Automation of Biological Research

Mid-Project Report

Team Member:

Ruijian Wang(Andrew ID: ruijianw)
Jianhe Luo (Andrew ID: jianhel)

Project Background(Image classification):

We got 5120 unlabeled image instances, and each image instance is represented as a 1 * 26 vector, which means we do not have to do the feature extraction for the raw image data. And the goal for this project is to classify these image instances into 8 labels: (i) Endosomes; (ii) Lysosomes; (iii) Mitochondria; (iv) Peroxisomes; (v) Actin; (vi) Plasma Membrane; (vii) Microtubules; and (viii) Endoplasmic Reticulum

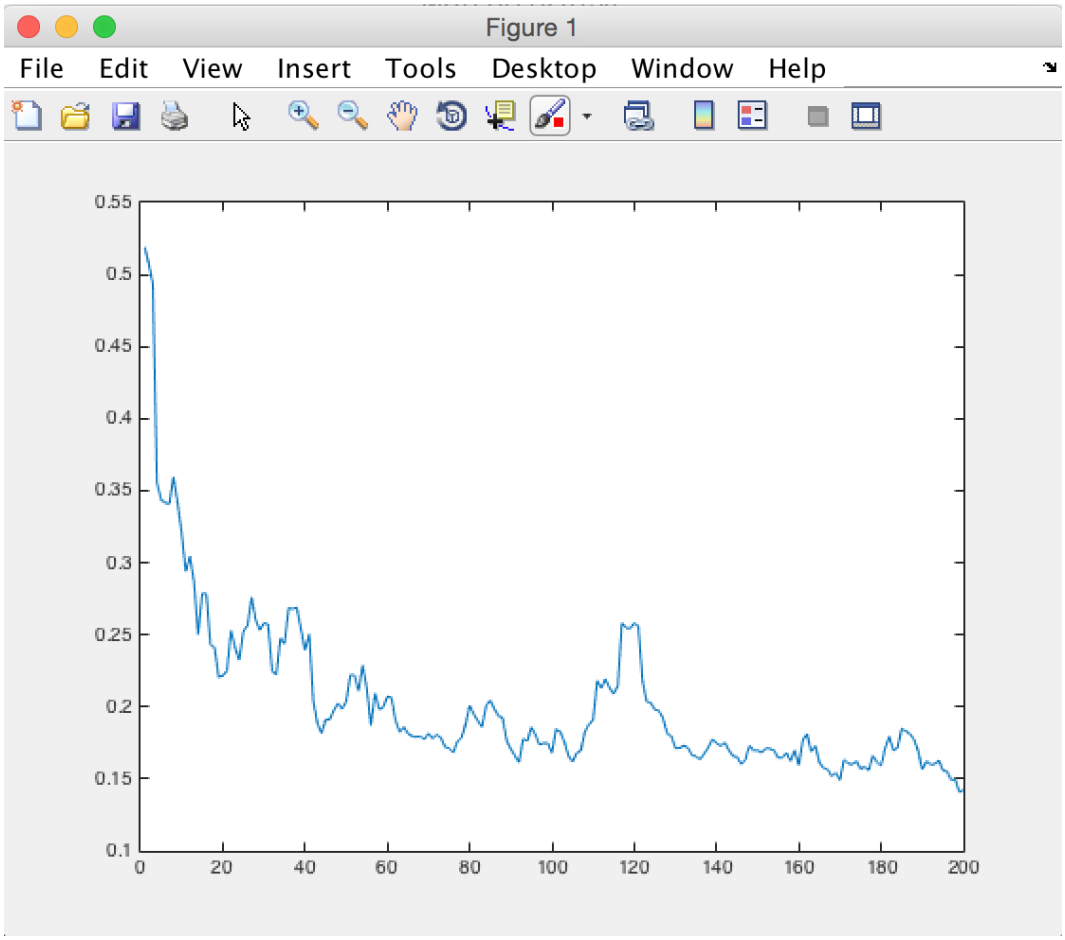
Preliminary Results:

For stream-based data access model, we implement DHM algorithm, and the followings are our specific steps for our algorithm.(we make some modifications for our process.)

1. We call nextImage() function to get 100 image instances and for these 100 image instances, we call oracle to get their true labels. Combining two of them, then we can get a 100 * 27 matrix. Then we get image instances, whose true label is i (i = 1, 2, 3, 4, 5, 6, 7, 8), and assign to data{i}; and then we randomly got same size as length(data{i}) from the rest of image instances and set their labels to be 0.After that we got 8 data{i}.
2. After the first step, we get 8 data{i}, each of them can be used as training data for binary classification. (if the image instance is belongs to current label, the label value will be i, else its label value will be 0.) For each data{i}, we can use svm classifier to train a original binary classification model. After that we get a 1 * 8 models vector, and each of this stand for a specific label classification model.
3. After step 1 and step 2, we can then train the our model. We call nextImage() to get our next image instance, and for each image instance, we make prediction about its label using all 8 models, so we got a 1 * 8 predictions vector, and each value of this vector stand for whether this image instance belongs to this label. For each image instance, we can make sure its label only when there is only one value in predictions vector which is not equal to 0. And then go to get next image instance. And in other case, we have to call the oracle to get the true label for this image instance and then retrain 8 svm binary classification model.
4. If 8 binary classification model can not make sure the image instance belongs to which label, we call oracle to get the true label at first. And then add the new_observation data point to 8 data{i}, and using 8 data{i} to retrain our binary classification model. After that use our 8 re-train models to make predictions for the all image instances. And then compute the loss.
5. After we using all of our budget to call oracle, we stop. and print out our loss vector. And the result is as following.

Algorithm 1 DHM Algorithm on image classification

procedure INITIALIZATION(*Data*)
    Sample 50 data points randomly, call oracles one 50 times to get their labels
    Categorize the data and create 8 sets training data for 8 binary classifiers
    for each training data **do**
        Train classifier and save the model
    end for
end procedure
procedure ACTIVE LEARNING(*Data*)
    for each data point in *Data* **do**
        Get predictions from 8 binary classifiers
        if none of the classifier gives positive prediction **then**
            Add this data point to training set, re-train 8 binary classifiers
            Assign label 0 to this data point ▷ 0 represents not recognized, recognized labels' range should be 1-8
        else if there is only one classifier gives a positive result **then**
            assign the positive classifier's label to this data point
        else ▷ if there is more than one classifier gives the positive result
            pick the first positive classifier's label for this data point.
            Add this data point to training set, re-train 8 binary classifiers
        end if
    end for
end procedure



Problems and solution:

1. Problem: for DH Algorithm, we are still trying to figure out how to evaluate purity of each node and how to estimate the majority labels in each cluster.
   Solution: Talk with the instructors to understand the DH algorithm better or switch to another pool based active learning Algorithm.
2. Problem: for DHM algorithm, what we learn and do in homework about DHM is binary classification. And in this image classification problem, we have to classify 8 labels.
   Solution: We transform our model from 8-labels classification to 2-labels classification. So we define 8 models, each stand for specific label.
3. Problem: for DHM algorithm, we have to figure out the selection strategy about when should we call the oracle.
   Solution: Every time we get a new image instance, we use 8 model to predict the label of this image, and we get a predictions vector. When we have only one value in predictions vector is not equal to 0, we can make clear prediction about its label, in other case, we have to call oracle and get new_observation, retrain the 8 model and compute the loss.
4. There is no label 2 in the true label file.
   Solution: replace label 1 with 2, and replace label 0 with 1. After this modification, we can use the label as index when processing the data.

Future Plan:

1. For DHM algorithm, we could use 8 binary classifiers to help choose the informative points and use one non-binary classifier to classify all models.
2. For DHM algorithm, the label selection strategy when 8 binary classifiers give the result needs to be refined. For now, we just use very naive selection method. In future, we may use some extra information from the classifier when it does the classification job, for example, posterior probability, cost and so on.
3. Look into more pool based active learning algorithm in case we need to switch to a new one from DH.
4. In DH algorithm, If there is no good way to evaluate the purity of each subtree, we might have to randomly choose the subtree to do pruning.

Reference:

Dasgupta, Sanjoy, and Daniel Hsu. "Hierarchical sampling for active learning."*Proceedings of the 25th international conference on Machine learning*. ACM, 2008.