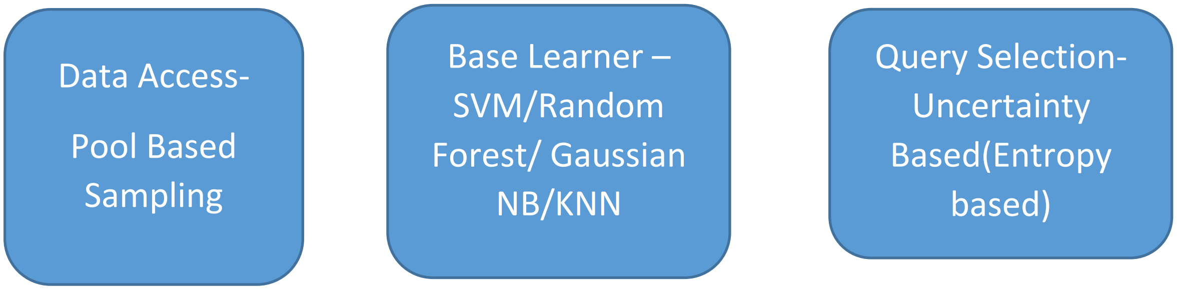
Image Classification Project

Name - Ramesh Oswal

Andrewid - ramesho

Algorithm used:



Data Access-

Pool Based Sampling

Base Learner – SVM/Random Forest/ Gaussian NB/KNN

Query Selection- Uncertainty Based(Entropy based)

Algorithm:

1. Import packages
2. Load data sets (train and test) into a numpy arrays and segment into features and labels.
3. Maintain two sets of feature instances. Labeled and Unlabeled
4. Pick (500) random instances from the Unlabeled set
5. On the initial pool using the SelectKBest (SelectKBest) algorithm retain the 26 best features.
6. Delete the corresponding features from the training, test and blinded feature set.
7. Repeat till cost limit is reached (2500)
    a. Get the labels for the selected instances from oracle and move the instances and labels to the Labeled set
    b. Update the cost by number of labels got from oracle
    c. Train a model (SVM/ Random Forest/ Gaussian_NB etc.) based on the labeled set of instances
    d. Predict the labels for Test dataset and calculate the error
    e. Pick a set (50) of the most uncertain values based on the current trained model from the unlabeled dataset.
        i. Uncertainty used is through entropy
        ii. Find the probability of all the predictions possible for a given feature
        iii. Find the entropy for each instance

2. Apply the above base classifier to predict the values for Blinded predictions.

Note: For Difficult Dataset:

Tools and Technologies Used:

IDE - Spyder (Spyder)

Python Libraries used – sklearn (al., 2011), numpy (others), matplotlib (Hunter), math (1990-2016, Python Software Foundation) and random (1990-2016)

Why we choosed the above strategy?

A.)Pool based sampling v/s Stream based Sampling?

As the entire pool of features is available at the start, using a pool based sampling strategy is better than stream based selective sampling. Pool based sampling strategy will converge much faster to the results than a stream based selective sampling.

In active learning we try to label the points which we are least sure about based on the current knowledge. In other words we will find the most uncertain points after every iteration which in case of stream based sampling will be computationally expensive (Settles).

B.)Classification v/s Clustering Algorithms for Base Learner?

“In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known” (Wikipedia, Classification Analysis)

“Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)” (Wikipedia, Cluster analysis)

Classification algorithms are a set of supervised learning algorithms whereas Clustering algorithms are a set of unsupervised learning algorithms. As in our case we can get the label from oracle for a particular instance Classification based algorithms are best suited for our data.

Also, looking at the labels we can say that only 8 set of categorical values are present in labels. Hence, a classification algorithm for supervised learning is best suited for our project.

Tested various Classification algorithms to check for their stability and test error with respect to the random learner.

C.)Query By Committee V/s Uncertainty based sampling.

“In order to implement a QBC selection algorithm, one must:

i. be able to construct a committee of models that represent different regions of the version space, and

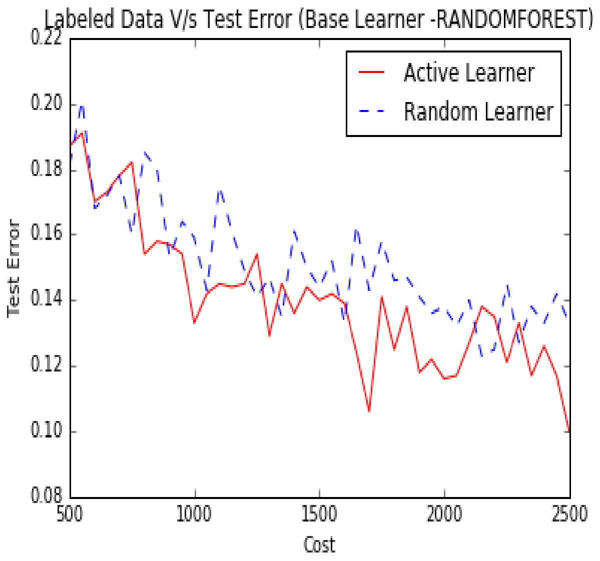
ii. have some measure of disagreement among committee members.” (Settles)

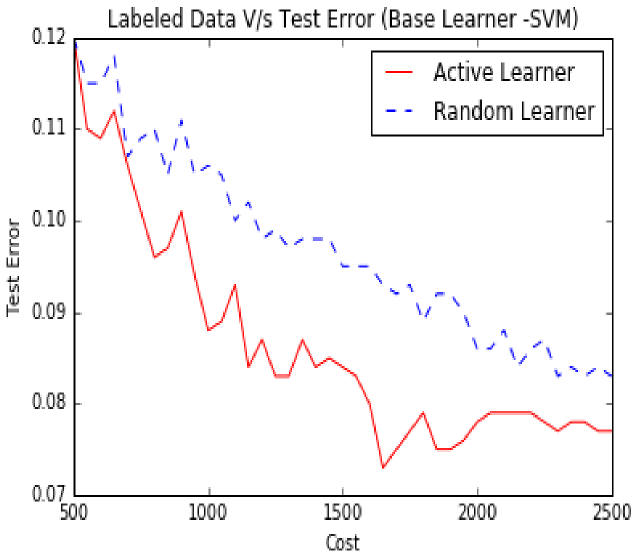
In uncertainty based sampling the active learner queries the instances which it is least certain about. It based on probabilistic graphical models.

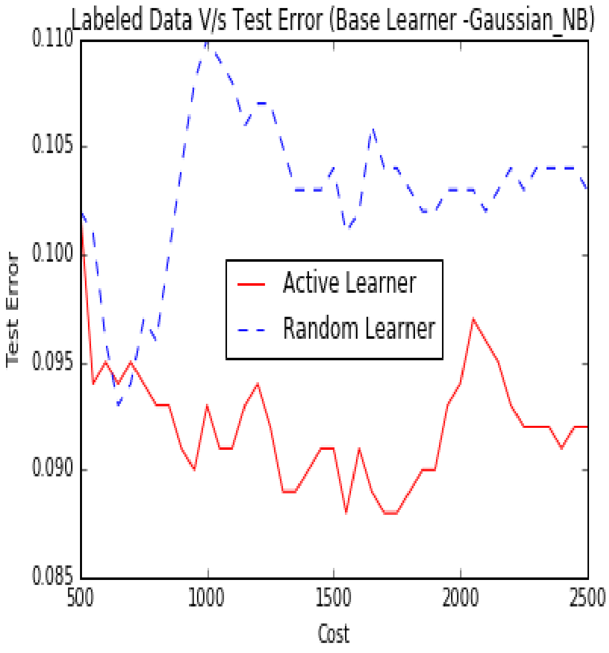
In our case as the label set has only 8 distinct features we can easily compute the probability of each label to be true. Using the probability, we can easily compute the instances which have maximum entropy.

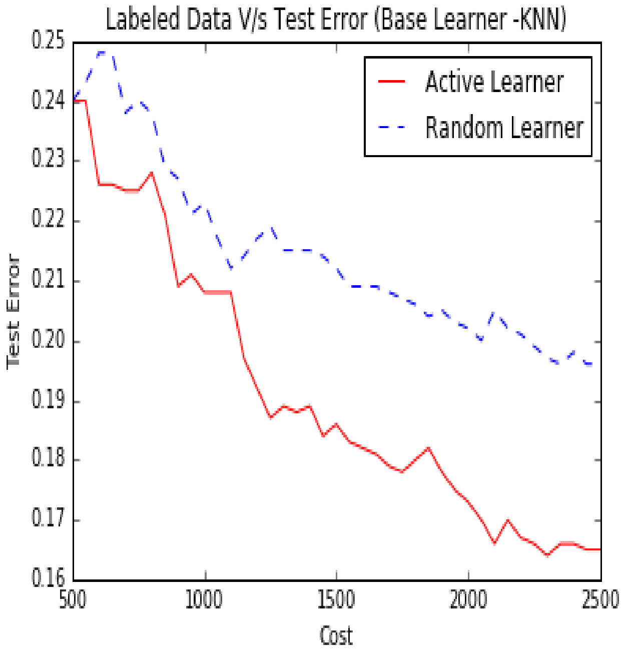
Figures/Graphs

EASY

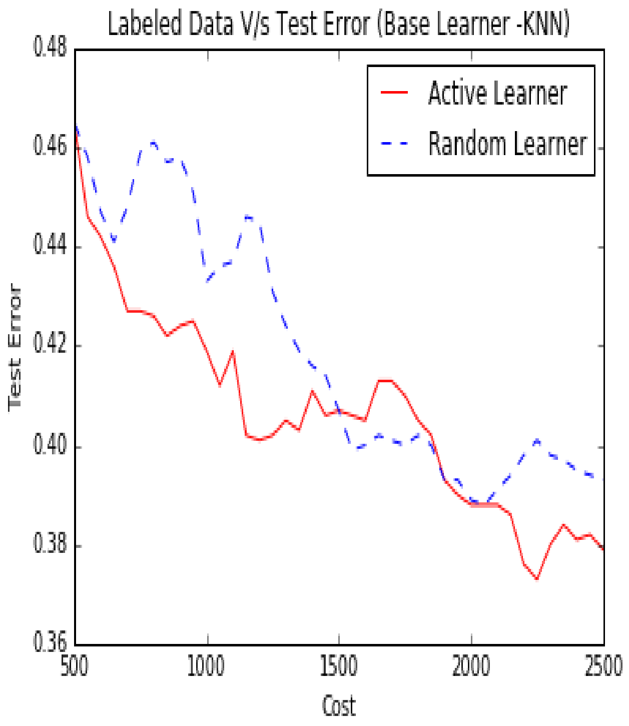


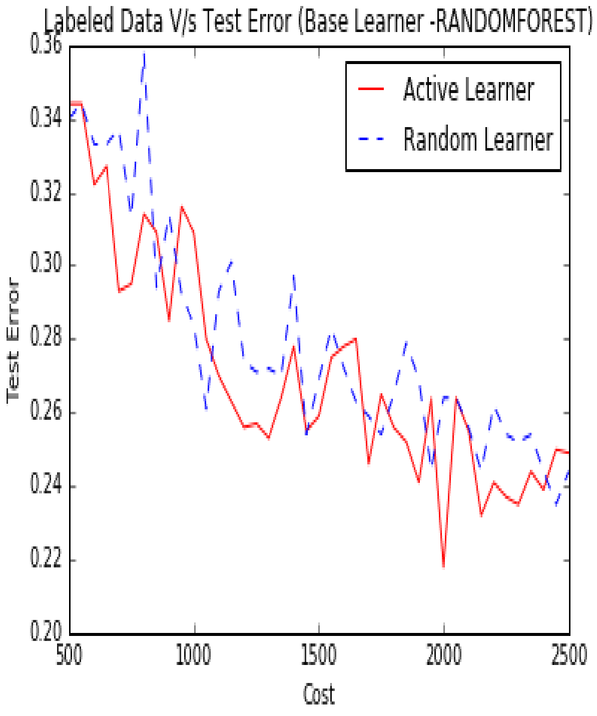


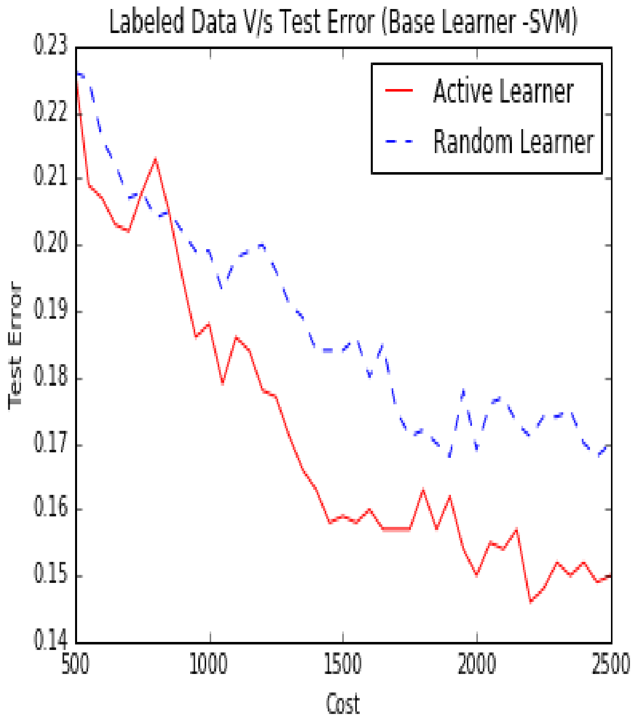


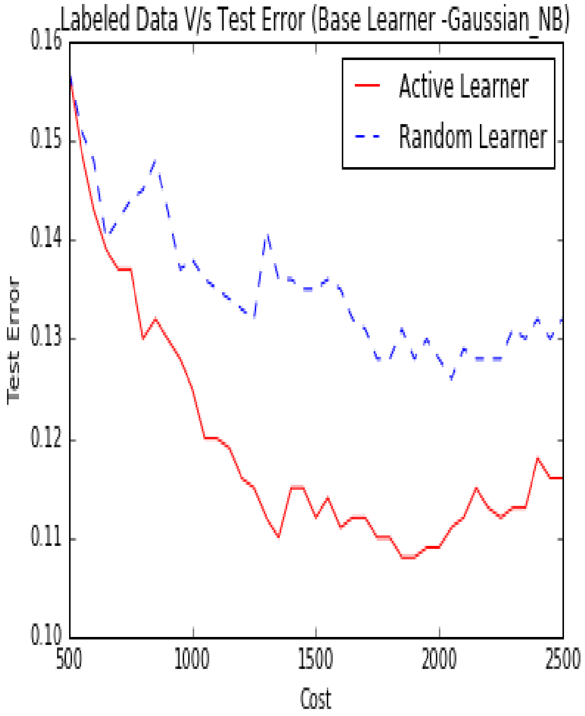


MODERATE

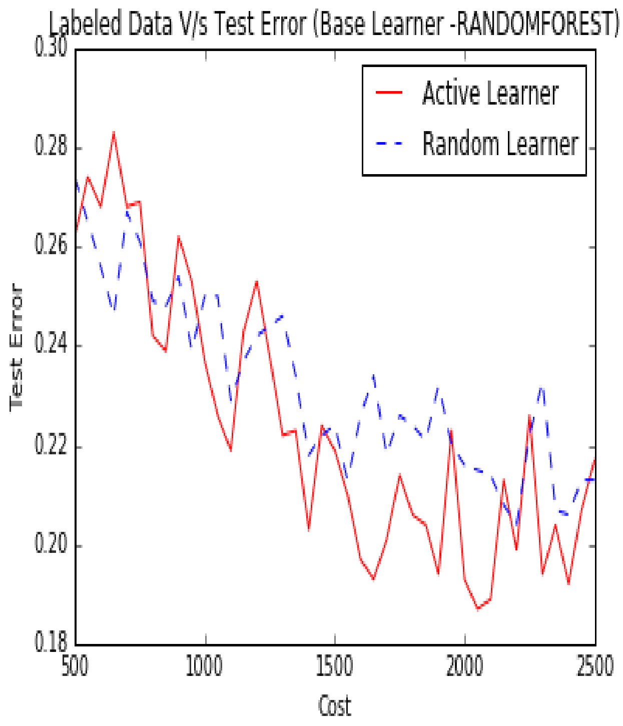

Labeled Data V/s Test Error (Base Learner -KNN)
Active Learner
Random Learner
Test Error
Cost
0.48
0.46
0.44
0.42
0.40
0.38
0.36
500
1000
1500
2000
2500

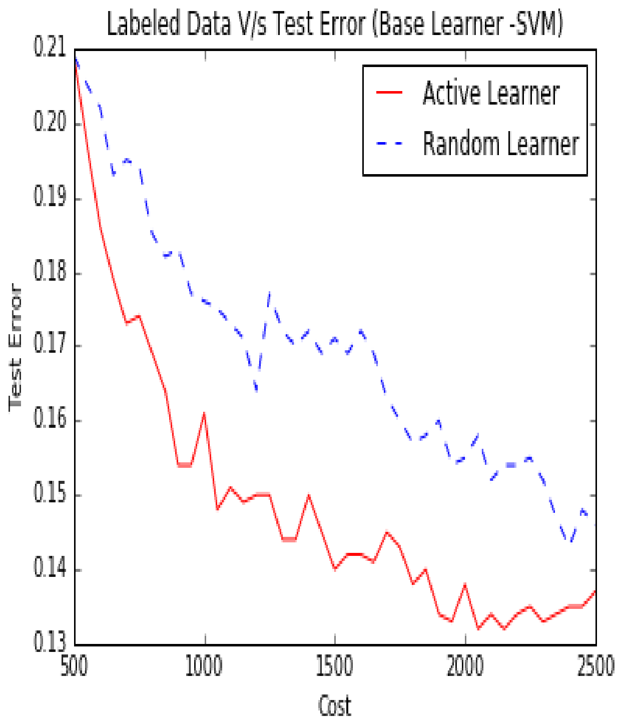

Labeled Data V/s Test Error (Base Learner -RANDOMFOREST)
Active Learner
Random Learner
Test Error
Cost
0.36
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0.20
500
1000
1500
2000
2500

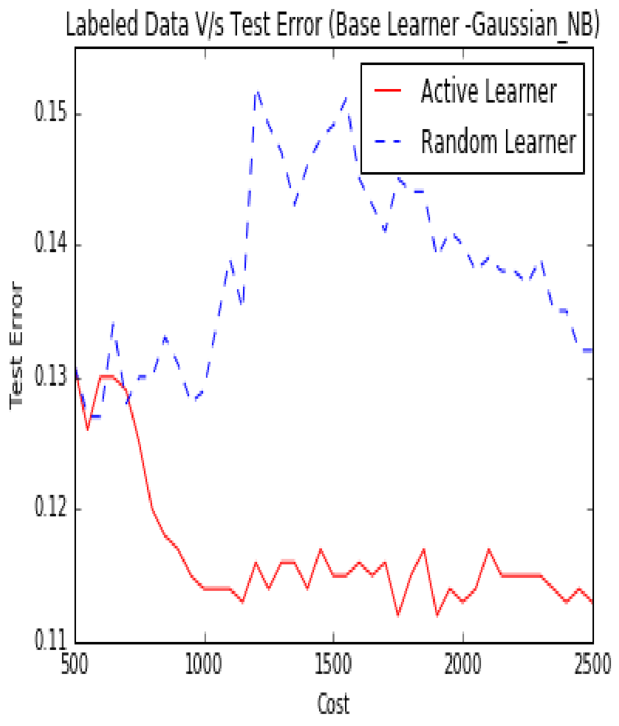

Labeled Data V/s Test Error (Base Learner -SVM)
Active Learner
Random Learner
Test Error
Cost
0.23
0.22
0.21
0.20
0.19
0.18
0.17
0.16
0.15
0.14
500
1000
1500
2000
2500

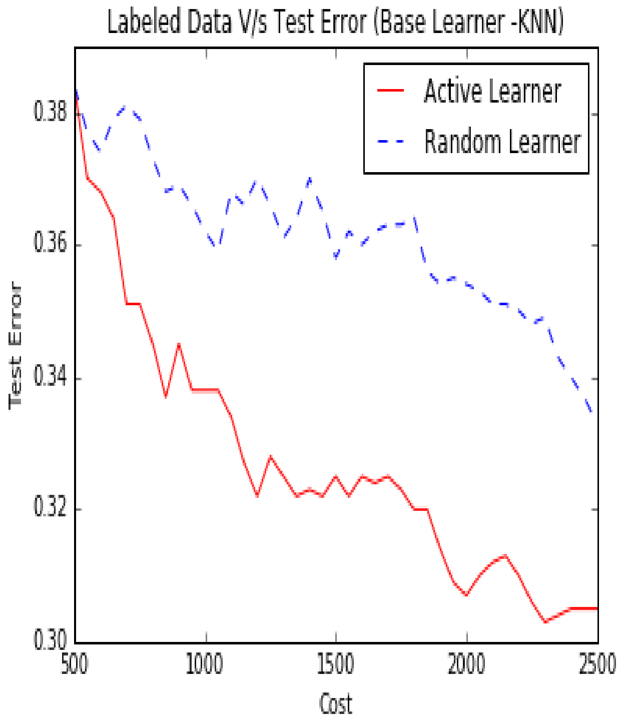

Labeled Data V/s Test Error (Base Learner -Gaussian_NB)
Active Learner
Random Learner
Test Error
Cost
0.16
0.15
0.14
0.13
0.12
0.11
0.10
500
1000
1500
2000
2500

DIFFICULT


Labeled Data V/s Test Error (Base Learner -RANDOMFOREST)


Labeled Data V/s Test Error (Base Learner -SVM)


Labeled Data V/s Test Error (Base Learner -Gaussian_NB)


Labeled Data V/s Test Error (Base Learner -KNN)

A brief summary of your findings

Parameters				Active Learner			Random Learner		
Algorithm	Dataset	Starting pool	Pool size	Min Test Error	Cost to get Min error	Test Error After 2500 cost	Min Test Error	Cost for Min error	Test Error After 2500 cost
RANDOMFOREST	EASY	500	50	0.1	2500	0.1	0.123	2150	0.133
RANDOMFOREST	MODERATE	500	50	0.218	2000	0.249	0.235	2450	0.244
RANDOMFOREST	DIFFICULT	500	50	0.187	2050	0.217	0.204	2200	0.213
SVM	EASY	500	50	0.073	1650	0.077	0.083	2300	0.083
SVM	MODERATE	500	50	0.146	2200	0.15	0.168	1900	0.17
SVM	DIFFICULT	500	50	0.132	2050	0.137	0.143	2400	0.146
Gaussian_NB	EASY	500	50	0.088	1550	0.092	0.093	650	0.103
Gaussian_NB	MODERATE	500	50	0.108	1850	0.116	0.126	2050	0.132
Gaussian_NB	DIFFICULT	500	50	0.112	1750	0.113	0.127	550	0.132
KNN	EASY	500	50	0.164	2300	0.165	0.196	2350	0.196
KNN	MODERATE	500	50	0.373	2250	0.379	0.388	2050	0.393
KNN	DIFFICULT	500	50	0.303	2300	0.305	0.333	2500	0.333

Based on the above table and graphs we can easily eliminate the KNN and Random Forest Base Learner as they have lot of instances when the Random Learner performs better than Active Learner.

Also, the algorithm performs much better on the easy data set than Gaussian Naïve Bayes. Hence, the blinded predictions were predicted considering SVM with kernel ‘rbf’ as the base learner.

While the Gaussian Naïve Bayes has better error than SVM, but looking at the graphs from Moderate and Difficult set we can see that there are lot of spikes in the active learner. This means that the Gaussian Active Learner Overfits the training data. Hence, SVM performs better on the EASY, MODERATE and DIFFICULT data set than Gaussian Naïve Bayes. Hence, the blinded predictions were predicted considering SVM with kernel ‘rbf’ as the base learner.