Hierarchical and Uncertainty Sampling for Active Learning

Zhe Yu, Rahul Krishna
Com Sci, NC State, USA
{zyu9, rkrish11}@ncsu.edu

Tim Menzies
Com Sci, NC State, USA
tim.menzies@gmail.com

ABSTRACT

In this paper, a clustering-based selection procedure for active learning, called hierarchical and uncertainty sampling, is proposed. It is a combination of hierarchical sampling and uncertainty sampling, aiming to complement each of the sampling method. By applying hierarchical sampling to obtain a well-distributed initial set, sampling bias can be avoided. By empowering the system with the ability of refining of uncertainty sampling, a better performance can be achieved at final stage comparing to hierarchical sampling.

KEYWORDS: Active Learning, Sampling Bias, Clustering-based Selection Procedure, Hierarchical Sampling, Uncertainty Sampling.

1. INTRODUCTION

Nowadays, it has been easier and easier to obtain huge unlabeled data set while the cost of labeling becomes a dominating part of the total cost of a classification task. Pool-based active learning is the strategy to stress this problem. By enabling the learner to choose the data point from which it learns, it can achieve a better performance with fewer labeled training data used. There are other scenarios where active learning applies to, i. e. query synthesis and stream-based selective sampling. However, pool-based is the most common scenario [7] and all the works in the following part of the paper will be based on this scenario.

The central problem of active learning is the selection procedure, which is usually reduced to the measurement of informativeness of an unlabeled data point [1]. Lots of selection procedures have been explored in the past decades, e.g. uncertainty sampling, impact sampling, sampling by disagreements, hierarchical sampling and so forth. Most of these are non-clustering-based selection procedures except for hierarchical sampling.

Uncertainty sampling is most widely applied among the non-clustering-based selection procedures due to its simplicity and speed [1]. All of these non-clustering-based selection procedures share a common problem, the sampling bias [2]. A typical active learning heuristic starts with an initial labeled training set and then iterate for convergence. When this labeled training set is not initially available, it is usually randomly sampled from the pool. However, according to the sampling bias effect described in [2], the performance of an active learner may highly rely on the initial labeled training set.

Hierarchical sampling is a clustering-based selection procedure proposed by Dasgupta and Hsu [2] to address the sampling bias problem. It tries to avoid sampling bias by first clustering on the unlabeled data set and each time sample more from less pure clusters. Although considered a good solution for sampling bias, hierarchical sampling has its own problem. There are evidences showing that hierarchical sampling manages to produce steeper learning curves early on than margin-based uncertainty sampling, but uncertainty sampling manages to achieve lower error rates later on [7]. This means that the later points sampled by hierarchical sampling are less informative than those by uncertainty sampling. This is intuitively natural since the hierarchical clusters can never be exactly related to the desired labels.

To sum up, if the initial training set is well distributed, less biased, uncertainty sampling performs better than hierarchical sampling. On the other hand, hierarchical sampling can avoid sampling bias, which leads to failure of uncertainty sampling, by focusing on less pure clusters. Consequently, will it be better to combine these two selection procedure to obtain both of the advantages? Here we propose our selection procedure as a combination of hierarchical sampling and uncertainty sampling.

2. UNCERTAINTY SAMPLING

Start with [3], there has been a whole group of selection procedure being regarded as uncertainty sampling which is very well studied and widely applied [4,5,9,11,12]. The basic idea of uncertainty sampling is to avoid querying data points that learner is already confident about [7]. By focusing on the most confusing data points, the learner is able to quickly shrink the version space in the case of using a margin-based classifier [7]. The basic selection procedure of uncertainty sampling is as follows:

1. Start with a pool of unlabeled data U and a initial set of labeled data L.
2. Train on L to get a model M.
3. Apply M on U.
4. If saturated (prediction no longer changes), end.
5. Else find the most uncertain data point $u \in U$, query the label of u and add u into L.
6. Go back to 2

There are various type of uncertainty measurement in uncertainty sampling. The most common ones are least confident, margin, and entropy. These are all measurements

of uncertainty based on probability model, while the uncertainty of a non-probability model can be measured by the distance from the decision boundary to the data point.

Uncertainty sampling is a great success due to its intuitive appeal combined with the ease of implementation. Engineers do not need much experience or specific knowledge to implement active learning with uncertainty sampling. However, uncertainty sampling suffers from the sampling bias much. Its performance heavily depends on the initial labeled training set. When provided an initial set with pool quality, uncertainty sampling has been observed to perform worse than random sampling [6, 8, 10].

3. HIERARCHICAL SAMPLING

Hierarchical sampling is a clustering-based selection procedure proposed by Dasgupta and Hsu [2] to address the sampling bias problem. Start with a hierarchical clustering on the unlabeled data pool, the algorithm keeps pruning clusters that are already pure and focusing on exploit less pure clusters. Since the clustering is only related to the hidden structure of the data, not necessarily the labels, the sampling will no longer rely on the trained model. Thus leads to two consequences, firstly, it is incrementally affected by sampling bias; secondly, it lacks the ability to refine what has been learnt. The basic procedure of hierarchical sampling is as follows:

1. Start with a pool of unlabeled data U.
2. Hierarchical clustering T on U.
3. Current Pruning $P = root\{T\}$.
4. Select a cluster node $v \in P$.
5. Pick a random instance x from T_v and query its label.
6. update counts for all cluster nodes u on path from $x \to v$.
7. choose the best pruning P'_v and labeling L'_v for T_v.
8. $P = P - \{v\} \cup P'_v$.
9. $L(u) = L'_v(u)$ for all $u \in P'_v$.
10. Go back to 4 until all pure.
11. Label every instance as the label of leaf node in P it belongs to.
12. Use the whole data set to train a classifier.

There are also theoretical guarantees that hierarchical sampling degrades gracefully into random sampling when clusters are not correlated with the labels [2]. The problem of hierarchical sampling is mainly the lack of ability to refine the classification model.

4. HIERARCHICAL AND UNCERTAINTY SAMPLING

Combining the advantages of both uncertainty sampling and hierarchical sampling, the proposed clustering-based selection procedure is as follows:

1. Hierarchical Sampling:
   (a) Start with a root node containing all the unlabeled data points.
   (b) For each leaf node, randomly sample $K - a$ points and query the label. a is the number of labels already retrieved in this node (from parent node), K is a predefined value.
   (c) If pure enough, go back to 1b, else split this node into two nodes by spectral clustering.
   (d) Stop when all the leaf nodes are pure enough. Go to 2a.
2. Uncertainty Sampling:
   (a) Train a classifier on the retrieved labeled data points.
   (b) Apply the classifier on the unlabeled pool. If saturated, end.
   (c) Else identify the most uncertain points for the classifier, query their labels.
   (d) Go back to 1b.

5. REFERENCES

[1] Z. Bodó, Z. Minier, and L. Csató. Active learning with clustering. *Active Learning Challenge Challenges in Machine Learning*, 6:141, 2011.

[2] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215. ACM, 2008.

[3] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156, 1994.

[4] D. Reker and G. Schneider. Active-learning strategies in computer-assisted drug discovery. *Drug discovery today*, 20(4):458–465, 2015.

[5] N. Roy and A. McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[6] H. Schütze, E. Velipasaoglu, and J. O. Pedersen. Performance thresholding in practical text classification. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 662–671. ACM, 2006.

[7] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[8] K. Tomanek, F. Laws, U. Hahn, and H. Schütze. On proper unit selection in active learning: co-selection effects for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 9–17. Association for Computational Linguistics, 2009.

[9] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[10] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 173–182. ACM, 2010.

[11] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua. Interactive video indexing with statistical active learning. *Multimedia, IEEE Transactions on*, 14(1):17–27, 2012.

[12] J. Zhu, H. Wang, T. Yao, and B. K. Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1137–1144. Association for Computational Linguistics, 2008.