Comparison among Various Active Learning Strategies and CNN on RFI data

Kalyan G - 140001011

Guide - Dr. Aruna Tiwari

OVERVIEW

Active learning is a special case of semi-supervised machine learning in which a learning algorithm is able to interactively query the oracle to obtain the desired outputs at an unlabelled data point. Since the learner chooses the examples, the number of examples to learn a concept can often be much lower than the number required in normal supervised learning. The way learner chooses the examples is called Query Strategy. Every active learning algorithm needs a query strategy.

Radio Frequency Interference

Radio-frequency interference (RFI) is an Electromagnetic interference (EMI) when the radiations are in radio frequency spectrum. It is a disturbance generated by an external source that affects an electrical circuit by electromagnetic induction, electrostatic coupling, or conduction. Radio frequency interference (RFI) often occurs as short bursts (< 1ms) across a broad range of frequencies, and can be confused with signals from sources of interest such as pulsars.

GOALS

We will compare the learning ability of active learning algorithm when implemented over the following query strategies:

Uncertainty sampling

- ➔ Entropy Sampling Query Strategy
- ➔ Least Confident Query Strategy
- ➔ Random Selection

Query by committee

- ➔ Kullback Leibler Divergence Query Strategy
- ➔ Vote Entropy Strategy

And also compare its precision with a **convolutional neural network** model

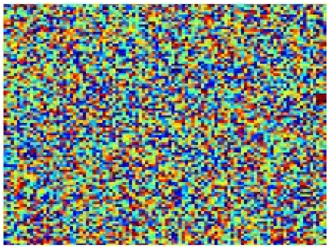
Dataset

RFI-Detection

Radio-frequency interference (RFI) is an Electromagnetic interference (EMI) when the radiations are in radio frequency spectrum. It is a disturbance generated by an external source that affects an electrical circuit by electromagnetic induction, electrostatic coupling, or conduction. This signal is converted into spectrogram (Image) using Short-time Fourier transform. This image can be used to determine if there exists a disturbance (RFI) or not.

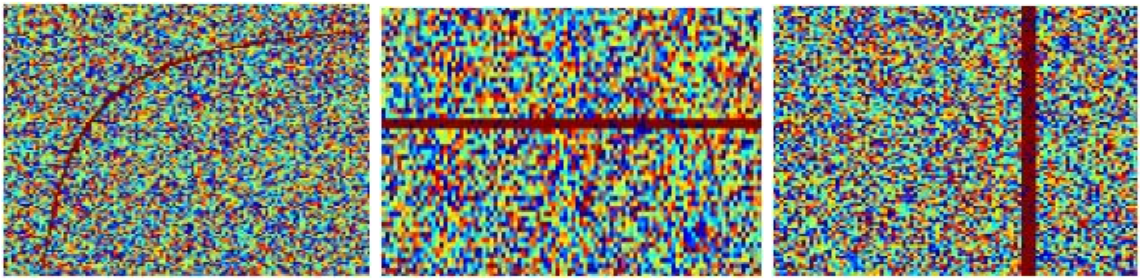
Spectrograms

Data without RFI :



Crude Outer Space Signal

Data with RFI :



Here I have used **simulated spectrograms** for training our active learning model. Our algorithm have to learn identifying patterns in an image, we can find which query strategy is good at doing this job of identifying patterns by comparing among them.

Algorithms Used

We are going to encounter two types of query strategies, Uncertainty sampling where we use a single learning model, here I have used **Naive Bayes classifiers**, while in case of Query by Committee strategies we need multiple models and we used **Naive Bayes classifiers and Sequential minimal optimization (SMO).**

Naive Bayes

naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

$$p(C_k \mid x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

Sequential minimal optimization (SMO)

Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines.

Consider a binary classification problem with a dataset (x_1, y_1), ..., (x_n, y_n), where x_i is an input vector and $y_i \in \{-1, +1\}$ is a binary label corresponding to it. A soft-margin support vector machine is trained by solving a quadratic programming problem, which is expressed in the dual form as follows:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j,$$

subject to:

$$0 \leq \alpha_i \leq C, \quad \text{for } i = 1, 2, \dots, n,$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

Query Strategies

Uncertainty Sampling Strategies:

label those points for which the current model is least certain as to what the correct output should be.

Entropy Sampling simplest and most commonly used query framework is uncertainty sampling. The algo queries the instances about which it is least certain how to label.A more general uncertainty sampling strategy (and possibly the most popular) uses entropy.

$$x_H^* = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x),$$

where y_i ranges over all possible labelings. Entropy is an information-theoretic measure that represents the amount of information needed to “encode” a distribution.

Least confident Query Strategy

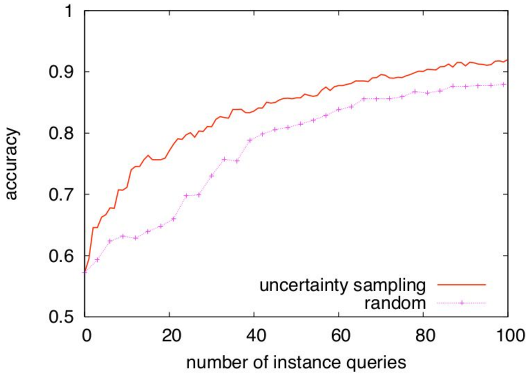
For problems with three or more class labels, a more general uncertainty sampling variant might query the instance whose prediction is the least confident:

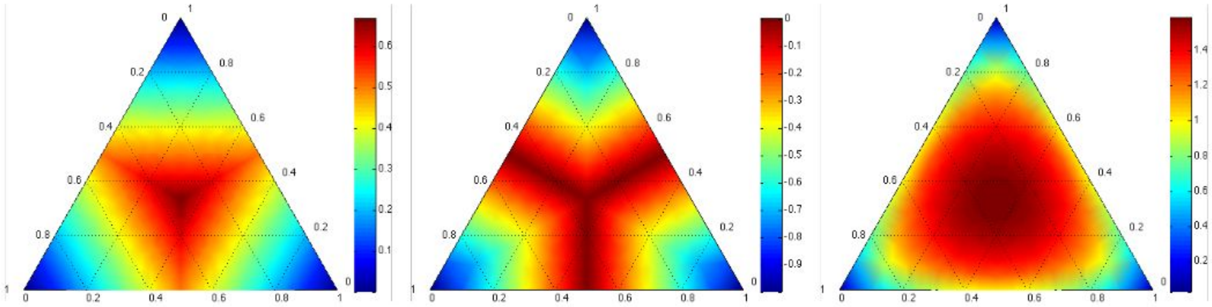
$$x_{LC}^* = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x)$$

where ŷ = argmax $_y$ P $_\theta$ (y|x), or the class label with the highest probability under the model θ .

Random Sampling

In this type of sampling there is no query strategy at all. The next instance to be queried is randomly selected. In fact there is no reason to study this but since here we are comparing among the query strategies it will be useful to study its behaviour comparing to others.





(a) least confident (b) margin (c) entropy

Heatmaps illustrating the query behavior of common uncertainty measures in a three-label classification problem. Simplex corners indicate where one label has very high probability, with the opposite edge showing the probability range for the other two classes when that label has very low probability. Simplex centers represent a uniform posterior distribution. The most informative query region for each strategy is shown in dark red, radiating from the centers.

Query by Committee

Another, more theoretically-motivated query selection framework is the query-by-committee (QBC) algorithm.The QBC approach involves maintaining a committee $C = \{ \theta^{(1)}, \ldots, \theta^{(C)} \}$ of models which are all trained on the current labeled set L.

The most informative query is considered to be the instance about which they most disagree.

A disagreement measure that has been proposed is **Kullback-Leibler** (KL) **divergence**

$$x^*_{KL} = \underset{x}{\operatorname{argmax}} \frac{1}{C} \sum_{c=1}^{C} D(P_{\theta^{(c)}} \| P_{\mathcal{C}}),$$

where:

$$D(P_{\theta^{(c)}} \| P_{\mathcal{C}}) = \sum_{i} P_{\theta^{(c)}}(y_i|x) \log \frac{P_{\theta^{(c)}}(y_i|x)}{P_{\mathcal{C}}(y_i|x)}.$$

Here $\theta^{(c)}$ represents a particular model in the committee, and C represents the committee as a whole.

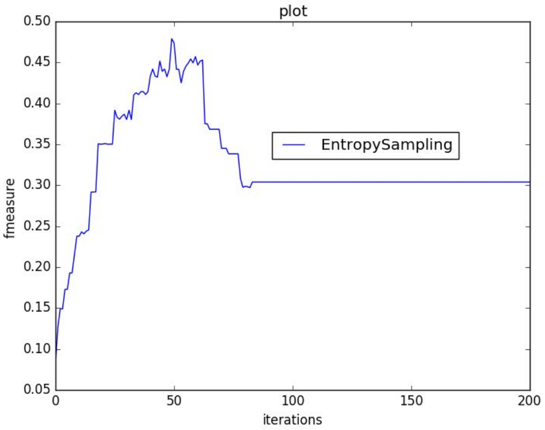
Another disagreement measure that has been proposed is **vote entropy query strategy.**

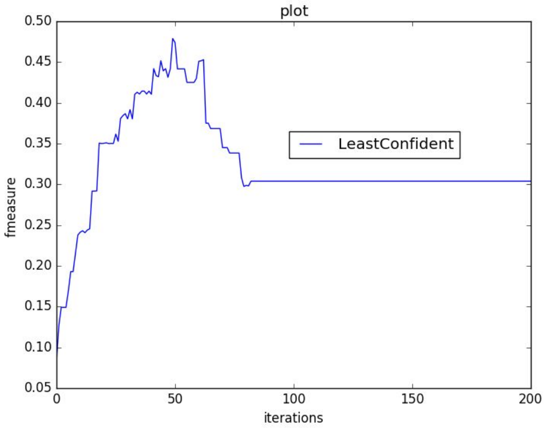
$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_{i} \frac{V(y_i)}{C} \log \frac{V(y_i)}{C},$$

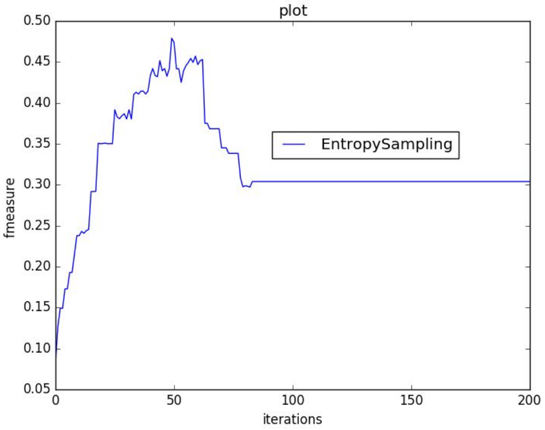
Experimentation:

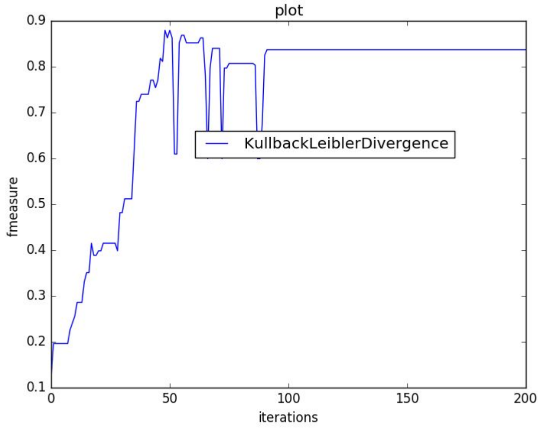
All the below trainings are done on the above discussed image dataset of size 200. Each image has a resolution of 20x20 = 400 pixels.

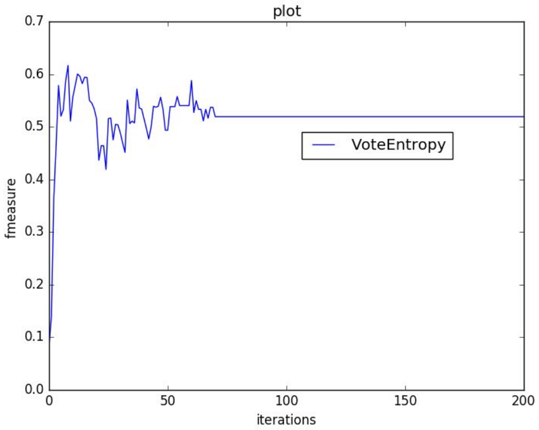
Learning process : Iteration vs Fmeasure









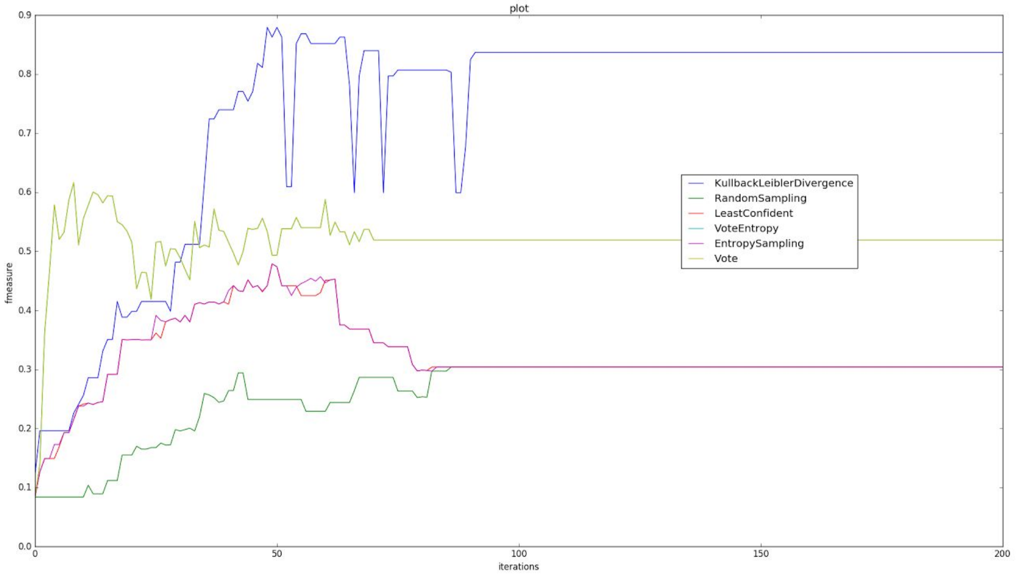


of

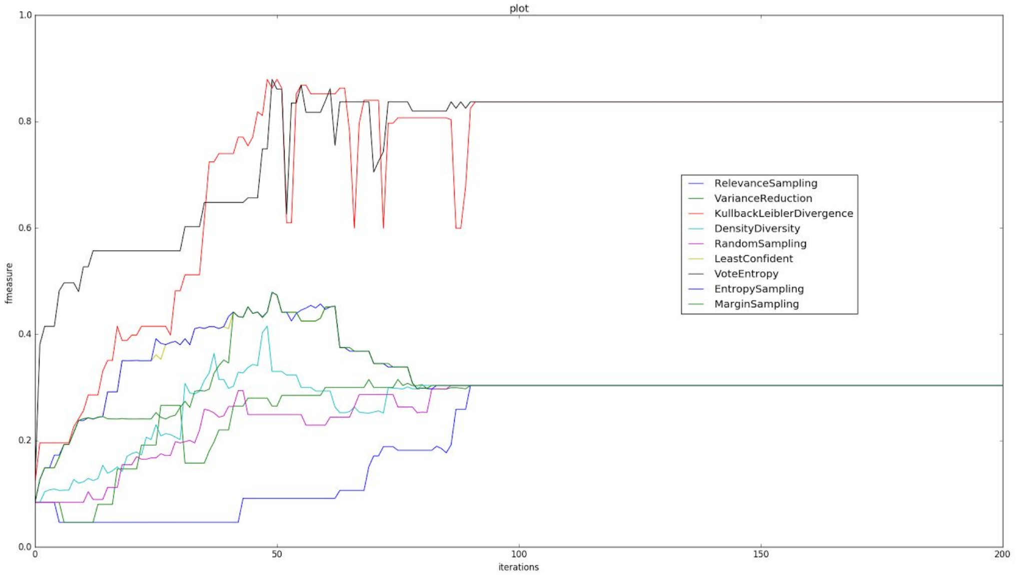
F-measure:

precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while **recall** (also known as sensitivity)is the fraction of relevant instances that are retrieved. And F-measure is the harmonic mean of both them

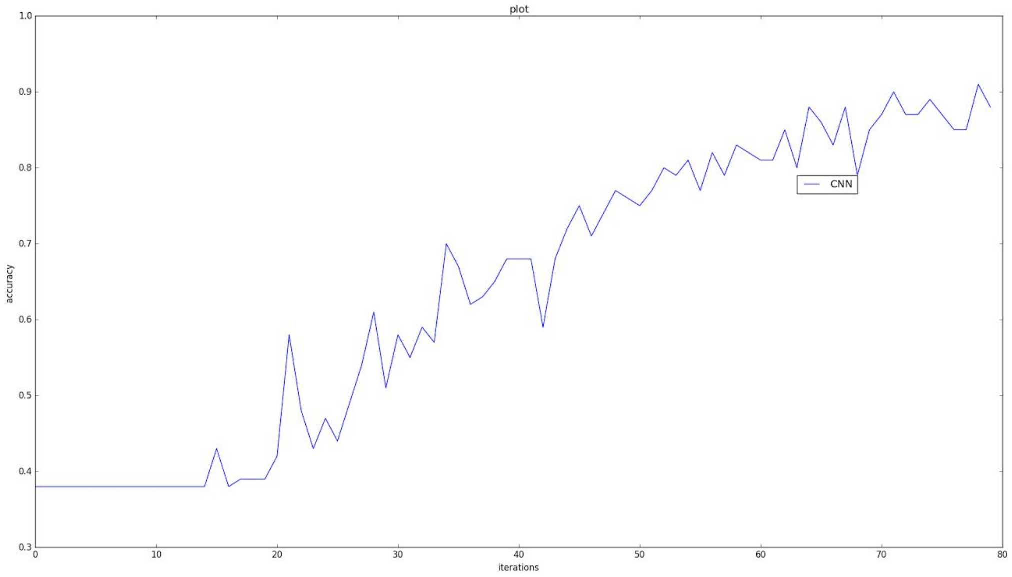
overall results learning:



On other dataset:



Convolutional Neural Networks Learning Results:



Results

We can observe from the above experimentation that Active learning with Query by Committee is

giving efficient results (accurate up to 88.9%) whereas the Uncertainty queries which use a single

learning model are not of much use for these image classification purpose.

Convolutional Neural Network models are also showing good results i.e., 90.1% accurate.