Active Learning to reduce the label complexity of classification in census income data

Anil Ramakrishna,
Department of Computer Science,
University of Southern California,
Los Angeles, CA
akramakr@usc.edu

Abstract

Active Learning is an important branch of Machine Learning that tries to reduce the label complexity associated with the task of constructing classifiers. In this work, we apply two major active learning techniques to the U.S. census income data and demonstrate the superior label complexity of active learning over supervised learning.

1 Introduction

Traditional machine learning techniques assume that a large number of labeled data points are available while training the classifier. However, in many practical applications, labeled data points are very expensive to obtain while unlabeled data points are available in plenty. For example, in the medical domain, we may have a large number of unlabeled data points that correspond to test results for different patients. Labeling these would require the intervention of doctors which may be an expensive task. In these situations, we need a way to construct a classifier using minimal number of labeled points. Several machine learning approaches have been developed to address this exact issue and **Active Learning** (AL) is a prominent technique among them. In this work, we evaluate the application of two major AL techniques: **Uncertainty Sampling** and **Hierarchical Sampling** to the US census income data set.

The census data set was released in 1996 and can be used to formulate a binary classification problem where the task is to predict whether or not a person earns more than $50000 per year. Using AL, we show that we can build a classifier that uses less than ten percent of the complete data but with the same prediction accuracy as one trained with full data.

The remainder of this paper is organized as follows. In section two we give an overview of work related to Uncertainty Sampling and Hierarchical Sampling.

We explain our approach in section three followed by the experimental results in section four. We conclude about the utility of our approach in section five.

2 Related Work

Active learning has been a focus of interest for researchers from several years and various algorithms have been developed in this regard. The term uncertainty sampling was first coined by [1] while referring to techniques that incrementally query for labels based on the uncertainty of unlabeled data points according a classifier learned using the available labeled points. They evaluate a heterogeneous approach where two different classifiers are used in each iteration.

The hierarchical sampling algorithm was first developed by [2] and has been successfully applied on several data sets. Unlike uncertainty sampling, it gives strong theoretical error bounds. For a thorough analysis of all the state of the art active learning techniques, the reader is referred to [3].

3 Approach

3.1 Data set

The data set contains about 32k training and 16k test records, each containing features such as age, education, occupation, etc. The output variable is a binary label where the positive class corresponds to persons with an annual salary of \$50000+. Additionally, we randomly split both the train and test data into two equal sized partitions and run our experiments on both parts.

3.2 Uncertainty Region Sampling

Uncertainty sampling is a simple yet powerful active learning strategy that iteratively performs two steps: learn a classifier on all available labeled data points, select the unlabeled point that has the highest amount of uncertainty among all available points according to the learned classifier. A general uncertainty sampling based active learning scheme is presented in algorithm 1 [3].

Algorithm 1 Algorithm for Uncertainty region sampling

$U \leftarrow$ a pool of unlabeled instances
$L \leftarrow$ a pool of labeled instances
while there are unlabeled points in U **do**
  $\theta \leftarrow train(L)$
  Select $x^* \in U$, the most uncertain point according to θ
  Query the label y^* of x^*
  Add the point $< x^*, y^* >$ to L and remove it from U
end while

The algorithm is run for a fixed duration or until all the available points have been labeled. Like all active learning algorithms, the uncertainty sampling algorithm assumes the presence of an **oracle** that can be used to query for labels of unlabeled data points. For example, in the medical domain, the unlabeled data points may be test results for different patients and the oracle may be doctors capable of labeling the condition as either life threatening or under control.

3.3 Cluster Based active learning

Cluster based techniques try to utilize the underlying structure of the data by identifying clusters that are *pure*, i.e. contain data points from one particular label class. The most effective of these is the **Hierarchical Sampling** algorithm due to [2]. The input to this algorithm is a hierarchical cluster tree of all the unlabeled data points. At any given point of time, the algorithm works with different *prunings* of the tree, which reflect the cluster structure it has learned based on the labeled data points seen so far. The full algorithm is presented in algorithm 2.

Algorithm 2 Hierarchical Sampling algorithm for Active Learning

Input: A hierarchical clustering T of the unlabeled data points
set initial pruning $P \leftarrow root(T)$ and chose a random label l
for time=1,2,... **do**
  Select node $v \leftarrow$ **select**(P)
  Pick a random point x from subtree T_v and query it's label from the oracle
  Update the counts, empirical probability estimates and error bounds of all nodes from x to v
  Choose P', the best pruning of T_v
  $P \leftarrow P \setminus \{v\} \cup P'$ and $L(v) \leftarrow L'(u)$ for all $u \in \mathrm{P}'$
end for
for all $v \in P$ **do**
  Update the label of x as $L(v)$ for all $x \in T_v$
end for

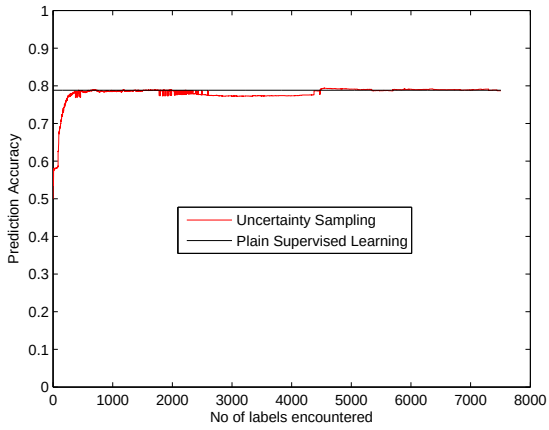
Here, P is the set of points in the current pruning of the tree T, T_v is the subtree of node v, $L(u)$ is the empirical estimate of the best label of node u, $p_{v,l}$ is the proportion of points in T_v with label l, $p_{v,l}^{LB}$ and $p_{v,l}^{UB}$ are upper and lower bounds for $p_{v,l}$, w_v is the weight of subtree v.

The algorithm maintains separate empirical counts and probability estimates for all the nodes in the tree. For each new labeled node it observes, all these estimates are updated in a bottom up fashion, starting at the newly labeled leaf node and ending at the pruning node v.

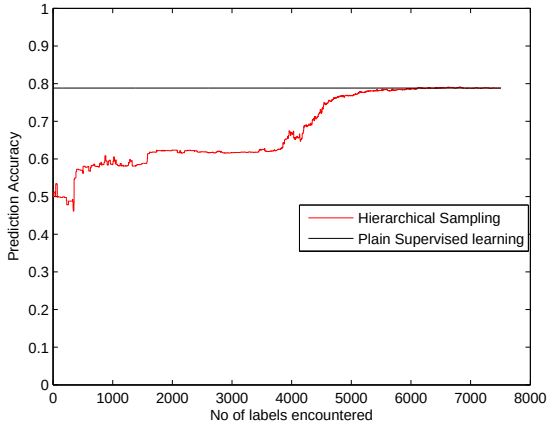
select(P) is the function that decides the active learning strategy used in the algorithm. For example, we could choose v with probability $\propto w_v$ which is equivalent to random sampling or with probability $\propto w_v(1 - p_{v,L(v)}^{UB})$ which samples more points from regions that are known to be impure.

4 Experimental Results

To evaluate the performance of the two active learners, we trained a classifier on all the available labeled data after each iteration and evaluated it on an independent test set. We report the prediction accuracy for both the learners run on two partitions of the data in the following graphs. For comparison, we also trained a classifier with all the labeled data available as a baseline result.



(a) Uncertainty Region Sampling

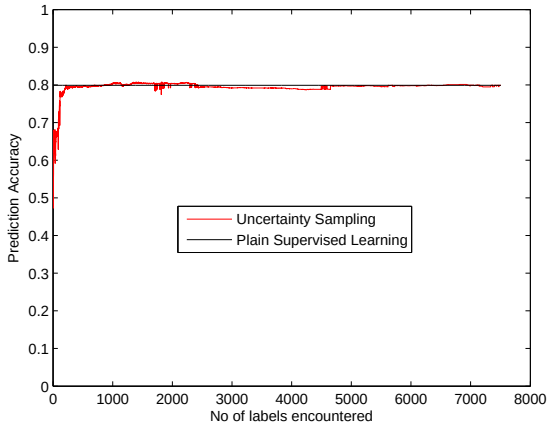


(b) Hierarchical Sampling

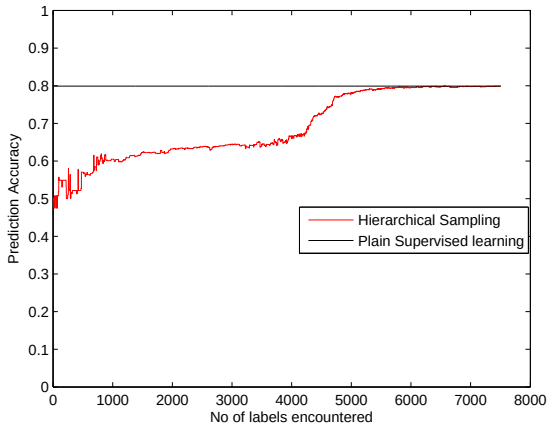
Figure 1: Prediction Accuracy on partition 1

The baseline supervised learner reported an accuracy of 78.81% and 79.89% respectively on the two partitions. As seen above, with both the partitions, uncertainty region sampling reaches this accuracy with just 500 labeled data

points. Hierarchical sampling however, takes relatively larger number of labels ($\sim 5,500$) to achieve this accuracy. It should be noted that, despite the early convergence of uncertainty region sampling, it does show a moderately high amount of variance with time. Hierarchical sampling however, remains stable.



(a) Uncertainty Region Sampling



(b) Hierarchical Sampling

Figure 2: Prediction Accuracy on partition 2

5 Conclusion

We applied two important active learning techniques to the US census income data and demonstrated that we can achieve the same classification accuracy

as a supervised learner with complete labeled data using less than 10% of the labeled data using the uncertainty region sampling.

Hierarchical sampling however, showed only moderate improvements when compared to uncertainty sampling. It should be noted that the performance of this algorithm depends largely on the cluster quality of the data set. By suitably modifying the features, we may be able to get better clusters and hence faster convergence. We leave this question open for future work.

References

[1] David D Lewis and Jason Catlett. Heterogenous uncertainty sampling for supervised learning. In *ICML*, volume 94, pages 148–156, 1994.

[2] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine learning*, pages 208–215, 2008.

[3] Burr Settles. Active learning literature survey. Technical report, 2009.