Active Learning via Multi-View and Local Proximity Co-Regularization for Hyperspectral Image Classification

Wei Di, Student Member, IEEE, and Melba M. Crawford, Fellow, IEEE

March 16, 2012

Outline

Data regularization based active learning

- Labeled data in D_L and unlabeled data D_U
- Regularization of data assumes an underlying smooth function
- A regularizer $\Re(f, \{D_L, D_U\})$ penalizes lack of smoothness of the classifier by labeled and unlabeled data

We adopt this idea into the active learning framework and define the loss at the τth query to reflect the overall inconsistency, i.e., the degree to which the current classifier violates the consistency assumption evaluated by all the unlabeled data $\mathbf{x}_{j,U} \in D_U^\tau$

$$Loss(\tau) = \frac{1}{N_U^\tau} \sum_{j=1}^{N_U^\tau} \Re(f^\tau, \mathbf{x}_{j,U}). \quad (2)$$

In AL, our purpose is to improve the learner by selecting n_Q new samples $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{n_Q}\} \in D_U^\tau$ for query each time to maximally reduce the loss

$$\arg \max_{\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{n_Q}\} \in D_U^\tau} Loss(\tau - 1) - Loss(\tau). \quad (3)$$

Data regularization II

The design of the regularizer is key to the success of the active learning strategy. According to the consistency assumption, it should favor the changes of $p(y|\mathbf{x})$ in regions with lower values of $p(\mathbf{x})$, where the decision boundary may be located. Samples which are in close proximity, but violate the consistency assumption, i.e., similar samples with higher confliction in terms of the conditional probability $p(y|\mathbf{x}_i)$ and $p(y|\mathbf{x}_j)$ should be queried first. Also, in the AL scenario, the information should be incorporated from both the labeled and unlabeled data, as well as the chosen classifier. Thus, we propose the following co-regularizer

$$\Re(\mathbf{x}) = \Re_{\mathrm{AMD}}\Re_{\mathrm{LIC}}(\mathbf{x}), \mathbf{x} \in \mathrm{D}_U. \tag{6}$$

The first factor is the multi-view adaptive maximum disagreement (MV-AMD) regularizer $\Re_{\mathrm{AMD}}$, and the second is the local inconsistency (LIC) regularizer $\Re_{\mathrm{LIC}}$ which represents the lack of smoothness measured on a local graph in the manifold space. Both are defined in Sections III and IV, respectively.

Multi-view generation for hyperspectral data

- In the multi-view setting, available attributes are decomposed into N_v disjoint sets: $X^1 \times \ldots \times X^{N_v}$
- It is assumed that each view is sufficient to learn the target concept
- The idea is to minimise disagreement between the outputs of different views
- In hyperspectral image analysis, views are built segmenting disjoint contiguous sub-band sets. Each sub-space is assumed to be highly correlated, but with little correlation to other sub-spaces.

Adaptive Maximum Disagreement Regularizer (AMD)

$$d\left(f_v^1, f_v^2, \ldots, f_v^{N_v}\right) \triangleq \frac{1}{N_U} \sum_{t=1}^{N_U} \left(\sum_{i=1}^{N_v} \sum_{j=i+1}^{N_v} 1_{\left(f_v^i(\mathbf{x}_t^i) \neq f_v^j(\mathbf{x}_t^j)\right)} \right) \tag{7}$$

To evaluate the contribution of each sample $\mathbf{x} \in D_U$, it is decomposed into a sample based distance measure as:

$$d\left(\mathbf{x}, f_v^1, f_v^2, \ldots, f_v^{N_v}\right) \triangleq \sum_{i=1}^{N_v} \sum_{j=i+1}^{N_v} 1_{\left(f_v^i(\mathbf{x}^i) \neq f_v^j(\mathbf{x}^j)\right)}. \tag{8}$$

This sample-wise distance represents individual uncertainty as it contributes to the overall confusion. To incorporate global information from the entire unlabeled data pool, we define the maximum disagreement $MaxD(\mathrm{D}_U)$ as

$$MaxD\left(D_U\right) = \max_{\mathbf{x} \in D_U} d\left(\mathbf{x}, f_v^1, f_v^2, \ldots, f_v^{N_v}\right). \tag{9}$$

Inconsistency by different views is then defined as

$$\varphi\left(1 - \frac{d(\mathbf{x}, f_v^1, f_v^2, \ldots, f_v^{N_v})}{MaxD}\right) \tag{10}$$

AMD II

$$\mathfrak{R}_{\mathrm{AMD}} = \delta \left(1 - \frac{d\left(\mathbf{x}, f_v^1, f_v^2, \ldots, f_v^{N_v}\right)}{MaxD} \right).$$

Regularization via generalized manifold space

- Manifold spaces$\simeq$projection of high dimensional data to a lower dimension space (?)

A k-nearest neighborhood graph for each unlabeled sample is then defined upon its closest k samples. The graph $G(N_{k,i}, E_{k,i})$ contains vertices (nodes) and edges where $N_{k,i}$ represents the set of nodes, and $E_{k,i}$ represents the edges of the graph. The predicted labels of the unlabeled data and the true labels from the labeled data are all represented on the graph. Let L_{ij} be the length of the edge from node i to node j, which represents the inconsistency between these two nodes

$$L_{ij} = w(\mathbf{x}_j)\|\mathbf{z}_i - \mathbf{z}_j\|_2 \left(1 - \delta(f(\mathbf{x}_i) - y'_j)\right) \qquad (13)$$

where

$$w(\mathbf{x}_j) = \begin{cases} w_L, & \mathbf{x}_j \in D_L \\ w_U, & \mathbf{x}_j \in D_U \end{cases}, y'_j = \begin{cases} y_j, & \mathbf{x}_j \in D_L \\ f(\mathbf{x}_j), & \mathbf{x}_j \in D_U \end{cases}$$

and $w(\mathbf{x}_j)$ is the weight used to differentiate the confidence assigned for the true label and the estimated label from the training data and the unlabeled data, respectively; generally $w_L > w_U$.