Anirudh Ravishankar & Gustavo Arismendi
COSC 7362 Advanced ML
Dr. Ricardo Vilalta

Project - Related Works

Active learning is a machine learning technique in which the learning algorithm selects instances to send to an expert (or oracle) to obtain the correct label and converge on the decision boundary more quickly than it would without the help of the expert (**Settles, 2010**). It shares elements with both supervised and unsupervised learning. Like supervised learning, the goal is ultimately to learn a classifier. But like unsupervised learning, the data come unlabeled. More precisely, the labels are hidden, and each of them can be revealed only at a cost. The idea is to query the labels of just a few points that are especially informative about the decision boundary, and thereby to obtain an accurate classifier at significantly lower cost than regular supervised learning. For instance the Learner selects the instance for which it has the least confidence in its most likely label. Some of the motivations for active learning include situations in which there's an abundance of unlabeled data but few labeled samples and situations where it may be very costly to obtain labels (**Settles, 2010**). Currently most of the algorithms in active learning fall into one of the three following categories: uncertainty sampling, query by committee, and expected reduction (**Padmakumar et all, 2018**). A research was done to guide sampling by a Cluster-based framework (**DasGupta08**). Other research has combined the areas of reinforcement learning with active learning to develop a policy that evaluates a number of information indicators to choose queries appropriate for the given task (**Padmakumar et all, 2018**).

1. Our approach involves using the labeled data to come up with an initial decision boundary and upon picking the right features and representing them in multiple dimensions we use distance as a measure among the labeled data.
2. We choose a pair of sampled data having contradicting class labels which has maximum distance among them, we proceed to take the mean of the distance between them.
3. Choose k-nearest neighbors (even numbers) to the mean. Labeling of the k-nearest sample data is done and added to the labeled set and retrain our model to generate a new temporary decision boundary.
4. We proceed to choose the next pair of opposite labeled data samples to generate the new mean and repeat step 3.
5. Finally we stop the process after all sampled data has been labelled and the global decision boundary has been obtained.

References

▪ Active Learning: Literature Survey, by Burr Settles, Technical Report, University of Wisconsin-Madison, 2010.
▪ Hierarchical Sampling with Active Learning by Dasgupta and Hsu, Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008.
▪ Balcan, M.-F., Broder, A., & Zhang, T. (2007). Margin based active learning. COLT.
▪ Castro, R., & Nowak, R. (2007). Minimax bounds for active learning. COLT.
▪Aishwarya Padmakumar, Peter Stone, Raymond J. Mooney In Proceedings of the 2018. Learning a policy for Opportunistic Active Learning. Conference on Empirical Methods in Natural Language Processing (EMNLP-18), Brussels, Belgium, November 2018.
▪ Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. ACL.