Prediction of ADRs using NMF and Weighted NMFs

Arshad Ahamed Shajahan, Sai Shasank Dandu

[1] School of Mathematical and Statistical Sciences, University of Galway

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].

Aim: To predict adverse drug reaction (ADR) profiles by integrating chemical fingerprints and drug–gene interaction using advanced statistical modeling approaches.

1. Explain the ADR profile prediction problem and its statistical challenges in imbalanced, noisy health data.
2. Explore various statistical methods for ADR prediction.

Datasets

1. **Drug-gene interaction pair**: Intersection of drugs from **DGIdb 4.0** and SIDER 4.1, generated to the binary matrix form. (Dim:778X2022)
2. **Chemical fingerprints**: Data from **PubChem** database, generated to the binary matrix form. (Dim:778X1024)
3. **Drug & side effects**: Drug along with it's side effects are extracted from SIDER 4.1 database. (1020 Drugs)

Descriptive Statistics

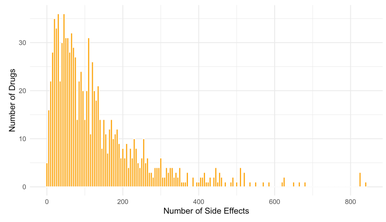


Figure 1: Distribution of side effects per drug

The above graph mentions the drugs majority of the drugs have more than 200 side effects, emphasising the importance of this study.

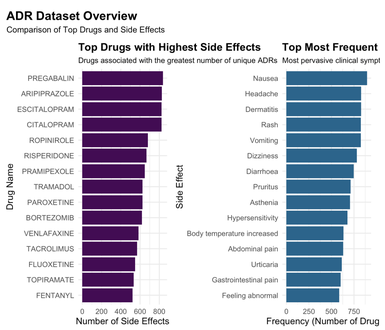
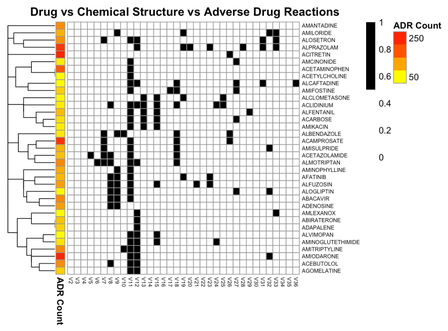


Figure 2: Top 15 drugs with most side effects and top 15 most frequent side effects

From the dataset, these are the drugs which possess the most side effects displayed alongside the most common side effects.



Methods

ADR Profile Prediction Methods Using Drug–Gene Interaction Features:

1.Naïve Frequency Model - Predicts ADRs based solely on their observed prevalence in the dataset, serving as a baseline.
2.Kernel Regression (KR) - Models the relationship between drug features and ADRs using a similarity-based kernel approach.
3.Linear SVM - Classifies ADR presence using a linear hyperplane in feature space.
4.RBF-Kernel SVM - Employs a non-linear radial basis function kernel to capture complex relationships between drug features and ADRs.
5.VKR (NMF + Kernel Ridge Regression) - Combines low-rank latent factor decomposition (NMF) with kernel ridge regression to predict ADRs in sparse and imbalanced datasets.

Early Results

Preliminary analysis on the figure.5 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.

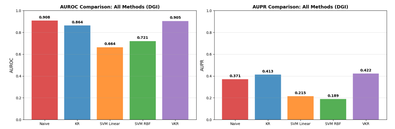


Figure 3: Early Performance of ADR Prediction Methods

Future Work

- Extended Data set with latest drug available from SIDER 4.1 & DGIdb database [2]
- New methods using weighted NMF, SVD and weighted SVD need to be explored. [3]

Acknowldgeme

We would like to thank Dr.Yezhao Zhong, Dr.Cathal Seoighe , Dr.Haixuan Yang for their work in ADR prediction and sharing the code and data through the github page.

References

[1] Zhong, Y., Seoighe, C., & Yang, H. (2024). Non-Negative matrix factorization combined with kernel regression for the prediction of adverse drug reaction profiles. Bioinformatics Advances, 4(1), vbae009.

[2] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, Peer Bork, The SIDER database of drugs and side effects, Nucleic Acids Research, Volume 44, Issue D1, 4 January 2016, Pages D1075–D1079,

[3] Lv X, Wang W, Liu H. Cluster-Wise Weighted NMF for Hyperspectral Images Unmixing with Imbalanced Data. Remote Sensing. 2021; 13(2):268.

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