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 modify the strategy used in [4] for query selection, and use SVM to learn our model. We also perform a gradient experiment to learn the optimal balance for allocating queries between initial and active learning.

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
- Each data(row) represents potential compound that may bind to the target
- Each feature(column) represents structural or chemical

As with real world drug target screening, we began with assumption that only an extremely small fraction of compounds will bind to the target. Also, only a small fraction of the features would be relevant in predicting the label (are actual sites that the drug can potentially bind to).

Our challenge is to first find which features are relevant in accurately predicting whether the compound will bind to the target or not, and then learning on those features.

Because of the imbalanced dataset (only small portion of the molecules actually bind), it would have been dangerous to use active learning strategy without any prior. We decided to split our learning into two different phases : Initial and Active. During initial learning, we randomly select molecules without any assumptions or selection criteria, other than choosing molecules we have not seen before. Once initial learning is over, we begin the active learning phase where points are selected per *largest positive* strategy as discussed in [4], elaborated below. Total number of queries made during both initial and active phase total 2500, the allotted budget.

We performed gradient experiment, and measured the performance based on the number of queries made between initial learning and active learning, as the total number of queries is capped at 2500.

2.1 Base Learner Strategy

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

TODO: add summary 2D chart, about the initial/active learning and the corresponding F1 scores

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