Image Classification Project Report

Shefali Umrania

02-750: Fall 2017

1. Introduction

Biological image classification is unlike other traditional image classification problems. Modern computer vision models are trained on millions of natural images like ImageNet however acquiring an equal number of medical images can be a challenge [1]. Moreover due to this scarcity of data, the amount of labeled data is often many orders of magnitude smaller. Although the features obtained by training on natural images can be fine-tuned for medical images [2], this is only an incremental performance improvement. Active Learning is an area of machine learning which deals with improving models by choosing and annotating only the most informative data points thus eliminating the need for large amounts of data for improved performance. There have been a number of algorithms using active learning for various computer vision tasks [3, 4]. Some of these are also used for medical image classification [5, 6] where amount of unlabeled data is very low.

In this project report, I look at one such active learning method which uses a Query-by-Committee (QBC) based learning approach [7]. The task here is the multi-class classficiation of fluorescent microscopy images based on their subcellular localization patterns. The active learner selects only those features which would reduce the classficiation error the most and then requests the oracle for their labels. Once their labels are queried for, these features are then removed from the unlabeled pool of features and a new feature is chosen to query. In comparison to a random learner, it can be seen that the active learner requires fewer calls to the oracle and also has a lower test error. This drives home the point that active learning helps in reducing the expense associated with labeling large amounts of data by ensuring the model labels only the most informative feature.

2. Data

There are three datasets that have been used for this project:

1. Easy: A low-noise data pool
2. Moderate: This pool has some noise (labels and features)
3. Difficult: The points in this pool have a larger number of features than those in the easy and moderate pools. Some of these features are irrelevant.

Each dataset has 4120 training images and 1000 test images and is represented by a feature vector. Each feature is classified into 8 labels: Endosomes, Lysosomes, Mitochondria, Peroxisomes, Actin, Plasma Membrance, Endosomes, Microtubules, Endoplasmic Reticulum.

The budget for the maximum number of calls to query the oracle is 2500. The initial pool size for the algorithm is set to be 100 which increment by a batch size of 50.

3. Algorithm

The algorithm I used is a QBC based approach as explained by Nigam et al [7]. Ideally an active learning algorithm chooses only those points which will minimize the error upon querying their label. In QBC, this error is approximated by a group of features. A committee member is sampled for each class from an appropriate distribution. The model then calculates the classification error for that committee. For each feature, the disagreement value is also calculated. In my modification of the algorithm, I have used the absolute L1 distance metric as the disagreement equation. The features with the maximum distance are then queried for labels and subsequently placed in the already labeled set of features and are also removed from the unlabeled set. A new committee is then sampled from the remaining unlabeled set of features and this process is continued until the budget runs out. To assist with the multi-class classification, Scikit Learn's OneVsRestClassifier was used. In this case, a different classifier is used to fit each class.

The random learner was implemented by using SciKit Learn's SVM multi-class classifier. For randomly choosing points to be queried, Python's random package is used. For data manipulation, Python libraries such as Pandas and Numpy are used.

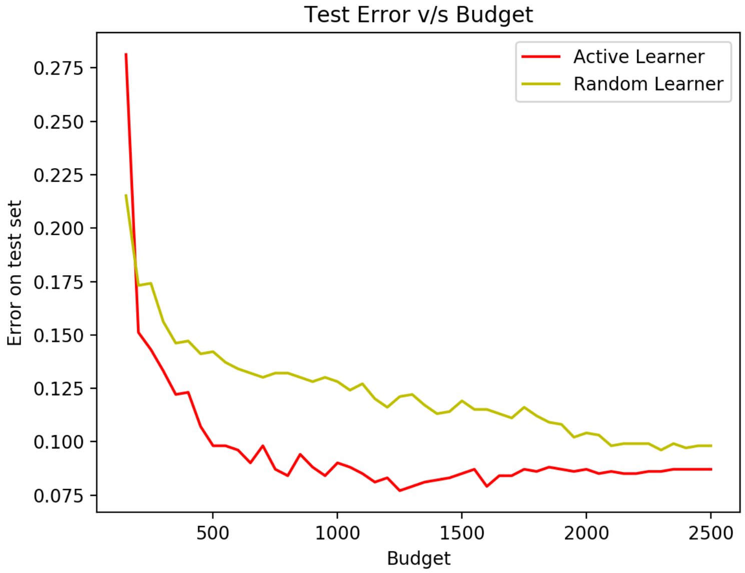

Figure 1: Test Error vs Number of Calls to Oracle on EASY dataset

I think the algorithm I chose is suitable for medical images such as the ones provided in this dataset, is because it offers a convenient way of dealing with very few labeled images. Moreover, it is also able to produce accurate results which are better than a random learner for the same budget. As there are no deep neural networks invovled, the computing resources required are also minimal and training time is not very large.

4. Results

As seen in Figures 1, 2, 3, the active learner performs better on the test set for each data pool compared to the random learner. The test classification error is consistently lower with each call to the oracle. In Figure 3 it can be see that in the initial calls to the oracle, the random learner performs better than the active learner. This can be attributed to the fact that the dataset contains a lot more noise and a larger number of features for the acitve learner to classify using very few labels.

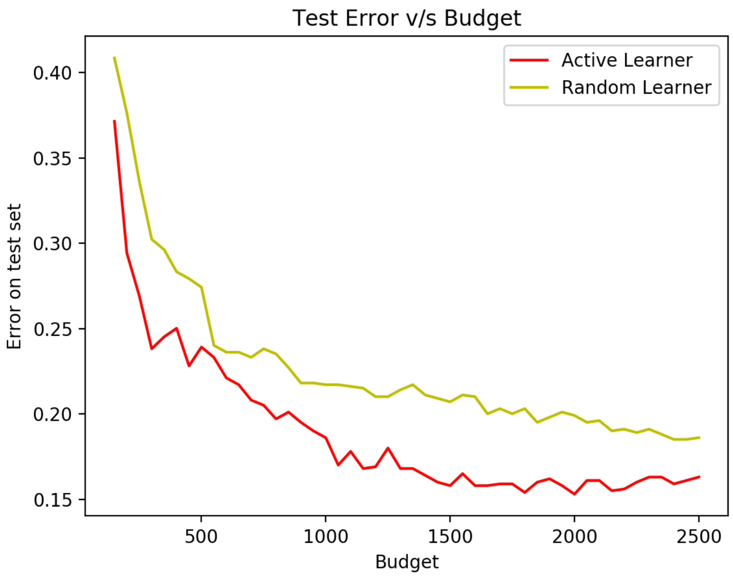

Figure 2: Test Error vs Number of Calls to Oracle on MODERATE dataset

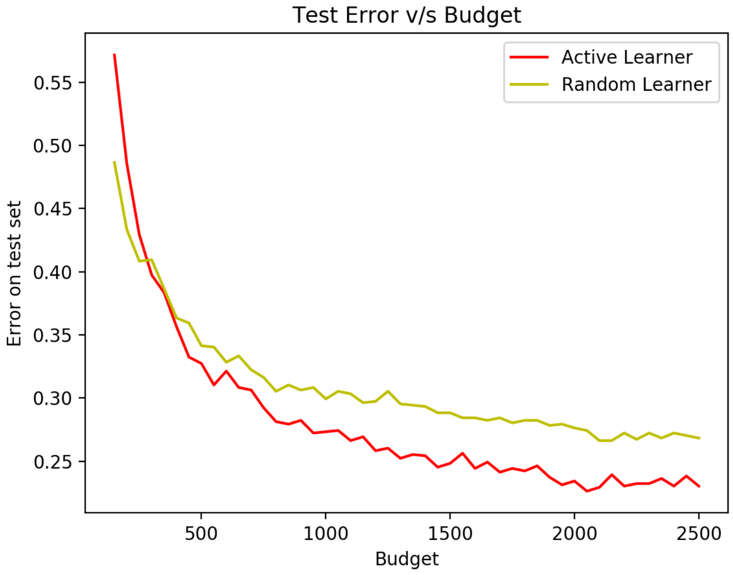

Figure 3: Test Error vs Number of Calls to Oracle on DIFFICULT dataset

5. Conclusion

As part of this project, I have shown that by using active learning for the multi-class classification of medical images we can obtain minimal test error in comparison to a random learner. The active learning algorithm used in this report - pool based query by committee uses the L1 absolute distance function to calculate the disagreement region and then queries only the features with the maximum distance. In this way, the model uses only 2500 labeled examples instead of required that all 4120 examples be labeled for traditional supervised learning.

6. References

[1] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning?, IEEE transactions on medical imaging 35 (2016) 1299–1312.

[2] N. Wang, P. Chu, Transfer imagenet for medical image analyzing using unsupervised learning (2016).

[3] B. Zhang, Y. Wang, F. Chen, Multilabel image classification via high-order label correlation driven active learning, IEEE Transactions on Image Processing 23 (2014) 1430–1441.

[4] F. Sun, M. Xu, X. Jiang, Robust multi-label image classification with semi-supervised learning and active learning.

[5] N. Kutsuna, T. Higaki, S. Matsunaga, T. Otsuki, M. Yamaguchi, H. Fujii, S. Hasezawa, Active learning framework with iterative clustering for bioimage classification, Nature communications 3 (????) 1032.

[6] S. C. Hoi, R. Jin, J. Zhu, M. R. Lyu, Batch mode active learning and its application to medical image classification, in: Proceedings of the 23rd international conference on Machine learning, ACM, pp. 417–424.

[7] A. K. McCallumzy, K. Nigamy, Employing em and pool-based active learning for text classification, Citeseer.