Using MEGA to Predict Molecular Bio-Activity

Arun Qamra, [1] King-Shy Goh, [2] Edward Y. Chang, [3]

Keywords: drug design, structure-activity relationships, machine learning, MEGA

1 Introduction

The discovery of a new drug typically requires over 10 years and up to a billion dollars. Machine learning algorithms have the potential to reduce experimental time and cost by intelligently guiding the discovery process. Drug molecules work by binding to protein molecules in the body and modulating their actions. Hence, the first step in drug design is to find compounds that bind with the desired "target" protein. Traditionally this is done by experimentally evaluating activity of a large number of compounds against the target. It is known that the chemical behavior is (largely) dictated by a compound's structure. Machine learning methods can thus be used to learn structure-activity relationship models, which can then be used to virtually screen compounds for activity. Recently, techniques such as Neural Networks, Bayesian Networks [2], and SVMs [3] have been applied to do so. However, this problem presents unique challenges to machine learning, such as the large number of features, limited training data, and significant positive/negative imbalance. We propose the use of MEGA [1] for learning activity models, demonstrate that MEGA can accurately predict activity, and using intelligent sampling, it can do so with much less training data. Another significant advantage is that the model MEGA learns can be interpreted.

2 The MEGA Algorithm

MEGA models concepts in k-CNF. A k-CNF expression, $c_1 \wedge \cdots \wedge c_L$, is a conjunction of terms c_i, where each c_i is a disjunction of at most k predicates (all features and their negations are used as predicates). MEGA also maintains, and refines at each iteration, a k-DNF expression (disjunction of k-predicate conjunctions) that represents the candidate sampling space. MEGA starts with the most specific k-CNF and the most general k-DNF, and for each training instance, removes terms so as to generalize the k-CNF or specialize the k-DNF, depending on whether the instance is positive or negative. The k-CNF expression left after training is used as the learnt model to classify unseen data. MEGA can perform Active Learning by iteratively selecting for labeling, the most informative unlabeled samples. MEGA selects these points based on the candidate sampling space and the learnt concept. This allows an accurate concept to be learnt in a few iterations from a few most informative training instances. Please refer [1] for details. MEGA is appropriate for the drug discovery problem for a number of reasons. The active learning approach used by MEGA is very suitable for the typically iterative drug discovery process, and allows learning from few training instances, thus drastically reducing experimental costs. A significant advantage of MEGA is that the model learnt by MEGA is interpretable since it is a logical expression. Interpretability of the learnt model can provide valuable insights into molecular bio-activity and aid the design and discovery of appropriate drugs. Another advantage is that MEGA can start learning even in the absence of positive instances (unlike most other methods), since

[1] Computer Science, Univ. of California Santa Barbara, arun@cs.ucsb.edu

[2] Electrical & Computer Eng., Univ. of California Santa Barbara, kingshy@engineering.ucsb.edu

[3] Electrical & Computer Eng., Univ. of California Santa Barbara, echang@ece.ucsb.edu

negative instances can shrink the sampling space and increase the probability of a positive instance being sampled in future iterations. This again is well suited to drug discovery since finding initial active compounds may not be easy.

3 Experiments, Results and Discussion

We evaluated MEGA's performance with the Thrombin dataset from the KDD Cup 2001 competition [2], and compared it with that of the KDD Cup winner. The task is to predict binding to thrombin. The training set contains 1909 instances, including 42 actives, while the test set contains 634 instances. 139, 351 binary features are used to describe structural and physical properties of each compound. The problems of high dimensionality, training data scarcity, and positive/negative imbalance, are evident here.

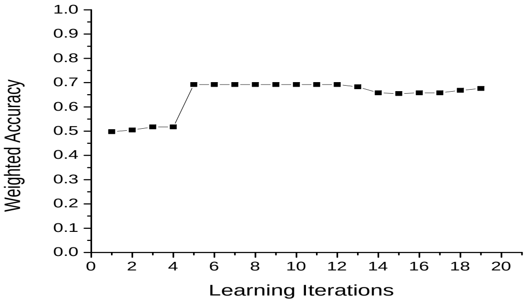

Figure 1: Weighted Accuracy vs Learning Iterations

Before using MEGA to learn the concept from the training data, we used Mutual Information to chose the 100 most important features. For the test set, which contains 150 positives and 484 negatives, the classifier learnt classified with an overall Accuracy of 76.5% and a Weighted Accuracy of 67.6%, generating 75 false positives and 74 false negatives. In comparison, the KDD Cup winner (a Bayesian classifier) gave an overall Accuracy of 71.1% and a Weighted Accuracy of 68.4%, generating 128 false positives and 55 false negatives. Weighted Accuracy is defined as the average of prediction accuracy for positives and that for negatives. Space does not permit us to present comparisons with other techniques, but performance comparable to the KDD Cup winner is encouraging, and given the advantages enumerated above, we can say that MEGA is definitely a promising technique. We next conducted experiments to use MEGA's intelligent sampling to iteratively select samples from the training set for active concept learning. Figure 1 shows the Weighted Accuracy achieved at each iteration, where 100 samples were sampled at each iteration. From the graph, we see that the classifier shows high accuracy after just 5 iterations. With intelligent sampling, we can learn an equally good classifier with just 500 instances instead of the 1909 originally used, thus resulting in drastically reduced experimental costs. Learning here is limited by the size of the given training dataset. In a real scenario, further experiments could be conducted to create more informative training data based on MEGA's recommendations, and thus potentially achieve even higher accuracies at low cost.

References

[1] Chang, E., and Li, B. 2003. MEGA — The Maximizing Expected Generalization Algorithm for Learning Complex Query Concepts. In: *ACM Transactions on Information Systems (TOIS)*

[2] Cheng, J., Hatzis, C., Hayashi, H., Krogel, M.-A., Morishita, S., Page, D., and Sese, J. 2001. KDD Cup 2001 Report. In: *ACM SIGKDD Explorations*, 3(2), pp.47-64

[3] Warmuth, M. K., Ratsch, G., Mathieson, M., Liao, J., and Lemmen, C. 2002. Active Learning in the Drug Drug Discovery Process. In: *Advances in Neural Information Processing Systems*