Project Report – Image Classification

Automation of Biological Research: 02-750, Fall 2017

Prajwal Prakash Vasisht – prajwalp@andrew.cmu.edu

Base Learner analysis:

The base learner is a vital part of any active learning algorithm. Hence, choosing a suitable base learner for the application domain we are targeting is very important. As part of the analysis, the base learners implemented are:

1) Linear Support Vector Machine:
    a. Support Vector Machines can easily model high dimensional data.
    b. SVMs with linear kernels are robust against noise.
    c. Low computation cost.
2) Random Forest Classifier:
    a. Ensemble models have a low bias.
    b. Prevents over-fitting.
    c. Similar to query-by-committee
3) Gaussian Naive Bayes Classifier:
    a. Naïve Bayes classifiers work on the principle that input features are conditionally independent given the class label. This was an avenue I wished to explore with the given data.

Random forest as an agent to mimic query by committee

In the project, the driving reason to use an ensemble model like the Random Forest classifier is due to its similarity to Query by committee sampling. By setting the value of n_estimators of sklearn's Random Forest classifier implementation, I can mimic a committee of size 100 and use the model to compare two query sampling strategies, namely, Uncertainty Sampling and Query-by-committee.

Results:

Table 1 summarizes the accuracy of the base learners on the given data sets.

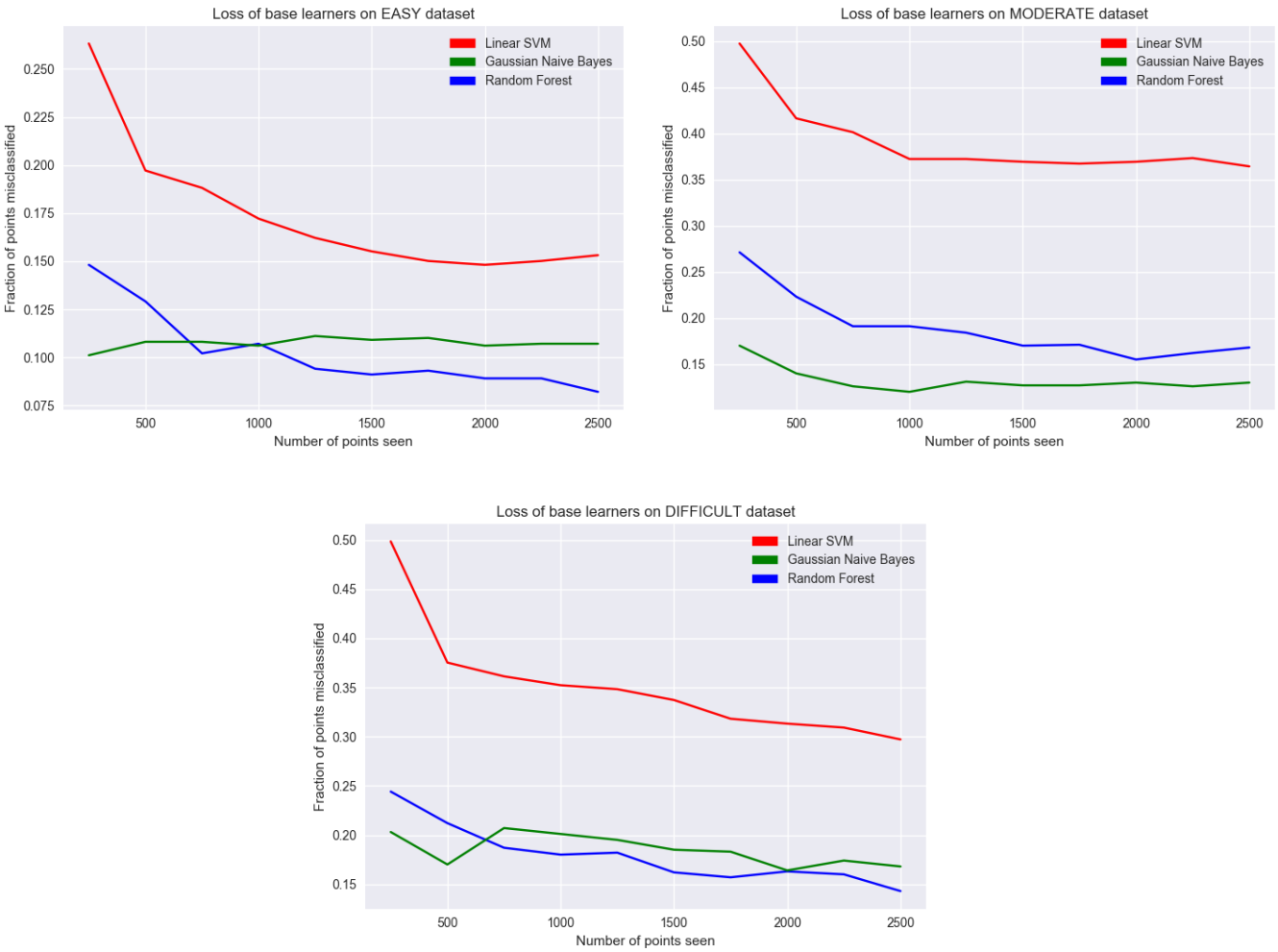
- The first 2500 (the stipulated budget) samples from the training set were used to train the models. This was done to mimic a streaming y data source.
- As expected, the random forest classifier has the best overall performance across the datasets.
- An interesting find is in the moderate dataset where the Gaussian Naïve Bayes classifier outperforms the random forest. The moderate dataset contains some noise features and labels and it could be possible that ignoring dependencies among the features helped the model to perform better

TABLE 1 - TEST ERRORS OF BASE (PASSIVE) LEARNERS

Test Dataset Base Learner	EASY (Accuracy in %)	MODERATE (Accuracy in %)	DIFFICULT (Accuracy in %)
Linear SVM	84.68	66.86	70.27
Random Forest	**91.79**	82.98	**84.68**
Gaussian Naïve Bayes	89.68	**86.88**	82.88

The following plots represent the percentage of misclassified points as a function of the number of training examples seen by each base learner.

Figure 1 - Loss curves vs Number of calls to the oracle



The plots below represent the accuracy of base learners as a function of the number of training examples seen. The plots help us understand the label complexity of each of the base learners and aid in the selection of an appropriate base learner for future active learning algorithms for this application domain.

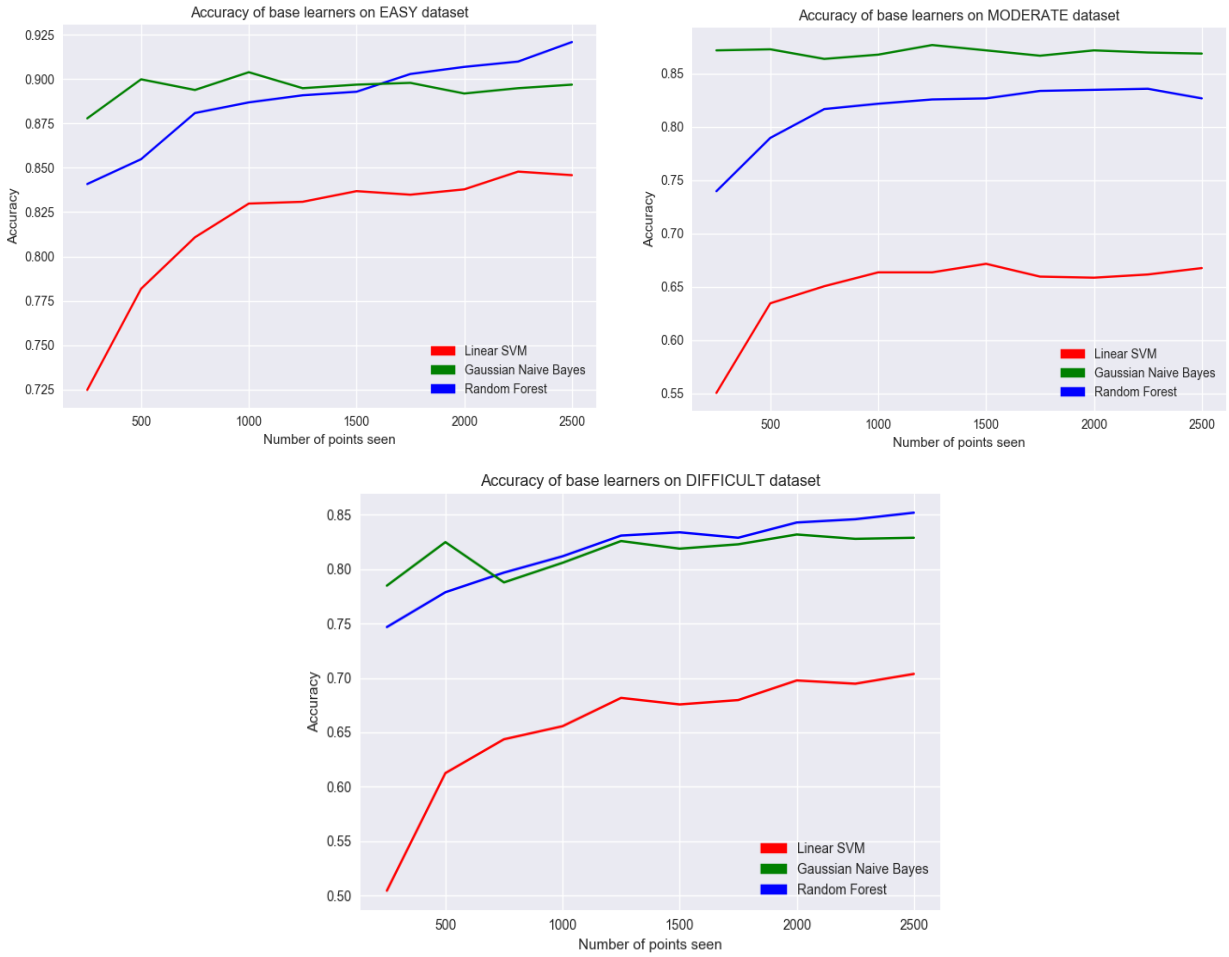


Figure 2 - Accuracy of base learners vs Number of training points seen

Conclusion

From the results of the analysis, we can conclude that using a Random forest classifier or a Gaussian Naïve Bayes classifier as a base learner for active learning algorithms would be appropriate for our dataset

The Active learning algorithm

After analysis of base learners, I decided to implement both the Random Forest and Gaussian Naïve Bayes classifiers as they corresponded to two different sampling methods, Query by committee and Uncertainty Sampling respectively.
Using sklearn's Random Forest Classifier, I constructed a random forest with a committee size of 100 and used the entropy measure of each unlabeled point in the training set and chose the point with maximum entropy in each iteration.
I implemented the Gaussian Naïve Bayes classifier using the same information measure i.e. entropy to sample the point with maximum entropy in each iteration.

The passive learner has been implemented in such a way that at each iteration a random starting point is chosen and then next 2500 consecutive samples are chosen for training.

The point of max entropy

To measure the uncertainty the model has about a data point I used the concept of entropy.
To calculate the entropy per point, I predict the probabilities of a point being classified as a certain label. If we have n labels, let us assume the probabilities of the point being classified as a given label are $(p_1, p_2, \ldots\ldots p_n)$ respectively,

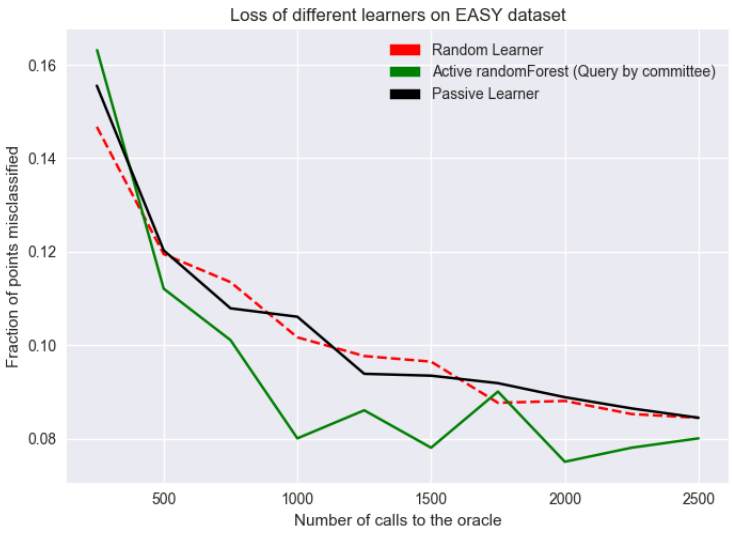
$$\text{Entropy of the point} = \sum_{i=1}^{n} p_i * \log(p_i)$$

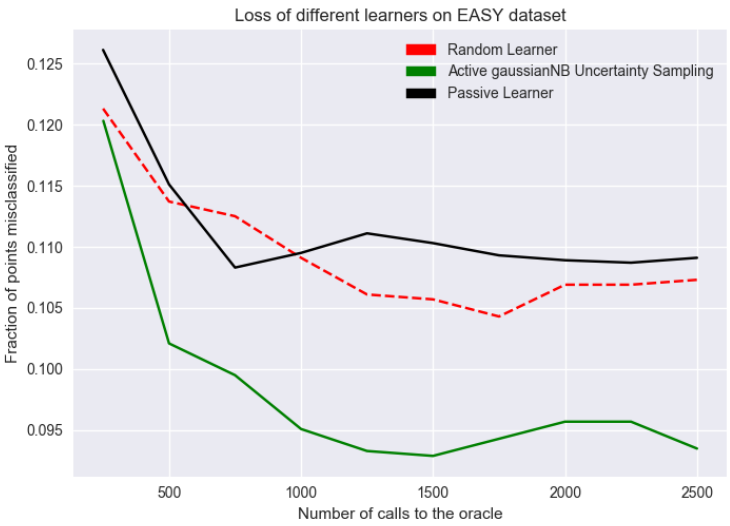
At each iteration, I choose the point with the maximum entropy and query the oracle and add the point to the labelled dataset. This process continues until we reach the stipulated budget and then the trained model is used to make predictions on the blinded dataset.

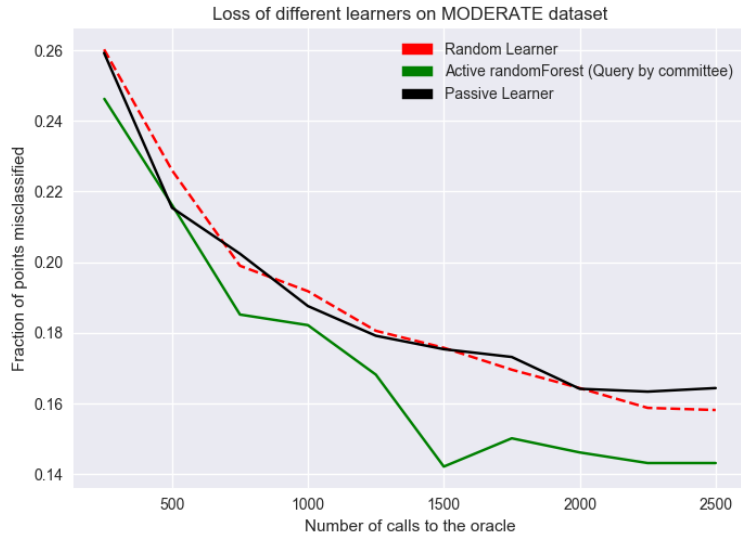
Active learning results

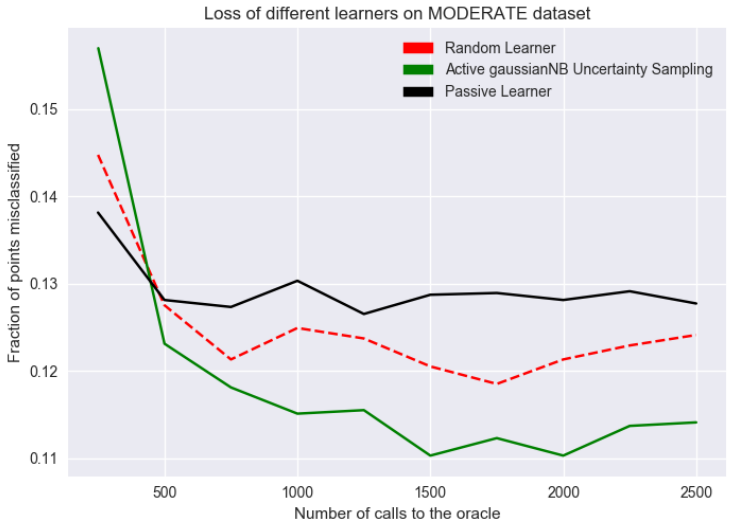
The results for all learners except that of the Active Random Forest classifier have been averaged over 5 iterations as the run time of the active random forest classifier is too high (~30mins per iteration). The averaging is required for the random and passive learners as by nature, the choice of samples chosen per iteration differ which can lead to a biased output if only one iteration is considered.
I have computed the loss of a model as a fraction of points which are wrongly classified by the model.

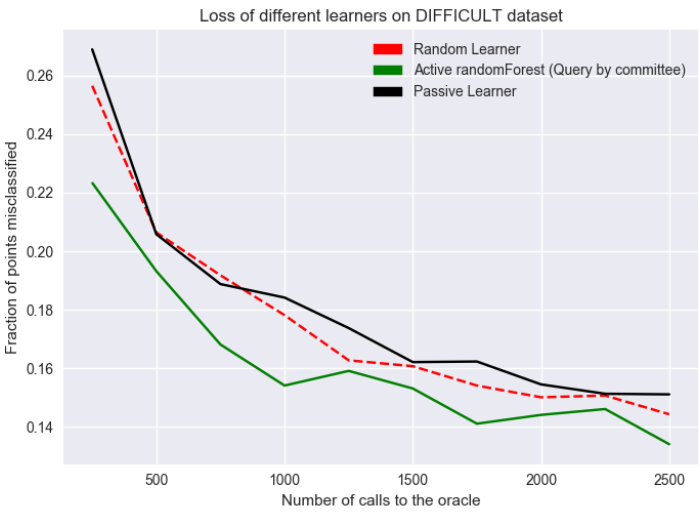
Following plots show the loss curves of the active learners on the test datasets as a function of number of calls to the oracle (here labelled-number of points seen).

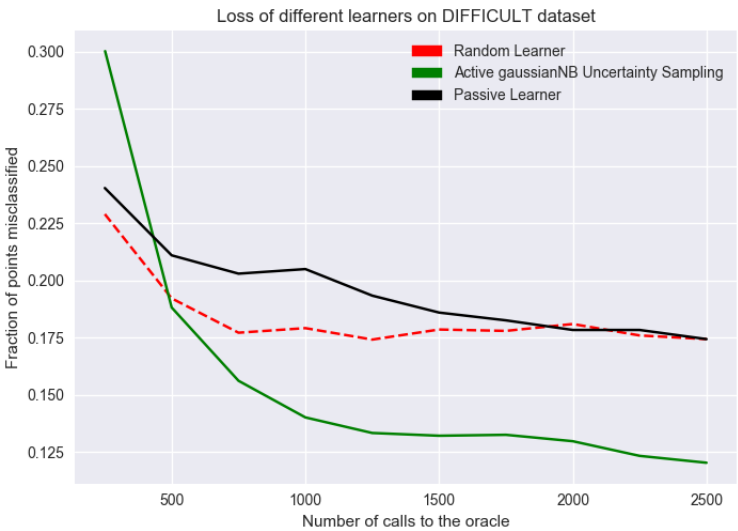

Loss of different learners on EASY dataset
Random Learner
Active randomForest (Query by committee)
Passive Learner
Fraction of points misclassified
Number of calls to the oracle


Loss of different learners on EASY dataset
Random Learner
Active gaussianNB Uncertainty Sampling
Passive Learner
Fraction of points misclassified
Number of calls to the oracle


Loss of different learners on MODERATE dataset
Random Learner
Active randomForest (Query by committee)
Passive Learner
Fraction of points misclassified
Number of calls to the oracle


Loss of different learners on MODERATE dataset
Random Learner
Active gaussianNB Uncertainty Sampling
Passive Learner
Fraction of points misclassified
Number of calls to the oracle


Loss of different learners on DIFFICULT dataset
Random Learner
Active randomForest (Query by committee)
Passive Learner
Fraction of points misclassified
Number of calls to the oracle


Loss of different learners on DIFFICULT dataset
Random Learner
Active gaussianNB Uncertainty Sampling
Passive Learner
Fraction of points misclassified
Number of calls to the oracle

Below table summarizes the performance of both active learners – Random Forest and Naïve Bayes classifiers on the 3 datasets:

TABLE 2 - TEST ERRORS OF ACTIVE LEARNERS

Test Dataset Active Learner	EASY (Accuracy in %)	MODERATE (Accuracy in %)	DIFFICULT (Accuracy in %)
Gaussian Naïve Bayes – Uncertainty Sampling	90.55	**88.52**	**88.48**
Random Forest – Query by committee Uncertainty Sampling	**91.89**	86.08	87.48

As portrayed by the loss curves and accuracy of the active learners on the 3 datasets, Gaussian Naive Bayes outperforms the Random forest classifier easily. In addition to the higher accuracy of the model, the label complexity of the Gaussian Naive Bayes classifier is lower than the random forest learner.
This can be attributed to the overall complexity of the random forest classifier. In the current implementation of the project, the random forest classifier is made up of an ensemble of 100 decision trees, each of which need to be trained accurately with samples representative of all the data in the domain we consider.
Another drawback of random forest classifier is that the model has become so large that it is very slow and run times for even the easy dataset can go up to 20 minutes which makes debugging and interpretation very hard.

Predictions on the blinded dataset

I have chosen to use the Gaussian Naive Bayes classifier to make predictions on the blinded dataset due to its simplicity, accuracy, speed and the overall lower cost of the model.

Conclusion

In conclusion, active learning algorithms tend to outperform passive learners and random learners by either:

- Achieving higher accuracy with a given budget.

  or

- Achieving a target accuracy with lower number of labelled data samples due to the query selection strategies.

References

[1] Smith and Horvath, "Active Learning Strategies for Phenotypic Profiling of High-Content Screens", J Biomol Screen June 2014 vol. 19 no. 5 685-695

[2] http://scikit-learn.org/stable/