THE INFEERENCE OF GROUND TRUTH FROM BIOLOGICAL LITERATURE

Alexander Belikov, James Evans, Andrey Rzhetsky
The University of Chicago

Introduction

It is known that in the process of scientific discovery is not straightforward. The intrinsic properties of certain scientific phenomena, such as

- the complexity of context, surrounding a statement
- the sociological bias towards positive results
- the complexity of the required experimental setup, prevent rapid convergence or even render certain statements unverifiable.

We try to find out if it is possible to infer:

- assess the correctness of a claim made by a scientific team relative to the "objective" biological interaction

given a set of claims made by scientific teams, what the likely sign of the interaction is

A typical record in the Geneways database has the form: "alb prevents "tett". We aggregate all positive and negative lexicographic actions (with the simplifying assumption that a claim with a positive action and negative flag set to true translates to a negative action) and project these expression onto the boolean space (true, false).

Features

Popularity

$c_i$ is the set of claims pertaining to interaction $i$. The number of claims made before time $t$ is the observed popularity $n_i(t) = |c_i(t') : c_i(t') \subset C_i, t' < t|$. Also density. Log-normal works for a small fraction (power-law distribution for interactions). WIP: capture the fluctuations.

Network measures


Time

Time passed from the first publication regarding the interaction $\Delta_i = t_i - t_{i_0}$.

Journal/Affiliation Quality

Using the Article Influence (proxy to PageRank) to rank journals (using Web of Science, more than 50M publications). Affiliation quality taken from Q8 World University ranking. Using hierarchical clustering to disambiguate affiliation (cf. App.).

Citation metrics

Citation data was fit with a log-normal distribution (alternatively we use the window of 20 years).

Herfindahl index

We compute "the share of the market" $s_i(t)$ for each interaction for institution and authors and from there derive the normalized HI $H_i(t) = \sum s_i(t)$. NIH tells us how uniform or monopolistic the current domain is. Social independence.

Appendix

- Disambiguation problem - hierarchical string distance: Affiliations are provided as ordered sets of phrases, e.g. "Dept. Anim. Sci., Rutgers University, New Brunswick, NJ 08903". Solution: given a string metric, compute word to word distance, define the alphabet of words, compute the distance on metawords, etc. Cluster on the set of phrases distances.

- Stratified random forest (Forest of Forests): Cross-validation is performed on the observed frequencies of classes, however, for the training (depending on the metrics we are optimizing) we want to sample more the underrepresented classes. Solution: grow many forests sampled from training set with less extreme proportions. Improves AUROC by 30-40%.

- Unsupervised approach, graphical models:

Assume there are latent states which have Markovian or coarse Markovian structure and which generate the observable variables.

Generalized bin-packing (applied to GM inference): Split the dataset in batches, so that the running times are similar, while sampling in stratas for feature pools to facilitate the convergence of the posteriors.

Datasets

- Geneways: 130K unique interactions, 198K unique claims, from 109K publications
- Literome: 144K unique interactions, 279K unique claims from 220K unique publications
- LINCS L1000: 77 cell-lines, 5K perturbagens, 11K downstream genes
- Geneways-LinCs projection: 23K unique interactions, 44K unique claims from 33K unique publications.
- Literome-LinCs projection: 36K unique interactions, 75K unique claims from 69K unique publications.

The overlap between Geneways and Literome (aligned with LINCS) is 89 claims (0.32 corr). Before alignment the overlap is 0.9K claims (0.35 corr). Correlation between LINCS and Geneways or Literome $\approx 0$.

Setup

Index $\pi$ refers to a specific interaction. $\pi^*$ is the continuous strength of the interaction (score of over-expression from LINCS, which is also aggregated and tested for consistency). We discretize it, splitting the cdf of $\pi^*$ into 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and assign to the intervals “negative”, “indifferent” and “positive”. We fix $\epsilon = 0.2$ and define in this manner $\pi^*$ taking discrete values 0, 0.5 and 1.

After projecting GW and Literome on LINCS, $\pi^*$ acquires a bias (mostly positive interactions in literature) for Literome we have 50.4K claims, 4.8% pos, 1.5% neg, 93.7% ind; for Geneways 15.5K, 10.5% pos, 1.5% neg, 88.5% ind.

TPF3 (7.75%) on CDK11A1 (1036) interaction excluded from Literome dataset.

We define the distance $\alpha_i^* = |\pi_i^* - \pi_i^*|$ and train classifiers $P_i(c_i^* | f_i, \alpha_i^*)$, where $t' < t$ assuming conditional independence at time $t$. NB: some of the features explicitly depend on $c_i^*$ at earlier times. By symmetrizing, we reduce the number of parameters (rare classes are not numerous).

Having defined $\alpha_i$, we can infer the experimental strength from a sequence of claims $(c_i, f_i)$, inverting the conditional probabilities using Bayes law under the simplifying assumption (index $\alpha$ omitted):

$$P_i(\pi | c_i, f_i) \propto \frac{P_i(c_i | f_i) P_i(\pi)}{\prod_j P_i(c_j | f_i) P_i(\pi)} = \prod_j P_i(c_j | f_i) P_i(\pi) \sum_c P_i(c_i | f_i) P_i(\pi)$$

Results

- Geneways
- Literome

Conclusions

a) positive features are the size of the community, the social attention, the latency
b) publication metrics have no effect on the correctness

c) the correlations are zero between Geneways, Literome and LINCS

d) Geneways and Literome independently select the same features in the same order

www.github.com/alexander-belikov
June 27, 2018
belikov@uchicago.edu