Prediction of ADRs using NMF and Weighted NMFs

Ollscoil na Gaillimhe
University of Galway

Arshad Ahamed Shajahan, Sai Shasank Dandu

Background and Objectives

Adverse drug reactions (ADRs) are a common problem in clinical and pharmacovigilance research and can lead to serious patient harm and biased conclusions if modelled poorly. Sparse, noisy, and highly imbalanced drug ADR data often cause standard machine learning methods to perform no better than naïve frequency-based approaches, unless appropriate low-rank and kernel-based methods are used with clear assumptions. [^1]: https://doi.org/10.1093/bioadv/vbae009

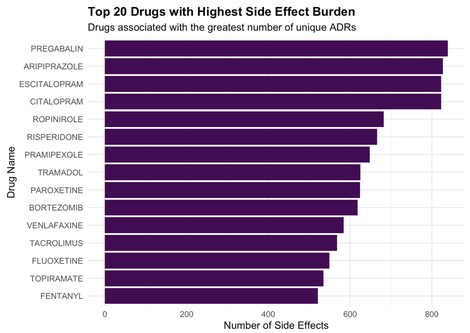
Objectives

1. Explain the ADR profile prediction problem and its statistical challenges in imbalanced, noisy health data.
2. Explore low-rank and kernel-based methods (e.g. NMF+kernel regression and alternatives) for ADR prediction.
3. Apply these methods to updated drug–ADR datasets with additional molecular and interaction features.

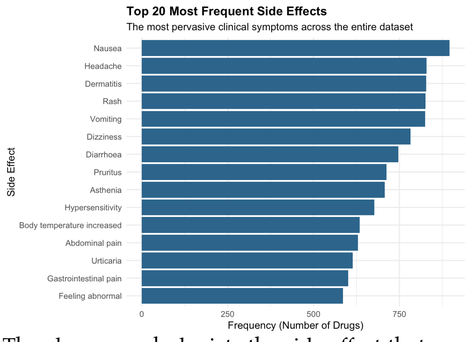
Data Sources and Descriptive Statistics of Datasets

The primary source of data are DGIdb 4.0, SIDER 4.1 and PubChem database.

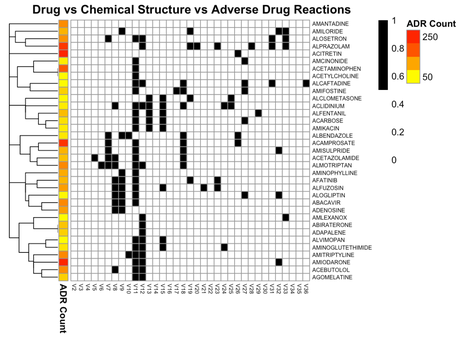
1. Drug-gene interaction pair: Intersection of drugs from DGIdb 4.0 and SIDER 4.1, generated to the binary matrix form.
2. Chemical fingerprints: Data from PubChem database, generated to the binary matrix form.
3. Drug & side effects: Drug along with it's side effects are extracted from SIDER 4.1 database.



The above graph mentions the drugs that has the most side effects.



The above graph depicts the side effect that was most common among the drugs.



Early Results / Descriptive Statistics of Datasets

Usually you want to have a nice table displaying some important results that you have calculated. In `posterdown` this is as easy as using the `kable` table formatting you are probably use to as per typical R Markdown formatting.

You can reference tables like so: Table 1. Some basic summaries of the dataset are below:

Table 1: Table caption.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
5.4	3.9	1.7	0.4
4.6	3.4	1.4	0.3
5.0	3.4	1.5	0.2
4.4	2.9	1.4	0.2
4.9	3.1	1.5	0.1

Figure 1, and Figure **??** below show the patterns in our dataset. Make sure that all the details in your plots will be legible when printed (legend text, axis text, and any labels)

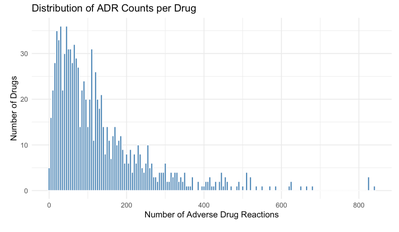
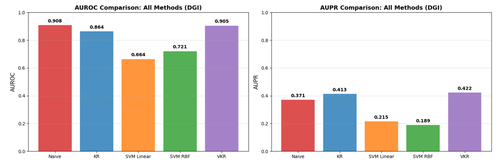


Figure 1: Early Performance of ADR Prediction Methods

Preliminary analysis on the DGI dataset shows that the Naïve baseline and VKR achieve the highest AUROC (≈0.91), while KR and VKR achieve the best AUPR (≈0.41–0.42), clearly outperforming SVM variants on both metrics. VKR therefore provides the best overall trade-off between discrimination (AUROC) and rare ADR detection (AUPR), motivating its use as the main reference method in further experiments.



Next Project Steps

We plan to conduct further analysis using:

- Extended Data set [1]
- New methods using weighted NMF, SVD and weighted SVD [2].

We will use the `PYTHON` Programming for this.

GitHub

The code and datasets for this project can be viewed at our GitHub repository here: https://github.com/arshad4387/ADR-Prediction.git

References

1. Massey et al. 2005 doi: 15.36.413↩
2. https://doi.org/10.3390/rs13020268↩