1 What is active learning? Why do we need it?

Active learning (from now on AL) in the simplest terms is to learn a classifier in the settings where data comes unlabeled, but still certain data points can be queried for their labels at a certain cost [1, 2]. Considering the fact that there is a huge amount of unlabeled data in various domains (documents off the web, speech samples, images, videos, fault logs etc.) and the almost impossible to afford cost associated with labeling all these data, AL appears as a promising solution [1–4]. AL picks up particular instances from a pool of unlabeled data points on the basis of their informative power and asks a human annotator for the labels of these points [5]. With this approach the amount of data points required to train an accurate classifier is greatly reduced. In terms of prediction capability and sample complexity, it has been shown that under certain constraints (hypothesis class of homogenous linear separators and uniformly distributed seperable data on the unit sphere), AL achieves the same error rate as a linear supervised learner with an exponentially lower sample complexity. In this introduction, we will provide a summary of the theory behind active learning as well as proposed methodologies for this theory and we will also include a section providing the state-of-the-art technologies for the application of active learning.

2 Theory and Methodologies

The aim in AL settings is to learn an unsupervised learner that is as accurate as supervised learner, yet with a considerable less sample complexity. For attaining this goal, recent literature proposes two methodologies [2, 6]:

- **Efficient search through hypothesis space:** We can think of the possible set of classifiers as a hypothesis set. Our aim in this methodology is to choose the points that would help us shrink this hypothesis space the most, i.e. at each new point, the set of learners that are roughly consistent with the data points seen so far shrinks more.
- **Exploiting the cluster structure in data:** This method is based on the assumption that unlabeled data points form clusters and by effectively discovering these clusters, we can use querrying strategies to label instances in each cluster. A very simple example is to form clusters of data (via clusterer x), querry random points in each cluster and assign the majority label of queried points as the label of the cluster.

Current theory tells us that AL can be applied anywhere, where supervised learning can be applied. However, although AL seems like an easy to apply approach, it faces a unique problem called sampling bias, which separates it from other learning models.

- **Sampling bias:** A standard AL scheme may start with randomly selected points to get an intuition of the decision boundary and continually select points that are closer and closer to its current decision boundary. Thereby, in the later stages of learning, AL scheme may significantly diverge from discovering the underlying distribution of data and may fall victim of its assumed decision boundary [7].

3 Technologies

After providing a summary of active learning and commonly proposed methodologies, in this section we will provide the technologies proposed in the recent years for AL related problems.

- **Hierarchical sampling for active learning:** Dasgupta et. al. make use of cluster structure in the data [6]. The summary of their approach is: Hierarchicaly cluster data into a tree structure, querry random instances in the tree, keep statistics (purity/impurity) for each subtree, eliminate the pure clusters (since we think they are pure enough), then query some more instances and follow the same procedure. For assigning a label to a cluster use majority voting of the labeled instances in the cluster.
- **Active learning for biomedical citation screening:** Authors use different oracles to predict the same point (3 oracles) and ask the experts only the points where the oracles differ. The original idea in this paper is to let expert not only tell the labels but also weight or even delete some of the features (feature weighting) [8].
- **Multilabel learning by exploiting label dependency:** Zhang et. al. propose building Bayes-Net structure on the data labels to discover the dependencies among different labels. Then they remove features responsible for dependencies among different labels, thereby making each class indepent of one another. Finally they build another Bayes-Net on each independent class for prediction. (I do not think that this is directly an active learning solution, since it actually uses labels, i.e. supervised learning. However, it was one of the conference papers that you asked us to read.)
- **Proper unit selection in active learning:** Due to sampling bias missed class effect (complete miss of discovering a class) occurs and authors show that missed class effect can be avoided by changing the granularity level of instance sampling [5]. They work on text mining and change granularity level from single words to sentences or word groups.
- **Multi-view multi-label active learning:** This work is a solution for multilabel problems (attributes define more than one labels/classes, as in bio-metrics and image processing) [9]. Authors propose making use of multiple learners and exploiting their agreement/disagreement for the same label set for finding which points to query.
- **Agnostic active learning:** Agnostic sampling is a more robust version of selective sampling. Selective sampling may delete a hypothesis/learner depending on a single query. However, agnostic learner keeps history of learners and requires multiple query fails for a learner to be deleted from hypothesis set [4]. Reported to work particularly well in the noisy data, provided that sample points are i.i.d.
- **Coarse sample complexit bounds for active learning:** This is a rather theoretical work on sample complexity of active learning. Authors generalize the complexity measure of homogenous linear classifiers to non-homogenous classifiers [1]. They propose a complexity measure, which takes into account: Distribution of input space, specific target learner and desired error rate.
- **Analysis of greedy active learning:** Dasgupta proposes using a greedy method for choosing the learner among the hypothesis space [1]. Greedily choose the set of learners that most evenly divides the instace space. This is a sub-optimal solution

and is limited to low dimensional spaces and does not guarantee the lowest number of queries, but it reduces the instance complexity.
- **The use of simulated experts:** Since experts are not very easy to find, the study proposes replacing an expert with a learner like Naive Bayes [10]. The idea behind is that experts cannot give a conclusive answer regarding their labeling, hence they can be simulated by algorithms. since it uses algorithms for labeling, a certain initial amount of data is necessary.

4 Summary of the summary

The most basic list of technologies exploited by all the above papers (all after 2000 and the important ones before 2000) is:

- Hierarchical clusterers
- Majority voting of multiple learners
- Bayes-net to discover data structure
- Change of sampling granularity (as solution for sampling bias)
- Agreement disagreement among multiple learners
- Agnostic sampling (multiple fails of learner to be deleted)
- Sample complexity measurement
- Greedy selection methods for hypothesis space
- Replace human expert with a learner

References

1. L. S. Dasgupta, "Coarse sample complexity bounds for active learning," in *Advances in Neural Information Processing Systems 18*, 2005.
2. S. Dasgupta and J. Langford, "Tutorial summary: Active learning," in *ICML*, p. 178, 2009.
3. M. Hasenjäger and H. Ritter, "Active learning with local models," *Neural Process. Lett.*, vol. 7, no. 2, pp. 107–117, 1998.
4. M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," in *ICML '06: Proceedings of the 23rd international conference on Machine learning*, (New York, NY, USA), pp. 65–72, ACM, 2006.
5. K. Tomanek, F. Laws, U. Hahn, and H. Schütze, "On proper unit selection in active learning: co-selection effects for named entity recognition," in *HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, (Morristown, NJ, USA), pp. 9–17, Association for Computational Linguistics, 2009.
6. S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *ICML '08: Proceedings of the 25th international conference on Machine learning*, (New York, NY, USA), pp. 208–215, ACM, 2008.
7. D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 1994.
8. B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Active learning for biomedical citation screening," in *Knowledge Discovery and Data Mining (KDD)*, 2010.
9. X. Zhang, J. Cheng, C. Xu, H. Lu, and S. Ma, "Multi-view multi-label active learning for image classification," in *ICME'09: Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, (Piscataway, NJ, USA), pp. 258–261, IEEE Press, 2009.
10. P. Compton, P. Preston, and B. Kang, "The use of simulated experts in evaluating knowledge acquisition," in *University of Calgary*, pp. 12–1, 1995.