Exploring the Frontier of Uncertainty Space

Rong Hu and **Patrick Lindstrom** and **Sarah Jane Delany** and **Brian Mac Namee**
Dublin Institute of Technology, Dublin, Ireland

Abstract

We aim to investigate methods balancing exploitation with exploration in active learning to improve the performance of uncertainty sampling. Two exploration guided sampling methods are compared to uncertainty sampling on various real-life datasets from the 2010 Active Learning Challenge. Our initial experiments seems to indicate that combining exploration with uncertainty sampling improves performance on certain datasets but not all.
Keywords: Active learning, Uncertainty sampling, Exploration

1. Introduction

Active learning (AL) methods with Support Vector Machines (SVM) have gained wide acceptance because of their significant success in numerous real-world learning tasks. Uncertainty sampling is commonly used to select the most informative instances to present to an oracle for labelling, which in the case of a SVM are the instances closest to the decision hyperplane.

In our study, we use two exploration guided sampling methods for SVM active learning based on the following ideas: 1) Instances near the hyperplane are likely to more informative; 2) Not all instances close to the hyperplane are needed as they may be similar to each other or similar to previously labelled instances. Instead a small percentage of the most representative instances is chosen by exploring the "frontier" of uncertainty space.

2. Method

We based our approach on *SIMPLE* (Tong and Koller, 2001), an uncertainty sampling method which uses a SVM to select instances closest to the hyperplane. Our selecting approach complements SIMPLE by exploring the instances chosen by SIMPLE and selects a representative sub-sample. This reduces the possibility of selecting duplicates and/or noisy instances.

The first approach uses affinity propagation clustering (APC) (Frey and Dueck, 2007) to cluster the N instances sampled by SIMPLE and chooses the cluster centers to present for labelling. We call this approach *SIMPLE+APC*. APC is a state of the art clustering algorithm and it has been shown to perform well in a variety of clustering problems. The second method employs EGAL (Hu et al., 2010) for the purpose of exploration. We call this method *SIMPLE+EGAL*. EGAL is an approach that combines measures of density and diversity to select instances which are in high density areas of the uncertainty space but also are the least similar to the existing labelled dataset.

We compare our augmented approaches to using SIMPLE alone on datasets from the 2010 Active Learning Challenge[1]. Similar to the challenge the prediction performance is measured by using the Area under the Learning Curve (ALC). At each iteration k instances are sampled for labelling. For SIMPLE the k instances closest to the hyperplane are selected for labelling. For the hybrid approaches SIMPLE+APC and SIMPLE+EGAL, SIMPLE selects the N where $N \simeq 10 * k$ instances closest to the hyperplane from which a subset of k instances are chosen for labelling by the exploration algorithm. The value of k was increased logarithmically. After eight iterations only SIMPLE is used.

3. Results & Discussions

The overall ALC performance figures with their ranks shown in parentheses are included in Table 1. It shows that both SIMPLE+APC and SIMPLE+EGAL outperform SIMPLE alone on IBN_SINA, NOVA and SYLVA. However on ORANGE and E, all approaches perform poorly with SIMPLE performing the best.

Table 1: Final Global Score (ALC) on Six Datasets

Approach	**IBN_SINA**	**NOVA**	**SYLVA**	ORANGE*	E*
SIMPLE	**0.7793 (3)**	**0.5036 (3)**	**0.8693 (3)**	**0.1516 (1)**	**0.2648 (1)**
SIMPLE+APC	**0.8340 (1)**	**0.5750 (2)**	**0.8833 (1)**	**0.1459 (3)**	**0.2099 (2)**
SIMPLE+EGAL	**0.8189 (2)**	**0.5846 (1)**	**0.8819 (2)**	**0.1460 (2)**	**0.1689 (3)**

This shows that our approach, which enhances the SIMPLE uncertainty sampling approach with an exploration guided technique, improves the prediction performance on some datasets but not all. One possible reason could be that our base classifier is not suitable on some datasets, which prevents it from finding a "good" area of uncertainty space to explore. In this type of scenario it would be better to first perform exploration in a wide area until a classifier which can identify good areas of uncertainty space can be built.

Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/RFP/CMSF718.

References

Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

Rong Hu, Sarah Jane Delany, and Brian Mac Namee. EGAL: Exploration guided active learning for TCBR. In *Proceedings of ICCBR 2010 (to appear)*, 2010.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2001.

1. http://www.causality.inf.ethz.ch/activelearning.php