing the following objective function over the one to one cluster permutation mapping  $p : \{1, \dots, K\} \rightarrow \{1, \dots, K\}$ ,

$$f(p) = \sum_k d_{k,p(k)} = \sum_{i,k} |v_{ik}^1 - v_{ip(k)}^2|, \quad (7)$$

where  $k$  goes over all clusters, and  $i$  over all voxels.

To learn the parameter  $s_{ik}$  for a given class, we solve the correspondence among all the training images of that class. To calculate the CCLF given by Eq. (5), we then solve the correspondence problem between the given test image and each class of training images.

## 2.2 HMRF Dirichlet finite mixture model

In this section, we review the finite Dirichlet Gaussian mixture model with spatial dependences modeled by hidden Markov random fields and then derive the collapsed VB approach.

Suppose the observation  $\mathbf{x} = \{\mathbf{x}_i, i = 1, \dots, N\}$  can be described by a Gaussian mixture model with hidden states  $\mathbf{c} = \{c_i\}$ ,

$$p(\mathbf{x}, \mathbf{c}, \Theta) = \prod_i^N p(x_i | \Theta_{c_i}) p(\mathbf{c}) p(\Theta) \quad (8)$$

where  $p(x_i | \Theta_k) = \mathcal{N}(\mu_k, \Gamma_k)$  follows a Gaussian distribution. Each data point is generated from one of the  $K$  hidden Gaussian models whose parameters are drawn from the parameter space  $\Theta = \{\mu_1, \Gamma_1, \dots, \mu_K, \Gamma_K\}$  according to the hidden indicator variable (or class label)  $\mathbf{c} = \{c_1, c_2, \dots, c_N\}$ .

The hidden cluster variable follows the Generalized Bernoulli distribution and has a spatial constraint mod-

eled by a Markov random field,

$$p(\mathbf{c} | \boldsymbol{\pi}) \propto \prod_{k=1}^K \pi_k^{m_k} \exp(-\gamma U(\mathbf{c})) \quad (9)$$

where  $m_k$  is the number of samples whose hidden state is the  $k$ th state and  $U$  is an energy function of the random field  $\mathbf{c}$ . It is defined as the interaction between neighboring data points  $U(\mathbf{c}) = \sum_{i=1}^N \sum_{j \in \text{neighbors of } i} (1 - \delta(c_i, c_j))$ . To lower the energy function, the neighboring data points are desired to have the same class labels.

Conjugate prior is Gaussian for the means  $p(\boldsymbol{\mu}_k | \boldsymbol{\Gamma}_k) = \mathcal{N}(\boldsymbol{\rho}_0, \beta_0 \boldsymbol{\Gamma}_k)$ , and Wishart for the precisions,  $p(\boldsymbol{\Gamma}) = \mathcal{W}(\nu_0, \boldsymbol{\Phi}_0)$ . The conjugate prior for  $\boldsymbol{\pi}$  is Dirichlet  $p(\boldsymbol{\pi}) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^K \pi_k^{\alpha-1}$ .

After integrating over  $\boldsymbol{\pi}$ , we obtain the marginal distribution of  $p(\mathbf{c})$  and  $p(\mathbf{c}_{-i})$  from which, the distribution of the class label of one data point conditioned on other points can be calculated as,

$$p(c_i = k | \mathbf{c}_{-i}) = \frac{p(\mathbf{c})}{p(\mathbf{c}_{-i})} \propto \frac{m_{-i,k} + \alpha}{N - 1 + K\alpha} \exp(-\gamma U_i) \quad (10)$$

where  $m_{-i,k}$  is the number of data points assigned to class  $k$  and excluding the point  $i$ .

## 2.3 Variational Bayesian inference algorithm

In this section, we review the variational Bayesian inference algorithm and introduce the closed analytical form for the collapsed VB solutions to HMRF model.

In Bayesian framework, one needs to estimate the joint posterior  $p(\mathbf{c}, \Theta | \mathbf{x})$  to infer class labels  $\mathbf{c}$  and the corresponding parameters  $\Theta$ . VB approach considers an approximate variational posterior  $q(\mathbf{c}, \Theta)$  with a factorized form  $q(\mathbf{c}, \Theta) = \prod_{i=1}^N q(c_i | \mathbf{x}) \prod_{k=1}^T q(\Theta_k)$ .  $q$  is computed by minimizing the Kullback Leibler (KL) divergence between the variational posterior  $q(\mathbf{c}, \Theta)$  and the real joint posterior  $p(\mathbf{c}, \Theta | \mathbf{x})$ .

$$\begin{aligned} D(q(\mathbf{c}, \Theta) || p(\mathbf{c}, \Theta | \mathbf{x})) \\ = - \sum_{\mathbf{c}} \int d\Theta q(\mathbf{c}, \Theta) \log \frac{p(\mathbf{x}, \mathbf{c}, \Theta)}{q(\mathbf{c}, \Theta)} + \log p(\mathbf{x}) \end{aligned}$$

By minimizing the KL divergence, we get the update equations,

$$q(\Theta_k) \propto p(\Theta_k) \exp(\mathbb{E}_{\mathbf{c}}[\log p(\mathbf{x} | \Theta, \mathbf{c})]) \quad (11)$$

$$q(c_i = k) \propto \exp(\mathbb{E}_{\mathbf{c}_{-i}}[\log p(c_i | \mathbf{c}_{-i})] + \mathbb{E}_{\Theta_k}[\log p(\mathbf{x}_i | \Theta_k)]) \quad (12)$$

Subject	1	2	3	4	5
CD (%)	68.3	58.3	65.0	61.7	63.3
ML-KMeans	51.7	50.0	48.3	50.0	51.7

**Table 1. Classification rates for the CD and ML-KMeans classifiers on images segmented with HMRF method.**

Parameter posteriors are functionally identical to the priors but have different parameter values, i.e.,  $q(\boldsymbol{\mu}_k|\boldsymbol{\Gamma}_k) = \mathcal{N}(\boldsymbol{\rho}_k, \beta_k \boldsymbol{\Gamma}_k)$ ,  $q(\boldsymbol{\Gamma}_k) = \mathcal{W}(\nu_k, \boldsymbol{\Phi}_k)$ , see [6] for details about updating the parameter values.

Since there is no conjugate prior for the collapsed distribution of the cluster variable in Eq. (10), the first term of Eq. (12) needs careful treatment as shown in [6].

### 3 Results

In this section, we evaluate the performance of our classifier on fMRI images that were segmented using both the deterministic (k-means) and probabilistic (HMRF) algorithms. We compare the classification rates of the CD classifier and the ML-KMeans classifier (see [5] for more detail). In all experiments  $k = 5$ .

The fMRI data used in this work were recorded while the subjects were looking at face images (either disgusted or fearful) and trying to detect their emotional expressions [4]. Each face was shown for very brief period of time, 70ms, making the task very difficult. The number of fMRI images for each class was 30. The activation of each voxel was represented with 7 points and in order to reduce the dimensionality of the data, we modeled the voxel activation curve by fitting it to a polynomial function of the second order [6]. As a result, activation of each voxel was represented not with a scalar but with a 3D vector.

In the first experiment, we compare the classification rates of the CD and ML-KMeans classifiers on the fMRI images that were clustered using HMRF model and the collapsed VB inference algorithm. As one can see from Table. 3, the classification rates when using the HMRF clustering are consistently higher than the rates when the images were clustered with the k-means algorithm. In the second experiment we used as a clustering method the k-means algorithm. As shown in Table. 3, the CD classifier outperforms the ML-KMeans classifier for most subjects.

Subject	1	2	3	4	5
CD(%)	58.3	55.0	48.3	51.7	65.0
ML-KMeans(%)	53.3	50.0	51.7	50	48.3

**Table 2. Classification rates for the CD and ML-KMeans classifiers on images segmented with k-means method.**

### 4 Conclusions

In this work we introduced a new model for classifying segmented images and tested it on real fMRI images. We demonstrated that the CD classifier is very general in the sense that it can deal not only with images that were segmented with deterministic algorithms, such as the k-means algorithm, but also with probabilistic clustering approaches, such as the HMRF algorithm. We contrasted our classifier with ML-KMeans based classifier, and showed that it outperforms the ML-KMeans classifier on almost all subjects. Furthermore, we showed that our classifier can be used not only on binary images, but also on images that contain multiple clustering labels which can be of great importance when analyzing medical data.

### Acknowledgments

This work is supported in part by the ARO under contract W911NF-04-1-0357.

### References

- [1] M. W. Kenichi Kurihara and Y. W. Teh. Collapsed variational dirichlet process mixture models. In *IJCAI*, pages 2796–2801, 2007.
- [2] H. Kuhn. The hungarian method for the assignment problem. *Nav. Res. Log. Quart.*, 2:83–97, 1955.
- [3] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, and X. Wang. Learning to decode cognitive states from brain images. *Mach. Learn.*, 57:145–175, May 2004.
- [4] L. Pessoa and S. Padmala. Decoding near-threshold perception of fear from distributed single-trial brain activation. *Cereb. Cortex*, 17:691 – 701, Apr 2007.
- [5] D. Pokrajac, V. Megalioikonomou, A. Lazarevic, D. Kontos, and Z. Obradovic. Applying spatial distribution analysis techniques to classification of 3D medical images. *Art. Intel. in Med.*, 33:261–280, July 2005.
- [6] L. Wu, P. Neskovic, and L. Pessoa. Dirichlet process mixture model with spatial constraints. IBNS-TR-2007-02, Brown University, Providence, Rhode Island, 2007.
- [7] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imag.*, 20(1):45–57, January 2001.