CARNEGIE MELLON UNIVERSITY - COMPUTATIONAL BIOLOGY DEPARTMENT

Active Learning for Drug Selection on Identified Target Protein

Christine Baek
christib@andrew.cmu.edu

Qi Chu
qchu@andrew.cmu.edu

1 INTRODUCTION

In this project, we explore three datasets of different noise level, for identification compounds, or drugs that bind to specific target protein associated with disease. We use DHM as our active learning strategy to determine when to query the Oracle, and SVM to train our model on the oracle-obtained as well as inferred labels.

2 METHODS

Input :

- Training data consists of 4000 training instances, each with 1000 features, and true label
- Test data consists of 1000 testing instances, each with 1000 features, and true label
- Blind prediction data consists of 1000 instances, each with 1000 features

TODO: discuss [4]

Based on this, we chose DHM as our base learner strategy, and SVM as our classifier, elaborated below.

2.1 BASE LEARNER STRATEGY

DHM was chosen as our base learner strategy [2]. DHM is a good general strategy in agnostic setting, and provides an extended version of selective sampling scheme of CAL algorithm

[1], [2]. We compare the performance of our algorithm against random learner to test the effectiveness of the algorithm in selecting meaningful datapoints for query into the Oracle.

TODO: talk about modifications made to the algorithm

2.2 Classifier Strategy

TODO: elaborate

SVM was chosen as our classifier for multiple reasons. SVM finds a hyperplane that maximizes the distance between the nearest data (support vectors) and the hyperplane. Given the high dimension ofs our input data (1000 features), it was important that a classifier that scales easily to high dimension, with no local optima. Gaussian kernel (default) was used for training.

3 Results

3.1 Easy

Figure 3.1: Error rate of easy test set plotted against number of queries made to oracle

Figure 3.2: F1 Score of easy test set plotted against number of queries made to oracle

3.2 Moderate

Figure 3.3: Error rate of moderate test set plotted against number of queries made to oracle

Figure 3.4: F1 score of moderate test set plotted against number of queries made to oracle

3.3 Difficult

Figure 3.5: Error rate of difficult test set plotted against number of queries made to oracle

Figure 3.6: F1 score of difficult test set plotted against number of queries made to oracle

4 Conclusion

TODO: briefly summarize

References

[1] Cohn, Atlas, Ladner, *Improving Generalization with Active Learning*, Machine Learning May 1994, Volume 15, Issue 2, p 201-221

[2] S. Dasgupta, D. Hsu, C. Monteleoni, *A general agnostic active learning algorithm*, NIPS, 2008

[3] S. Dasgupta, *Two faces of active learning*, `http://cseweb.ucsd.edu/~dasgupta/papers/`, 2010

[4] Warmuth, Liao, Ratsch, Mathieson, Putta and Lemmen, *Active Learning with Support Vector Machines in the Drug Discovery Process*, J. Chem. Inf. Comput. Sci. 2003, 43, 667-673