Floating FCM

S.P.K .Karri

October 2016

1 Introduction

An ideal feature descriptor is aimed to represent samples of same class more closely and retain linear separability between samples from different classes. But this is not the case for practical examples where manifolds are common ie., in a feature space, samples of distinct classes are spatially closer and vice-verse. This opened up three types of solutions: Constructing data subjective feature descriptors, transforming (metric learning) feature space to a new space where samples are linearly separable and finally, classifiers capable of constructing nonlinear boundaries have been explored. As every approach have subjective merits and demerits proposed algorithm falls into third approach. Supervised and unsupervised classification are commonly used in biomedical image analysis depending on availability of expert labels for samples. Traditional unsupervised approaches creates boundaries between dissimilar samples through an intrinsic loss usually a distance metric where as supervised approaches take advantage of expert labels during loss computation. The proposed approach tries to combine the merits of both approaches and strikes similar to adaboost. The primary contributions of the paper being label guided clustering, no knowledge about number of clusters and adaboosting FCM to our knowledge.

FCM: Given a data matrix (features as columns and samples as rows), initially random samples are treated as cluster means and FCM iteratively computes the possible means of the cluster along with membership of each sample falling into each cluster.

Proposed approach: Given: Data matrix (features as columns and samples as rows) and corresponding labels (C number of classes) Output: 'C' mean sets where each set consists centroids of corresponding class. Initialization: Data matrix is partitioned into 'C' number of sets where each set have samples from same class and 'C' null sets are treated as mean sets.

c^{th} set is divided into two clusters with FCM and identified means are concatenated to c^{th} mean set. Upon repetition of the above step for all C classes, original data matrix is retrieved. Euclidean distance between individual samples of data matrix and means of individual mean set is computed. For each sample, the closest mean is identified and corresponding set indexes are treated as estimated labels of the samples. Identify the samples wrongly classified along

with corresponding true labels and partition them into 'C' sets where each set belongs to same class. The process is repeated until any part ions have more than 2 samples.

Why 2 clusters? why not just means per class ie., 1 cluster? Consider a two class toy example where samples with 2 dimensional features from different classes fall onto different concentric circles. The means of both class samples gets aligned so all samples needs to be allotted to one class (possibly with high number of samples) and class with low number of samples fall under error. During second loop of proposed algorithm the mean of wrongly classified samples stay as previous.

2 Algorithm

Proposed approach takes a data matrix X and label matrix L with C classes as input and return 'M' a set of matrices where each matrix is collection of centroids representing respective class. Functionaries in the algorithm are illustrated below FCM is a standard fuzzy C means clustering. Pairwise distance is standard function in matlab with eucledian distance min_means is min value along the axis of means ie row min_Idx gives the index of min value along each row

Algorithm 1 Floating FCM

procedure FFCM(X, L) ▷ X is data and L is labels
  $M^i \leftarrow \{\} \forall i = \{1, 2, .., C\}$ ▷ C is #classes
  $D^i \leftarrow \{\ \forall X^n \in$ class 'i' $\}$
  $iter \leftarrow 1; limit \leftarrow 10$
  for `iter:=1` $\rightarrow$ `10` **do**
    for `i:=1` $\rightarrow$ `C` **do**
      if `#samples in` D_t^i`>2` **then**
        $T \leftarrow$ FCM(D_t^i,clusters=2)
        $M^i \leftarrow \left\{M^i | T\right\}$
      end if
    end for
    for `i:=1` $\rightarrow$ `C` **do**
      $dist^i \leftarrow min\_means(pairwiseDist(D, M^i))$
    end for
    $\hat{L} \leftarrow$ min_Idx($dist^1, .., dist^C$)
    err $\leftarrow \hat{L} \neq L$
    for `i:=1` $\rightarrow$ `C` **do**
      $D^i \leftarrow \{\ \forall X^n \in$ class 'i' AND err(n) $\equiv$ TRUE $\}$
    end for
  end for
  return **M** ▷ a set of all M^i
end procedure

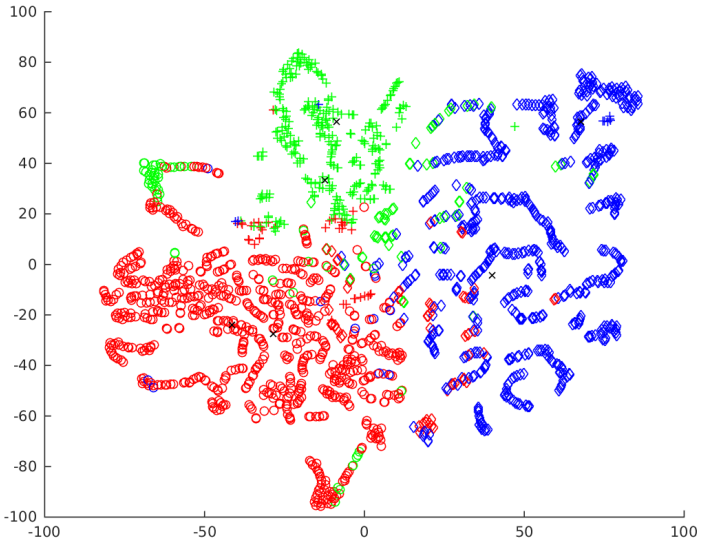

Figure 1: Iteration 1

3 Results

Duke OCT classification data is considered for experimentation. The dataset comprises of OCT images from three classes and 15 subjects per class. The images are processed and features are quantified according to the guidelines of srinivasan et. al., [1]. The train set includes 8 subjects per class and other subjects are treated as test set. The classification training performance along each *iter* is illustrated below. The samples are projected to 2D space thrugh t-SNE [2]. The shapes are true classes and colors are predicted classes. Circles, Diamonds and plus are true first, second and third classes respectively. Red, blue and Green are predicted first, second and third classes respectively. The black crosses are centroids.

During testing classical kmeans prediction is employed that is the class of closest mean vector is alloted to test sample. The test set accuracies for floating FCM and Linear SVM are 0.86 and 0.82 respectively

References

[1] Pratul P Srinivasan, Leo A Kim, Priyatham S Mettu, Scott W Cousins, Grant M Comer, Joseph A Izatt, and Sina Farsiu. Fully automated detection of diabetic macular edema and dry age-related macular degenera-

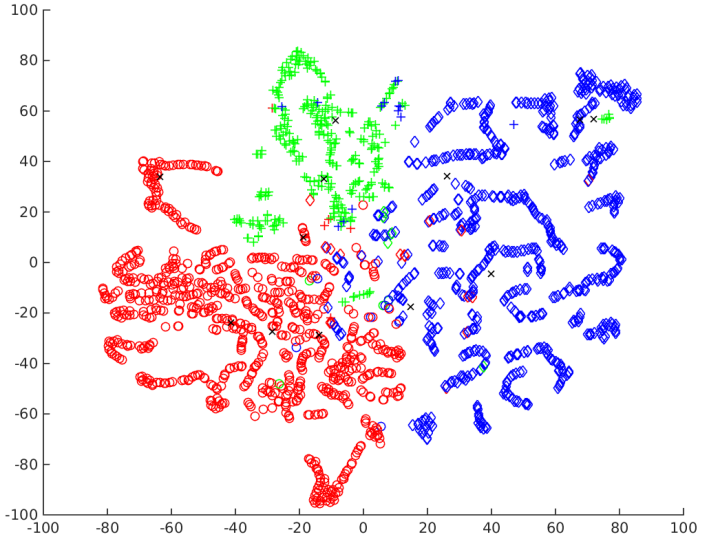

Figure 2: Iteration 2

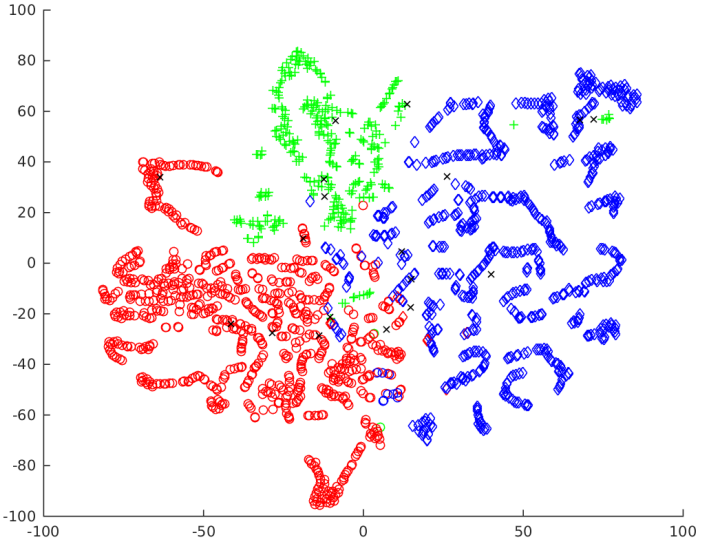

Figure 3: Iteration 3

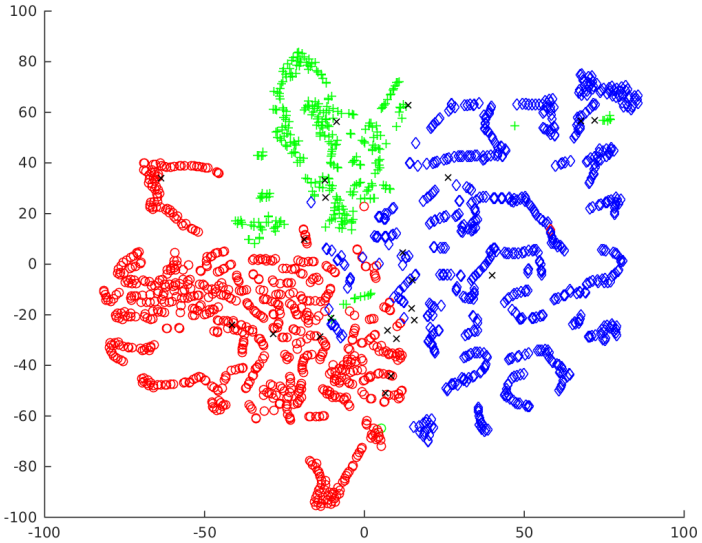

Figure 4: Iteration 10

tion from optical coherence tomography images. *Biomedical optics express*, 5(10):3568–3577, 2014.

[2] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *The Journal of Machine Learning Research*, 9(2579-2605):85, 2008.