SVM Based Active Learning with Exploration

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

Abstract

This paper proposes using exploration guided approaches to select both informative and representative instances to present for labelling in an active learning process.
Keywords: Clustering, Exploration, Support vector machines, Class imbalance

Framework: This paper proposes an exploitation/exploration framework for active learning (see Figure 1). The framework includes an initialisation stage and an active learning (AL) stage which iterates until a predetermined number of labels have been requested. The AL process has two steps - exploitation and exploration - both of which are parameterised allowing flexibility when dealing with different datasets.

Figure 1: System Architecture Overview Diagram

Initialisation Stage: The data is pre-processed to replace missing values with modes and means. The initial prediction of the first query is calculated using Euclidian distances to the labelled seed instance. Finally, the dataset is clustered using affinity propagation clustering (APC) (Frey and Dueck, 2007) and cluster centres are selected for labelling to seed the AL process.

Exploitation: The exploitation step identifies the N instances the classifier will benefit most from having labelled. The first selection strategy used is uncertainty sampling which uses an SVM classifier to query labels for the instances closest to the decision hyperplane. However, all datasets were suspected to be highly unbalanced, so we developed a second selection strategy namely *positive biased sampling* (posBias). posBias selects $N/2$ pairs of instances made up of $N/2$ instances closest to the hyperplane and $N/2$ positive instances for which the classifier gives the most certain classification.

Exploration: The optional exploration step selects the most representative of the exploitation-sampled instances. This avoids selecting similar instances or noisy instances, and explores

more of the uncertainty space. The exploration techniques used were APC and EGAL (Hu et al., 2010) which has been shown to be an efficient, exploration-based and classifier independent sampling method. Using one of these techniques a subset of the instances sampled by the exploitation step are selected for labelling.

Results: Figure 2 shows the learning curves for datasets F, D and B. On F (left) APC clustering was chosen as the exploration technique and used for all iterations. On D (middle) EGAL was used initially, then APC and finally no exploration. Results on F and D indicate that complementing exploitation with exploration can help to select the most informative instances to label. However the performance on B (right) is very bad. We hypothesise that the poor performance is due to too many changes during the AL process. Four key changes were highlighted: P1 - classifier parameters were changed; P2 - EGAL replaced APC; P3 - APC replaced EGAL; P4 - posBias was introduced and exploration was disabled.

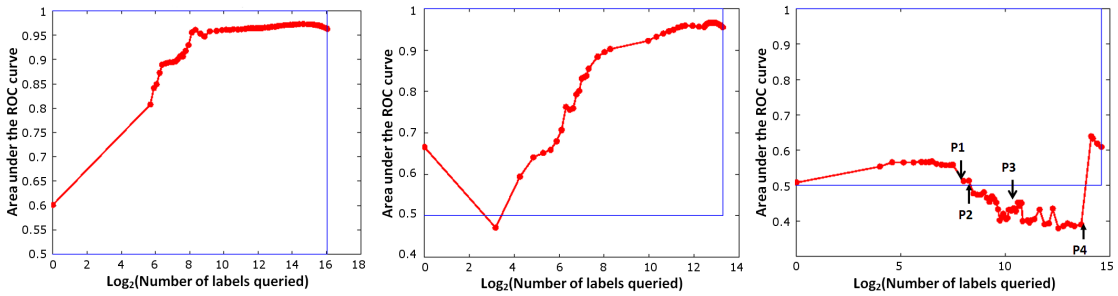

Figure 2: AL Learning Curves on F, D & B

Conclusions: Our preliminary findings show that augmenting uncertainty sampling with exploration can be beneficial. Our results also suggest that changes in the middle of the AL process can be detrimental. We believe this is due to the reusability problem: a set of labelled instances which is informative for one classifier is not necessarily informative for another classifier even for the same type of classifier with different parameters.

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.