Collaborative filtering for drug discovery

Dumitru Erhan

Department of Computer Science and Operations Research

University of Montreal

Credits

- Pierre-Jean L'Heureux
- Yoshua Bengio
- Olivier Delalleau
- Shi-Yi Yue & AstraZeneca

Motivation

- amazon.com
  - Customer ratings, tags
  - Customers also bought^{TM}
  - What do customers ultimately buy?
  - Items to go with / items instead of
- last.fm, radiolibre.ca
  - Ratings
  - Recommended artists, songs
- $$$

Motivation II

- Preferences of a user == 1 task
- Preferences of many users == many tasks
- Previous work:
  - Multi-task NNets (Caruana)
  - Bayesian approaches (Baxter)
  - Theory (Baxter)
  - “Modern” (==kernel) ways (Evgeniou)
- *Learn a new task by using intelligently information from other tasks*

Drug Discovery

- Complicated, expensive process
- High-Throughput Screening ($$$)
- Interested in QSAR (Quantitative structure-activity relationship)
- Bunch of molecules (“items”), a target (“user”)
- Virtual Screening (saves $$$)

Drug Discovery II

- Assume:
  - lots of molecules
  - a family of “related” targets
  - a new target
- Dense “user-item” ratings table
- Few rows
- *Predict new (very long!) row*

Formally speaking

- Data:
  - N molecules $x_i, i = 1 \ldots N$
  - M targets $t_j, j = 1 \ldots M$
  - Set of ratings/activity values R_{ij}
- Accuracy function: Lift
- Wanted: *a model that generalizes well to data for a new target wrt to the accuracy function*

The Lift

- $a =$ number of actives in general
- $N =$ number of molecules
- $a_s =$ number of actives at a given thr.
- $N_s =$ number of molecules at a given thr.
- Lift $= 100 \cdot \frac{a_s/N_s}{a/N}$

Testing hypotheses

- Given a multi-target dataset
- For each target:
  - TrainValidSet = OtherTData [+ DataFromThisT]
  - TestSet = TheRestOfDataFromThisT
- Two cases:
  1. tfraction = 0, is Lift > 100?
  2. tfraction > 0, is Lift > single-target Lift?

Neural Network Approach

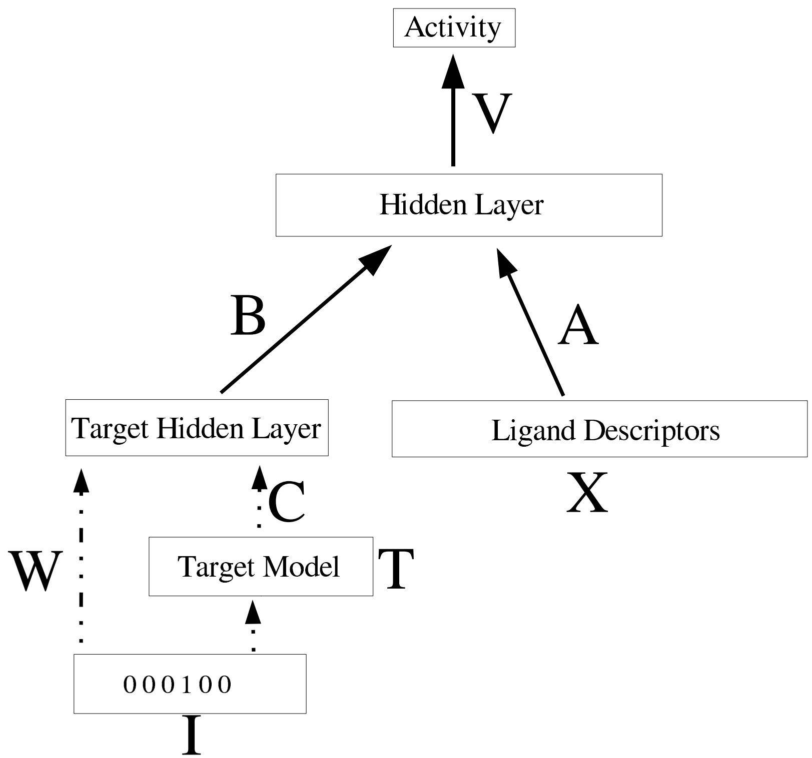


Figure 1: Collaborative Filtering Neural Network

Kernel Approach

- What if we defined two kernels:
  1. $K_X(x_i, x_k)$
  2. $K_T(t_j, t_m)$
- and combined them?
- Correct way of doing it:

$$K[(x_i, t_j), (x_k, t_m)] = K_X(x_i, x_k) \cdot K_T(t_j, t_m)$$

- Tensor product in the feature space
- Kernels: identity, gaussian, correlation, quadratic, polynomial, etc.

JRank

- Kernel Perceptron Ranker
- A set of learned parameters for each pair:

$$F((x_{new}, t_{new}); \alpha) = \sum_{i,j} \alpha_{i,j} K([(x_{new}, t_{new}), (x_i, t_j)]$$

- A set of learned thresholds used for ordinal regression (by binning F)
- Standard perceptron algorithm, updates when the bin is wrong
- Step-size of the update $\propto$ how wrong

Support Vector Machine

- Interpret dataset as classification problem
- Can then use the standard SVM routines:

$$W(\alpha) = \sum \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K[(\mathbf{x}, \mathbf{t})_i, (\mathbf{x}, \mathbf{t})_j]$$

- Problem 1: explosion in number of datapoints
- Problem 2: *quadratic* optimization problem
- Problem 3: choice of kernels limited

Target selection

- Problem: VERY BIG dataset
- Select “related” targets by “direct activity correlation”

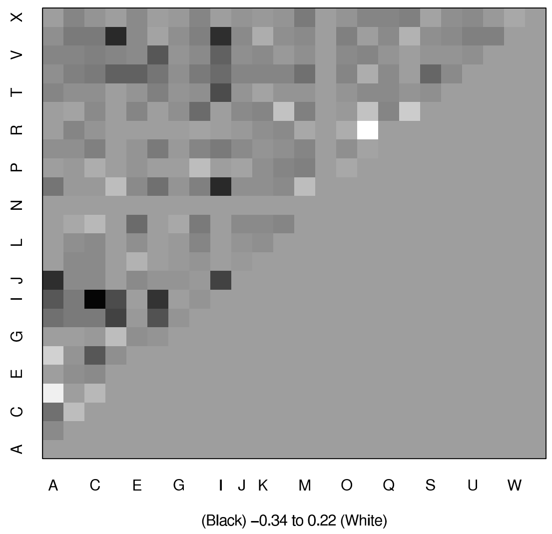
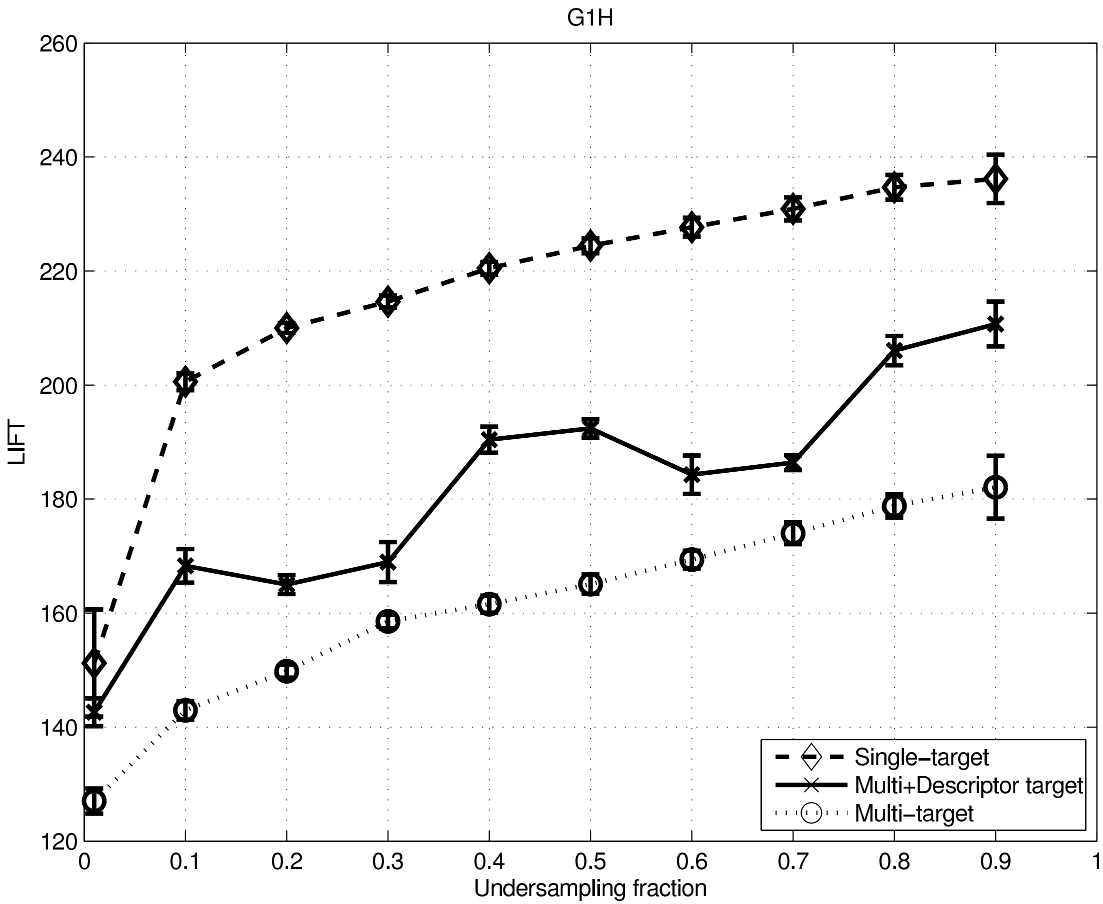
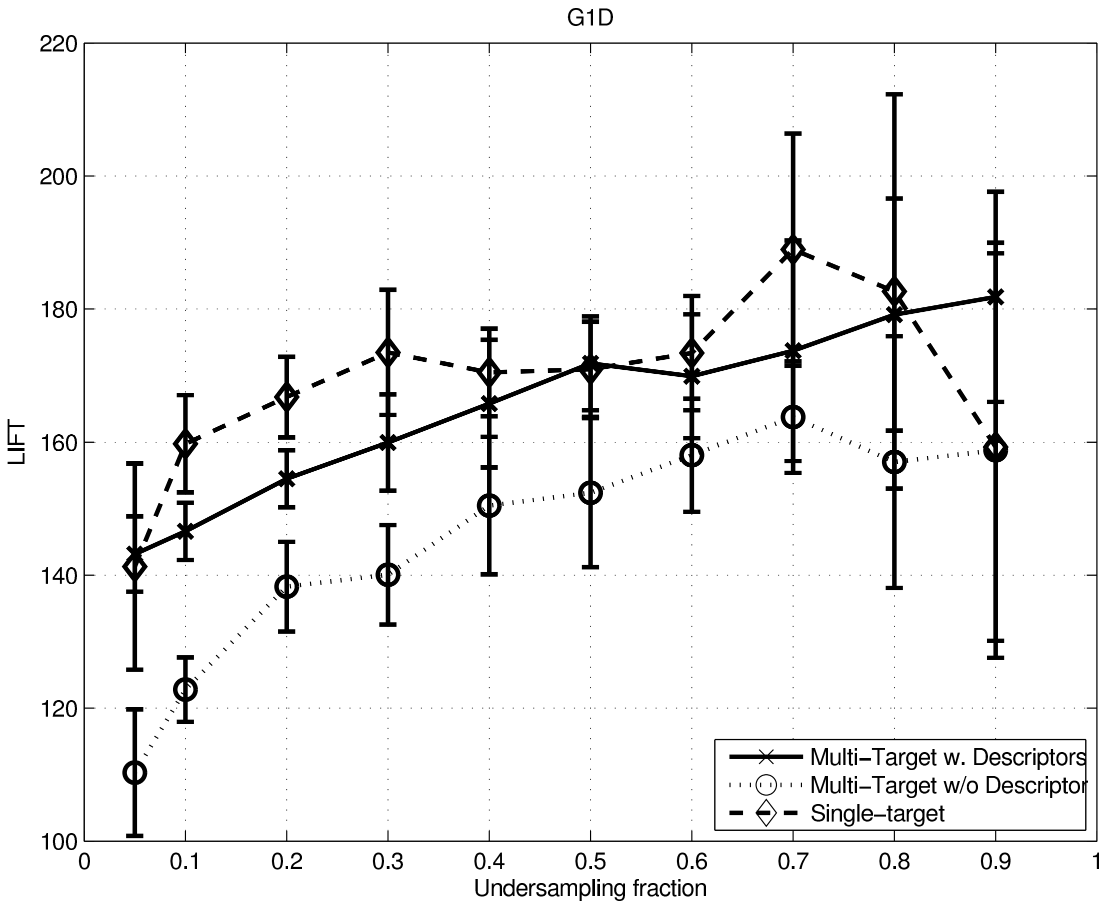


Figure 2: Pairwise Correlation of Biological Activity

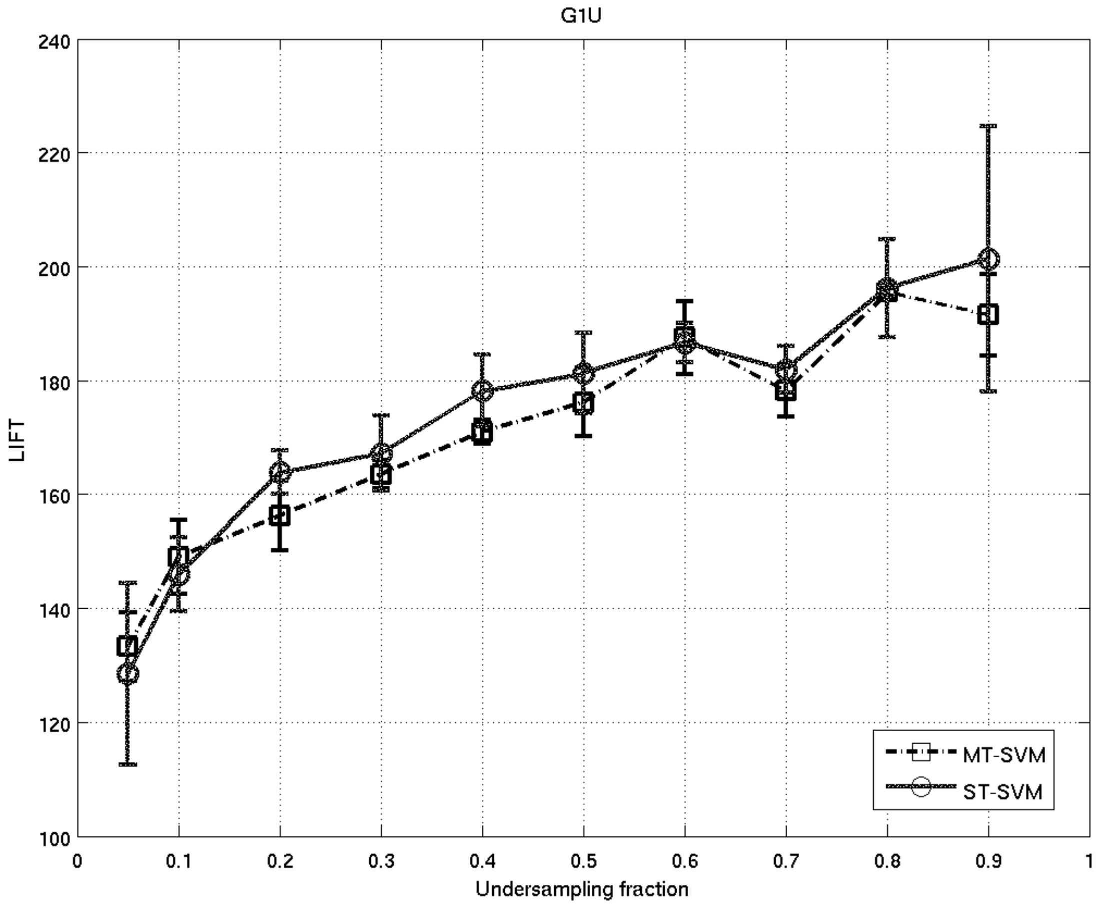
Experimental Results – NNet



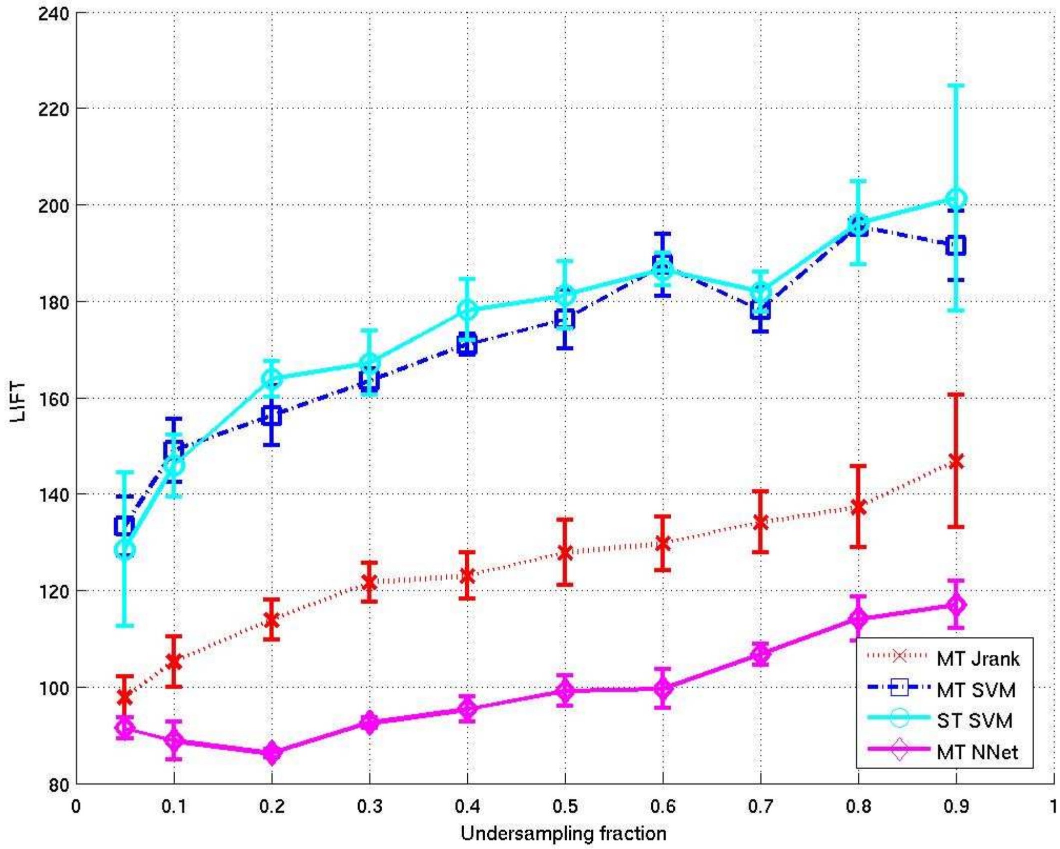
Experimental Results – JRank



Experimental Results – SVM

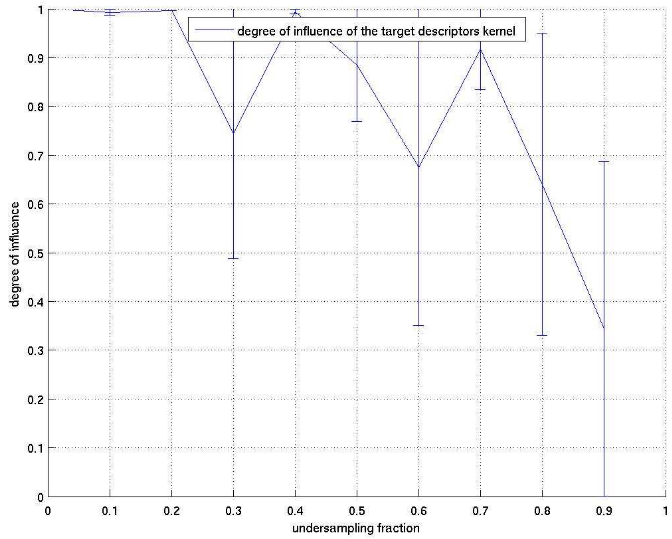


Experimental Results – Comparison



Experimental Results V

- tfraction = 0: 3 out of 7 targets with Lifts = 130...160
- NNet & JRank: lifts are close to 100
- Degree of influence



Discussion

- MT NNet and MT JRank don't work as expected
- MT SVM works as well as ST SVM (which is state-of-the-art)
- Encouraging fact: at tfraction=0 Lifts are > 100
- Means "inductive transfer" is happening!
- Need to find better way to exploit it

Future Work

- Other target descriptors + more targets
- Better NNet
- Better kernels
- Ordinal regression with SVM, other SVM approaches

Conclusions

- Collaborative filtering
- Learning a new task
- Drug discovery
- Virtual screening, new target
- NNets, JRank, SVMs

The Conclusion

It’s all about the Benjamins!

(with apologies to Puff Daddy)

References

- Rich Caruana, "Multitask Learning." *Machine Learning*, vol. 28(1), pp. 41–75, 1997.
- Jonathan Baxter, "A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling." *Machine Learning*, vol. 28(1), pp. 7–39, 1997.
- Jonathan Baxter, "A Model of Inductive Bias Learning.", *J. Artif. Intell. Res. (JAIR)*, vol. 12, pp. 149–198, 2000
- Theodoros Evgeniou and Massimiliano Pontil, "Regularized multi–task learning". *KDD 2004*, pp. 109–117
- Justin Basilico and Thomas Hofmann, "Unifying collaborative and content-based filtering." *ICML 2004*.