Active Learning For DBSI: DNA Binding Site Identifier

Max Horowitz-Gelb

December 12, 2016

Abstract

DBSI is a structure based method for predicting the positions of protein interaction sites associated with DNA binding [5]. Here I present a method for applying active learning to the training of DBSI. This method optimizes the training of DBSI in a way that considers the batch style form of labelled data collection necessary when creating a training set. This method shows slight improvements in efficiency in comparison to naive methods.

Introduction

For area under an ROC curve, DBSI has been shown to achieve 88%, a high degree of separability. The score was achieved by training DBSI on a set of 263 unique proteins. The model as a result of this training is now accessible to anyone on a public server [3]. The quality of this model could be improved further by training with more labelled data. But to collect more data practically we need to address practical issues.

Necessity for Active Learning

Unlike learning on data generated from sensors or internet activity, aquiring labelled data for DBSI requires considerable time and energy. Experiments must be done to probe each protein for interaction sites.

Find out how protein complexes are aquired to emphasize lab work.

Neccessity for batch style active learning

Due to the way training data for DBSI is aquired, standard active learning methods are not appropriate. This is because each training point given to DBSI corresponds to one residue in a protein. Therefore each protein that is probed will give hundreds of new training points. Because of this we must query training

points in batches as opposed to querying individual residues. This puts us in a unique situation where the batch with the most informative set training points as a whole may be different than the batch containing the one individually most informative training point. To address this issue I use a active learning querying method which scores batches rather than individual points in order to select the next query protein.

Methods

DBSI uses a support vector machine to classify residues. Such machine learning model has geometric properties that make applying active learning quite intuitive.

Non-Batch Active Learning SVM

When considering a standard binary classifier SVM as DBSI uses, our decision function is can be described by its training examples [1]:

$$g(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

and classifcation is simply the sign,

$$f(x) = sign(g(x))$$

Where x_i is the set of support vectors, x is our vector we want to classify, α_i is a model parameter, and k is our kernel function.

If our kernel satisfies Mercer's condition, it can be further rewritten [2]

Then some shit about version spaces [4]

[2]

Results

Conclusion

References

[1] Scholkopf B and Smola AJ. Learning with kernels: Support vector machines, regularization, optimization, and beyond. 2002.

[2] Brinker K. Incorporating diversity in active learning with support vector machines. In *In Proceedings of the 20th International Conference on Machine Learning*, pages 59–66. AAAI Press, 2003.

[3] Sukamar S, Zhu X, Ericksen SS, and Mitchell JC. Dbsi server: Dna binding site identifier. *Bioinformatics*, 2016.

[4] Mitchel TM. Generalization as search. *Artifical Intelligence*, 18:203–226, 1982.

[5] Zhu X, Ericksen E., and Mitchell JC. Dbsi: Dna-binding site identifier. *Nucleic Acids Research*, 2013.