

Carnegie Mellon University School of Computer Science

An Improved Matrix Completion Algorithm For Categorical Variables: Application to Active Learning of Drug Responses

Huangqingbo Sun, Robert F. Murphy

Workshop on Real World Experiment Design and Active Learning at ICML 2020





Thirty-seventh International Conference on Machine Learning

Drug Discovery Funnel





Active Learning - Multiple Phenotypes

- Solution is active learning of a predictive model of all compound effects on all targets
- But there are also many possible effects that compounds could have on a given target thus effects are categorical variables
- Assume that there are some similarities in effects among compounds and targets
- Predictive model: completion (imputation) of a very sparse (only a few observed entries) categorical matrix
- For active learning, uncertainty sampling is adopted, with 3 query strategies.



Experiment on Synthetic Data

• How fast does Active Learning comparing to random selection?

10% -	12%	5%	7%	3%	2%	4%	4%	2%	3%		
20% -	14%	7%	4%	4%	5%	6%	5%	5%	5%		
30% -	10%	6%	7%	4%	6%	6%	8%	8%	9%		
40% -	27%	7%	5%	6%	9%	9%	12%	13%	13%		
- %00 -	15%	9%	7%	8%	9%	11%	15%	16%	16%		
60% -	15%	11%	7%	11%	14%	16%	17%	17%	20%		
70% -	17%	9%	11%	10%	12%	14%	17%	18%	20%		
80% -	22%	8%	10%	10%	10%	12%	13%	16%	18%		
90% -	25%	4%	6%	4%	6%	8%	9%	9%	11%		
10% 20% 30% 40% 50% 60% 70% 80% Responsive											

	10% -	27%	36%	23%	18%	20%	17%	17%	23%	20%
	20% -	44%	28%	27%	55%	37%	59%	51%	61%	53%
	30% -	23%	40%	58%	58%	57%	45%	67%	64%	55%
	40% -	34%	46%	51%	50%	57%	60%	59%	57%	65%
Unique	50% -	20%	40%	48%	53%	53%	53%	54%	57%	57%
_	60% -	15%	35%	43%	41%	45%	46%	47%	47%	49%
	70% -	11%	19%	34%	34%	35%	35%	36%	37%	39%
	80% -	8%	12%	13%	23%	24%	23%	25%	27%	29%
	90% -	3%	7%	8%	12%	11%	12%	11%	14%	12%
		10%	20%	30%	40 ['] % Re	50%	60 ['] %	70%	80%	90%



Performance was measured as the difference in the number of batches to achieve 100% (right) or 90% (left) accuracy between active and random selection.

Experiment Using Microscope Images for Many Drugs and Targets



Learn the effect of 92 drugs on 94 GFPtagged proteins without doing experiments for all drugs and proteins with the help of Active Learning.





Image source: Naik et al.

Conclusions

- Improved clustering-based, "lazy learning" matrix completion algorithm for categorical matrices.
- Results in improved active learning performance over previous methods.

