AN OPTIMAL FEATURE SELECTION STRATEGY FOR FUZZY C-MEANS. APPLICATION TO LIP-TO-SKIN DISCRIMINATION

Mihaela Gordan[*], Costin Miron[*], Apostolos Georgakis[**]

[*]*Basis of Electronics Department, Technical University of Cluj-Napoca*
{mihag,miron}@bel.utcluj.ro
Constantin Daicoviciu Street, No. 15, Cluj-Napoca, RO-3400, Romania

[**] *Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, Box 451, GR-54006 Thessaloniki, Greece*
apostolos@zeus.csd.auth.gr

Abstract: One of the widely used algorithms for data clustering and image segmentation is the fuzzy c-means (FCM) algorithm. The objective function minimization in FCM guarantees the optimal data separation according to their feature values but there is no guarantee that this separation will match an available ground truth. However an optimal feature set can be found in respect to which the FCM classification matches the ground truth. We propose here a strategy for optimal feature set selection in FCM, making use of a small set of training data. This strategy will allow the selection of the best FCM "version" for a given image segmentation problem, in respect to the feature set. The strategy was developed for two-class segmentation and applied to the lip-to-skin discrimination problem in grey level mouth images. The experimental results show the improved performance of the resulting FCM segmentation as compared to FCM and to some semi-supervised FCM versions.

Key words: fuzzy c-means, feature selection, supervised clustering, lip detection

1. INTRODUCTION

In its standard form, fuzzy c-means (FCM) [1] is one of the widely used unsupervised clustering algorithms for image segmentation. The problem of selecting the best features to represent the data for a given application, so that the data considered to be similar according to some ground truth to have similar feature values and the data belonging to different classes to be strongly discriminated, is a general problem in clustering. Other researchers investigated this problem [2] and considered the possibility of selecting the optimal features to represent the data for a given application from a larger set than the minimal needed for a correct classification (actually computing as many features as possible). Then starting from a small number of features, the feature vector representing the data is increased by the inclusion of one feature at a time and for each new feature vector, an optimal classifier (in [2], an SVM) is trained and tested on the same data set. Only if a feature contributes significantly to the reduction of the test error it will be included in the final feature set. In this approach, only the possibility of crisp inclusion or rejection of a feature from the feature set is allowed; however in

practice, the features used to represent a data point can have different importance (not only 1) in describing the data. This corresponds to a "fuzzy" feature set, where every feature is assigned a degree of relative importance between 0 and 1, describing its significance in data grouping/separation; this is the approach we propose to select the optimal features for a data clustering application.

The selection strategy of the optimal features proposed here was developed to be used with the FCM algorithm and applied in particular to image segmentation in known classes (namely, lip region detection in mouth images only [3]). Since the optimal features are considered to be those that make the final classification result to match as close as possible an available ground truth, a set of training data is needed for the feature selection process. This a-priori information is usually available for a given application. Some researchers developed variants of FCM, where an available set of training data is used for supervising the segmentation results. These versions are known as semi-supervised versions of FCM. In the version proposed by Pedrycz [4], a set of available labeled data is included in the cost function to be minimized by the FCM algorithm as an error term proportional to the difference between the membership degrees of the data in FCM and the memberships given by labels [4]. A similar variant is proposed in [5], where the authors use a semi-supervised FCM on image data segmentation, with an extra-step where geometrical constraints are imposed to the image regions that are to be detected by FCM. Unlike these FCM versions that make use of learning to optimize the classifier's partition matrix under fixed features, we introduce the learning phase in the feature selection process, leaving the clustering algorithm itself unchanged. Thus the optimization strategy is kept simple. As the experimental results show, the final classifier performs very well on a given application of lip-to-skin discrimination.

2. THE STANDARD FUZZY C-MEANS ALGORITHM WITH MULTIPLE FEATURES

Let us consider a set of N data points represented by a number of F real-valued features, $\mathbf{x}_i = (x_{i1},...,x_{iF})^T$, $x_{if} \in \Re$, f=1,...,F; thus $\mathbf{x}_i \in \Re^F, i = 1,...,N$. Denoting by C the number of classes to which the data are to be assigned in some membership degree by the FCM algorithm, a membership matrix can be built, $\mathbf{U}[C \times N]$, with the u_{ji} element, j=1,...,C and i=1,...,N, representing the membership degree of the data $\mathbf{x}_i$ to the class j. **U** is constrained to be a fuzzy partition. The FCM algorithm aims to optimally define the C classes so that to minimize the uncertainty regarding the membership of every data $\mathbf{x}_i$, i=1,...,N, to each of the classes. This goal is achieved through the minimization of the cost function $J_m(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ji}^m \cdot d^2(\mathbf{x}_i, \mathbf{v}_j)$ where by **V** we denote the set of all prototypical class centers, $\mathbf{V}=\{\mathbf{v}_1,...,\mathbf{v}_C\}$, $\mathbf{v}_j \in \Re^F$, j=1,...,C. m is a parameter controlling the shape of the resulting clusters; the convergence of the optimization problem is guaranteed for m>1. d is a distance measure between the vectors $\mathbf{x}_i$ and $\mathbf{v}_j$. The minimization of $J_m(\mathbf{U},\mathbf{V})$ is solved iteratively, starting from an initial fuzzy partition matrix $\mathbf{U}^0$ or an initial set of prototypical class centers $\mathbf{V}^0$; the values of u_{ji} and of $\mathbf{v}_j$, j=1,...,C, are modified in each iteration to minimize $J_m(\mathbf{U},\mathbf{V})$. It can be mathematically proven that u_{ji} which minimizes $J_m(\mathbf{U},\mathbf{V})$, for a given **V**, and $\mathbf{v}_j$ which minimizes $J_m(\mathbf{U},\mathbf{V})$ for a given **U** are given by [1]:

$$u_{ji} = \left(\sum_{l=1}^{C} \left(\frac{d(x_i, v_j)}{d(x_i, v_l)} \right)^{\frac{2}{m-1}} \right)^{-1}, \; i=1,...,N, \; j=1,...,C \qquad v_j = \frac{\sum_{i=1}^{N} u_{ji} \mathbf{x}_i}{\sum_{i=1}^{N} u_{ji}} \qquad (1)$$

3. THE PROPOSED STRATEGY FOR OPTIMAL FEATURES SELECTION IN FUZZY C-MEANS BY SUPERVISED LEARNING

The minimization of the cost function $J_m(\mathbf{U},\mathbf{V})$ in the FCM algorithm guarantees the optimality of the resulting classifier in respect to the class selection and data assignment to the C classes, under a given set of features and a given distance metric d. However this does not guarantee that the class assignments match an available ground truth. Assuming we have a set of N data to be classified in C=2 classes with the FCM algorithm, and that based on some observations we know that the data $\{\mathbf{x}_1,...,\mathbf{x}_P\}$ should be assigned to C_1 and the data $\{\mathbf{x}_{P+1},...,\mathbf{x}_N\}$ should be assigned to C_2, we have no guarantee that after running FCM with C=2, this assignment will hold. Instead, we can assume that one can find an optimal feature set which will give the best match between the class assignments after FCM and the desired class assignments.

The proposed strategy for the optimal feature set selection in the FCM algorithm is as follows. Let us denote the feature set by $\mathbf{S}_F$, with $\mathbf{S}_F = \{f_1, f_2, ..., f_F\}$. Each $f_k \in \Re, k = 1,...,F,$ represents a feature value. To each feature $f_k, k = 1,...,F,$ we assign a degree of confidence in its significance for the data classification denoted by $w_k, w_k \in [0,1], k = 1,...,F$. Let us consider the set of N data $\mathbf{x}_i$, i=1,...,N, $\mathbf{x}_i \in \Re^F$, to be classified in C classes by the FCM. In the proposed approach each data vector $\mathbf{x}_i$ will have the form $\mathbf{x}_i = (x_{i1} \quad x_{i2} \quad ... \quad x_{iF})^T$, with $x_{ik} \in \Re, k = 1,...,F,$ given by:

$$x_{ik} = w_k \cdot f_{ik}, \quad k = 1,...,F; i = 1,...,N. \qquad (2)$$

Let us assume that, for a sub-set of the global data set $\{\mathbf{x}_i\}_{i=1,...,N}$, we have a-priori knowledge about its desired class assignments for a given application, and we aim to find the values w_k, k=1,...,F, that give the closest match in the classification output with these class assignments. We will discuss for simplicity the case of a binary classification, C=2. We denote the two classes by C_1 and C_2, the sub-set of data known to belong to C_1 as $\mathbf{S}_{C_1} = \{\mathbf{x}_1^{(C_1)},...,\mathbf{x}_{N_1}^{(C_1)}\}$, the sub-set of data known to belong to C_2 as $\mathbf{S}_{C_2} = \{\mathbf{x}_1^{(C_2)},...,\mathbf{x}_{N_2}^{(C_2)}\}$ and the total data set of N data as $\mathbf{S} = \{\mathbf{x}_1,...,\mathbf{x}_{N_1}, \mathbf{x}_{N_1+1},...,\mathbf{x}_{N_1+N_2}, \mathbf{x}_{N_1+N_2+1},...,\mathbf{x}_N\}$ with: $\mathbf{x}_i = \mathbf{x}_i^{(C_1)}$ for i=1,...,N_1; $\mathbf{x}_{i+N_1} = \mathbf{x}_i^{(C_2)}$ for i=1,...,N_2. The data from $\mathbf{S}_{C_1} \cup \mathbf{S}_{C_2}$ will be used for learning and validating the confidence degrees w_k, k=1,...,F, thus obtaining the optimal feature set to be used in FCM for classifying the data set. For a given FCM algorithm, with a fixed distance norm d and C=2, once the iterative algorithm of minimizing $J_m(\mathbf{U},\mathbf{V})$ has converged, we will have available the partition matrix $\mathbf{U}[2 \times N] = \{u_{ji}\}_{j=1,2; \; i=1,...,N}$. j=1 corresponds to data classification in C_1, whereas j=2 corresponds to data classification

in C_2. Then the classification error of this FCM in the data set $\mathbf{S}_{C_1} \cup \mathbf{S}_{C_2}$ can be described by the error function: $\varepsilon_{trn} = \varepsilon_{trn1} + \varepsilon_{trn2}$, where:

$$\varepsilon_{trn1} = \sum_{i=1}^{N_1} (1 - u_{1i}) + \sum_{i=N_1+1}^{N_1+N_2} u_{1i} \quad \text{and} \quad \varepsilon_{trn2} = \sum_{i=1}^{N_1} u_{2i} + \sum_{i=N_1+1}^{N_1+N_2} (1 - u_{2i}) \qquad (3)$$

To choose the optimal set of feature weights $\left\{w_k^{(opt)}\right\}$, $w_k^{(opt)} \in [0,1]$, $k = 1,...,F$, we consider a finite subset of possible values for every w_k, with an arbitrarily chosen step $\alpha \in [0,1]$, $w_k \in \mathbf{W}_\alpha = \{0; \alpha; 2\alpha; \ldots; 1\}$, k=1,...,F. The size of the step α should be chosen to ensure a tradeoff between the optimality of the feature set and the computational complexity, since the number of FCMs to be evaluated is $\left(\alpha^{-1} + 1\right)^F$. With every possible combination of weights from the set $\mathbf{W}_\alpha^F$ (the Cartesian product of $\mathbf{W}_\alpha$ in F dimensions), we will run a FCM algorithm on the data set $\left\{\mathbf{x}_i^{(q)}\right\}$, i=1,...,N, built according to Equation (2). After the convergence of each FCM, the value of $\varepsilon_{trn}^{(q)}$, q=1,..., $\left(\alpha^{-1} + 1\right)^F$, is evaluated as given in Equation (3). The optimal feature weights and thus the set of optimal features for the application will be the one minimizing ε_{trn}:

$$\left\{w_k^{(opt)}\right\} = \left\{ \left\{w_k^{(q)}\right\} \middle| \varepsilon_{trn}^{(q)} = \min_{r=1,...,\left(\alpha^{-1}+1\right)^F} \varepsilon_{trn}^{(r)} \right\} \qquad (4)$$

4. APPLICATION OF THE PROPOSED STRATEGY IN THE LIP-TO-SKIN DISCRIMINATION PROBLEM

We applied the procedure for selecting the optimal features to be used in the FCM algorithm for the lip region detection in gray level mouth images. We consider only images where the mouth is almost closed, so that the regions to be distinguished are the lip region and the skin region. The problem of lip region identification in gray level mouth images is not a trivial task, since these images usually present low contrast and variable illumination, and is an important task in visual speech recognition applications [6]. As we showed in a previous work [3], using FCM on the luminance component only for the segmentation of mouth images leads to poor results. When the image is divided into four sub-images (upper left, upper right, lower left and lower right) and a FCM is applied in every sub-image on pixel data including also geometrical features besides luminance (namely, the spatial position of the pixel), the algorithm leads to better results [3]. However even in this case the segmentation results are not always good, especially in images with low contrast. In these regions the geometrical features should get stronger importance; the importance should be "learned" for each sub-image. This goal can be achieved with the proposed feature selection strategy.

The procedure is described for the upper left sub-mouth image (as illustrated in Figure 1. (a)); for the remaining sub-images it should be applied in exactly the same fashion. Denoting by W_{UL} and H_{UL} – the upper left image width and height, we will have a set of $N = W_{UL} \cdot H_{UL}$ data to be segmented by the FCM algorithm into C=2 classes, namely, class C_1=Lips and class C_2=Skin. Each data $\mathbf{x}_i$ from the set, i=1,...,N, corresponding to an image pixel, will be described by two features, f_1 and f_2; f_1 is the luminance y and f_2 is the distance ρ of the pixel towards the origin; both f_1 and f_2 are

scaled between 0 and 100. With the notations from the previous section, F=2; $\mathbf{x}_i \in \Re^2$, i=1,...,N, and every $\mathbf{x}_i$ is given as: $\mathbf{x}_i = (x_{i1} \quad x_{i2})^T$, with the components $x_{i1} = w_1 \cdot y_i$; $x_{i2} = w_2 \cdot \rho_i$, where w_1 and w_2 represent the weights of f_1 and f_2, $w_1, w_2 \in [0,1]$ to be optimized. The set to be segmented by the FCM algorithm is $\mathbf{S} = \{x_i\}, i = 1,..., N$.

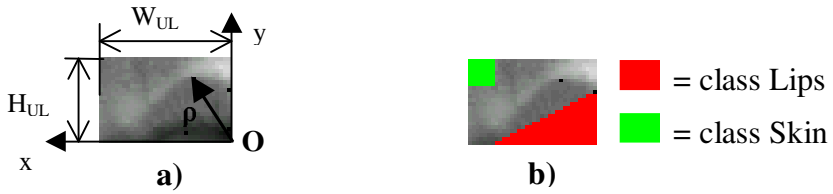

Figure 1. a) The upper left sub-image for an image in Tulips1 database, and the coordinate system for pixel position description; b) generation of the "training" set

The "training" data subset is built as follows: (1) the triangle with the corners (0, left corner of the mouth, middle of the upper lip) comprises the pixels belonging to C_1; (2) a rectangle of at least 7×7 pixels located in the upper left corner of the image belongs to C2. Thus we can construct the sets $\mathbf{S}_{C1}$ of size N_1 (the number of pixels inside the triangle) and $\mathbf{S}_{C2}$ of size N_2 (=49 in our case), as shown in Figure 1.(b). A label is attached to each pixel $\mathbf{x}_i$, denoted by l_i, $l_i \in \{-1;0;+1\}$, with the following meaning: if $\mathbf{x}_i \in \mathbf{S}_{C1} \Rightarrow l_i = +1$ ("training" data from class C_1); if $\mathbf{x}_i \in \mathbf{S}_{C2} \Rightarrow l_i = -1$ ("training" data from class C_2); if $\mathbf{x}_i \in \mathbf{S} \setminus \{\mathbf{S}_{C1} \cup \mathbf{S}_{C2}\} \Rightarrow l_i = 0$.

This label is not taken into account as a feature, but it is needed in the computation of the training error $\varepsilon_{trn}^{(q)}$ for the evaluation of each FCM. We choose α=0.2 (for a good precision-speed tradeoff), leading to a set of $(\alpha^{-1}+1)^F = 36$ combinations of weights. After running the 36 FCMs and computing the 36 training errors, the minimal error index will identify the optimal set of weights as $\{w_1^{(opt)}, w_2^{(opt)}\} \in \mathbf{W}_{0.2}^2$. The fuzzy partition of the "winning" FCM $_{(opt)}$ algorithm will be used to get the segmented image, by assigning each pixel $\mathbf{x}_i$ a different color, $\mathbf{cl}_1$ for C_1 if $u_{1i} > u_{2i}$, otherwise $\mathbf{cl}_2$ for C_2.

5. EXPERIMENTAL RESULTS

The validation of the proposed strategy was performed on the application of lip-to-skin discrimination. For our experiments, we use a set of closed mouth images from Tulips1 database [7]. We compare our results with those of the standard FCM, on the same feature set, with fixed weights assigned to the features independently on the processed image. In all the cases, the optimal weight vector differs from [1;1]; the set of weights [0.2;1] seems to be near optimal, but not for all the cases. Some examples of segmented images, both in standard FCM and in our method, are given in Figure 2.

The segmentation results obtained with the proposed strategy were also compared with results obtained with the standard FCM algorithm in our previous work [3] and with the results given by a semi-supervised version of FCM with geometric constraints [5]. As it can be seen from Figure 3, again our algorithm proves superior in the images under investigation. Thus the performance of the proposed solution is validated. The drawback of the proposed strategy is the extra-computational time needed in the optimal feature selection step. However for images with similar illumination and contrast the optimal weights have the same values; therefore the algorithm can be speeded up if a prior histogram analysis of the images to be segmented is performed.

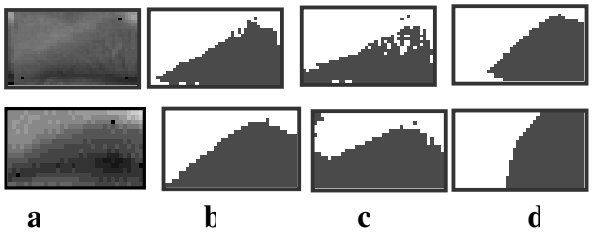

Figure 2. Comparison of FCM versions for different sets of weights, for two subjects in Tulips1 database: (a) original image; (b) optimal set ((w_1,w_2)=(0.6;0.4) for the first row;((w_1,w_2)=(1;0.4) for the second row); (c) (w_1,w_2)=(1;0.2) (fixed); (d) $w_1=w_2=1$

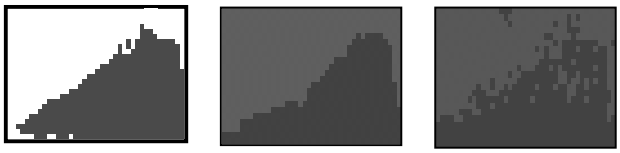

Figure 3. A comparison of FCM segmentation results in 3 algorithms: a) optimal feature set FCM; b) FCM with geometrical features from our previous work [3]; c) a semi-supervised FCM [5]

6. CONCLUSIONS

In this paper we proposed a strategy to select the optimal feature set in the FCM algorithm. Instead of allowing only the presence or absence of a certain feature, we assign weights between 0 and 1 to each feature, modeling in a fuzzy manner their relative importance in the segmentation. The experimental results validate the proposed method and encourage its further use. In our future work we will investigate solutions to reduce the computational burden introduced by running several FCM algorithms.

7. REFERENCES

1. J. C. Bezdek (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York

2. J. Park, H. Yae (2002), "Analysis of active feature selection in optic nerve data using labeled fuzzy C-means clustering", *Proc. IEEE Int. Conf. on Fuzzy Systems FUZZ-IEEE'02*, Vol. 2, pp.1580 – 1585

3. M. Gordan, C. Kotropoulos, A. Georgakis, I. Pitas (2002), "A New Fuzzy C-Means Based Segmentation Strategy. Applications to Lip Region Identification", *Proc. 2002 IEEE-TTTC A&QT-R 2002*, Cluj-Napoca, Romania

4. W. Pedrycz, J. Waletzky (1997), "Fuzzy clustering with partial supervision", *IEEE Trans. on Systems, Man and Cybernetics,* 27(5): 787-795.

5. J.C. Noordam, W.H.A.M. van den Broek (2000), "Geometrically Guided Fuzzy C-Means Clustering for Multivariate Image Segmentation", *Proc. Int. Conf. on Pattern Recognition*, pp. 462-465

6. I. Matthews, T. Cootes, S. Cox, R. Harvey, J. A. Bangham (1998), "Lipreading using shape, shading and scale", *Proc. Auditory-Visual Speech Processing*, Sydney, Australia, pp. 73-78

7. J. R. Movellan (1995), "Visual Speech Recognition with Stochastic Networks", *Advances in Neural Information Processing Systems, (G. Tesauro, D. Toruetzky, and T. Leen, Eds.)*, MIT Press, Cambridge, MA, Vol 7