Hierarchical clustering is very important in medical imaging as it helps in visualization of levels of dissimilarity among different cells and tissues. Semi-supervised clustering aims to cluster the data using the knowledge constraints, most research efforts in semi-supervised clustering use instance level must-link and cannot-link constraints, which cannot be used for hierarchical clustering since data samples are linked over different levels of hierarchy, we propose a 2-stage semi-supervised clustering model which gives the necessary hierarchical structure in the data but alleviates large memory requirement problem of hierarchical clustering, hierarchical clustering is done in 2nd stage using ultra metric transformation of dissimilarity matrix subject to triple wise relative constraints[3]. Our second method uses 3D-CNN to extract spectral-spatial features for classification. Experimental results demonstrate the efficiency and effectiveness of our proposed methods.

Medical data usually has high cost associated with annotating so using the labeled and unlabeled data through semi-supervised paradigm for classification is important. Medical HISI has a requirement for high accuracies and there is sufficient information content in the imagery to achieve high accuracies, as the hyper-spectral imagery has abundant information as each pixel vector is collected from large number of wavelengths across the electromagnetic spectrum.

**HSI DATASETS**

AVIRIS dataset collected by NASA using an airborne drone called Airborne Visible and Infrared Image Spectrometer, used as trade mark dataset to test classification models for HSI datasets.

HSI cell dataset contains 30 medical images with 4 classes, in each image (partially annotated), with each image pixel having 226 spectral bands corresponding to 900 wave numbers to 1800 wave numbers which are shown in Figure 1.1.

Class0: background + other cell structures
Class1: non-goblet cell epithelium (probably enterocytes)
Class2: stroma
Class3: goblet cell

**FEATURE EXTRACTION**

Semi-Supervised Non negative Matrix Factorization is used to extract the features. NMF builds a 2-factor decomposition of data matrix such that the cost function J = ||X - AS||^2 + ||L + (F - BS)|| is minimized, where norm is the frobenius norm, S is the feature matrix, X is data matrix, and Y is the label matrix, multiplicative updates are used iteratively for A, B and S.

**RESULTS**

Clustering of each image is done separately, each time 10% of labeled samples are used in the semi-supervised clustering by formulating triple wise relative constraints exhaustively, rest 20% are retained for classifying the clustering accuracy. Two stages in the clustering model restrict us from using F score as a metric for clustering accuracy. Accuracies over different images are weighted with ratio of number of test samples of that image to the total number of test samples and then mean value is calculated to get the average clustering accuracy of 97.16%

**REFERENCE**


**MODEL INSIGHTS**

Decision boundaries in a classification problem determines the class of the data samples, data samples close to decision boundaries are the ones that model the classifier and the samples that are deep inside the decision boundaries don’t effect the classifier to a great extent, so instead of storing and working on this large set of non crucial points we represent the sub-sets of points by a mean vector and covariance matrix. This information suffices to build a classifier and greatly reduces the memory requirements which is highly required in hierarchical clustering models.

This can also be understood as discretized version of probability density function estimation, covariance matrix models the density and the mean vector points to the location.