Performance Prediction of the Influence Relevance Voter

C.-A. Azencott, J. S. Swamidass, and P. Baldi
School of Information and Computer Sciences
Institute for Genomics and Bioinformatics
University of California, Irvine
Irvine, CA 926297-3435, USA
cazencot@ics.uci.edu

February 18, 2009

Virtual High-Throughput Screening (vHTS) is the cost-effective, in silico complement of experimental HTS. A vHTS algorithm uses data from HTS experiments to predict the activity of new sets of compounds in silico. Here we present two new algorithms for vHTS problems. The first is a new metric and visualization approach to better assess the results of vHTS experiments and overcome the limitations of existing metrics, such as ROC curves, which assess classification performance over the entire data but do not focus on the real quantity of interest, namely "early enrichment". The second is a new vHTS algorithm, the Influence Relevance Voter (IRV), which refines a k-nearest neighbor classifier by non-linearly combining the influences of a chemical's neighbors in the training set, using a highly constrained neural network. Influences in turn are decomposed, also non-linearly, into a relevance component and a vote component [1].

The IRV is benchmarked using the data and rules of two large, open, competitions, and its performance compared to the performance of other participating methods, as well as of an in-house Support Vector Machine (SVM) method. On these benchmark datasets, IRV achieves state-of-the-art results, slightly better than the SVM in one case (56% of the actives in the top 5% of its prediction-sorted list, compared with 55%), and significantly better than the SVM in the other, retrieving three times as many actives in the top 1% of its prediction-sorted list. On both benchmarking datasets, enrichment factors can be estimated for all possible cutoffs with an average absolute error of less than 5%. Moreover, the IRV can be used to correctly evaluate the total number of active compounds in each of the datasets within a margin of about 0.5%.

The IRV presents several important advantages over SVMs and other meth-

ods: (1) the output predictions have a probabilistic semantic; (2) the underlying inferences are interpretable; (3) the training time is very short, on the order of minutes even for very large data sets; (4) the risk of overfitting is minimal, due to the small number of free parameters; and (5) additional information can easily be incorporated into the IRV architecture. Combined with its performance, these qualities make the IRV particularly well suited for vHTS problems.

References

[1] S. J. Swamidass, C.-A. Azencott, T.-W. Lin, H. Gramajo, S. Tsai, and P. Baldi. The Influence Relevance Voter: an Accurate and Interpretable Virtual High-Throughput Screening Method. *J. Chem. Inf. Model.*, 2009. in press.

Topic Learning Algorithms

Preference Oral

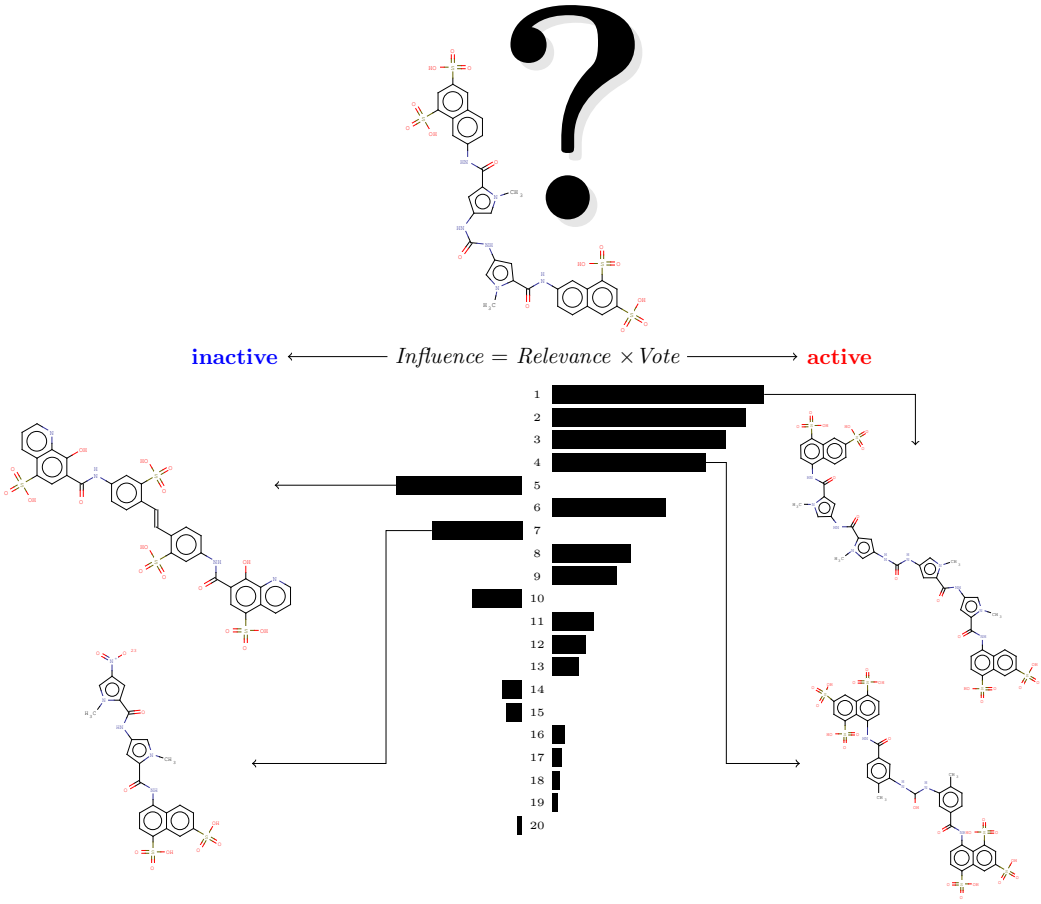

Figure 1: The Influence Relevance Voter (IRV) is an interpretable method that extrapolates information from high-throughput screening experiments to infer the probability that an unknown compound is active. In the image, the influences, computed from an IRV on an accurately predicted hit from the National Cancer Institute's HIV screening dataset, are displayed as a bar graph. The experimental data both supporting and countering the prediction is readily apparent. Compounds on the right are structurally similar to the hit and listed as active in the screen. Compounds on the left are structurally similar to the hit and listed as inactive in the screen. The structures of two active and two inactive neighbors are displayed.