CSE 584 HW 1

Maxwell Dailey

September 2024

1 Paper 1

The first paper I chose to read is called: Active Learning Using Pre-clustering [2]

This paper is trying to solve the open problem of choosing which data points are the most important to label using active learning. The paper addresses this problem by incorporating clustering into active learning. More specifically, they use clustering to determine the prior distribution of the data and make use of that information when determining which samples to select through the minimization of the prediction of future classification error.

The novelties and contributions from this paper are: a unique meaning of $p(y|k)$ where it is defined for all clusters with the same variables instead of within individual clusters; a novel active learning algorithm that goes: Initial Clustering, Estimating $p(y|k)$, Calculating $p(y|x)$, Check if the stopping criteria has been met, Selecting and Labeling an Unlabeled Sample, and Cluster Adjustment which loops back to the Estimating $p(y|k)$ step; a novel criterion for data selection that takes into account proximity to classification boundary and how representative a data point is of a cluster.

One downside of the work is that the algorithm was only tested on image datasets instead of a variety of datasets. Another downside of the work is that the method they propose is restricted to linear logistic regression, though this was done because the main purpose of the paper was to show the benefit of using clustering information.

2 Paper 2

The second paper I chose to read is called: Active learning applied to automated physical systems increases the rate of discovery [3]

This paper is trying to solve the translation problem that currently exists when applying active learning methods to physical systems because, as of now, the increased pace of discoveries these methods promise has yet to be realized. This paper addresses this problem through a novel general framework for active learning within large-scale physical systems that is applied to large-scale boundary layer wind tunnel experiments to demonstrate the feasibility of using the framework in the study of physical experimental systems to increase the rate of discovery.

The novelties and contributions from this paper are: a specific framework for conducting active learning for large-scale experimental investigations, with the steps being: Identifying Quantities of Interest, Experimental Parameterization, Automated Actuation of the Test Environment, Automated Measurement of the Quantities of Interest, Data Processing and Analysis Utilities, and the Learning Function which loops back to the Experimental Parameterization step; three unique learning functions: Noisy U-function which worked to conduct experiments along the parameter surface separating the 2 order equivalent and non-equivalent regions, Noisy Expected Improvement for Global Fit (EIGF) which worked to conduct experiments that globally best approximate the performance function, and MUSIC which worked to conduct experiments that have both high prediction uncertainty and large differences from nearby experiments in conditional GP.

One downside of the work is that the framework was only applied to a single type of experiment (Boundary Layer Wind Tunnel (BLWT) experiments), and it has yet to be seen if the same framework would also work for other experiment types. Another downside is that the framework relies heavily on being able to automatically conduct physical experiments, something which can be hard to plan and execute.

3 Paper 3

The third paper I chose to read is called: Active Learning for Speech Emotion Recognition Using Deep Neural Network [1]

This paper is trying address the current problem with using deep neural networks for speech emotion recognition. Deep neural network solutions for speech emotion recognition rely on an incredible amount of labeled data to train, and labeling speeches with various emotional labels can be very costly. This results in relatively small datasets with only a few different speakers across the different recordings, which prevents the model from being able to generalize well to new domains. This paper endeavors to address this issue by exploring ways of making it feasible to train deep neural networks with limited labeled data through active learning.

The novelties and contributions from this paper are mainly the evaluation of the effectiveness of various acquisition functions for regression models in SER problems. Three classes of acquisition functions were tested: Uncertainty Based, Greedy Sampling (GS) Based, and Random Sampling (RS) based. For the Uncertainty Based class, a dropout approach was tested. For the GS Based class, feature space, label space, and feature and label space approaches were tested. For the RS Based class, a baseline was taken where a given number of samples from an unlabeled dataset were randomly selected.

One downside of the work is that it only tested a single Uncertainty Based acquisition function. Another downside is that Uncertainty Based acquisition functions that selected samples with medium uncertainty instead of the most uncertainty were not tested even though such functions have been shown to have potential in other papers. Lastly, the current approach does not do a good job of catching the uncertainty due to the lack of data versus the uncertainty of difficult samples.

References

[1] Mohammed Abdelwahab and Carlos Busso. Active learning for speech emotion recognition using deep neural network. *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7, 2019.

[2] Hieu Tat Nguyen and Arnold W. M. Smeulders. Active learning using preclustering. In *Proceedings of the 21st International Conference on Machine Learning*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.

[3] Michael Shields, Kurtis Gurley, Ryan Catarelli, Mohit Chauhan, Mariel Ojeda-Tuz, and Forrest Masters. Active learning applied to automated physical systems increases the rate of discovery. *Scientific Reports*, 13, 05 2023.