Prediction of Adverse Drug Reactions Using Advanced Statistical Methods

Arshad Ahamed Shajahan, Sai Shasank Dandu
Supervised by Dr. Davood Roshan
School of Mathematical and Statistical Sciences, University of Galway

Background

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 and Objectives

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

The primary objectives are to,
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 with drugs on rows and genes on columns. (Dim:778X2022)
2. **Chemical fingerprints**: Data from **PubChem** database, generated to the binary matrix form with drugs on rows and chemical fingerprints on columns. (Dim:778X1024)
3. **Drug & side effects**: Drug along with it's side effects are extracted from **SIDER 4.1** database. (Dim:1020x5)

Descriptive Statistics

Figure 1 displays the distribution of side effects per drugs. This histogram illustrates a right-skewed distribution, where the majority of drugs have a relatively low number of side effects, typically clustered between 0 and 200. As the number of side effects increases, the frequency of drugs drops significantly, with few extreme outliers reaching over 800 side effects.

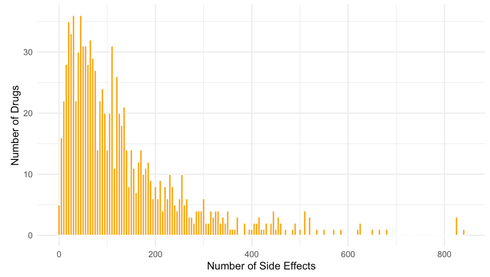

Figure 1: Distribution of side effects per drug

Figure 2 highlights a statistical concentration where a small group of medications exhibits over 800 unique adverse drug reactions (ADRs). Simultaneously, the data identifies pervasive clinical symptoms, such as nausea and headache, which are shared by nearly 900 different drugs.

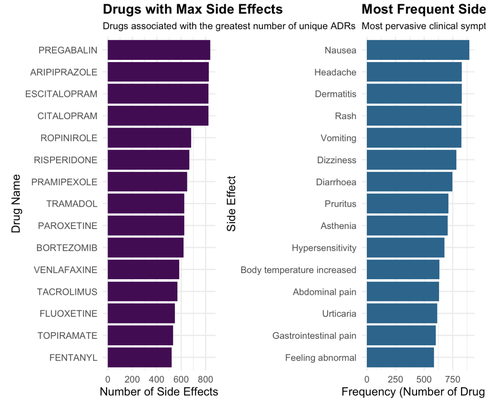

Figure 2: Drugs with most side effects and most frequent side effects

Figure 3 presents a heatmap illustrating statistically significant clustering where specific drug pairs share more than 80% of their side-effect profiles, indicating high pharmacological redundancy.

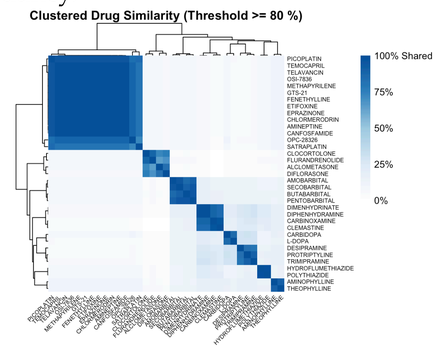

Figure 3: Drug and side effects similarity

Methods

ADR Profile Prediction Methods Using Drug–Gene Interaction Chemical fingerprint 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 4** 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.

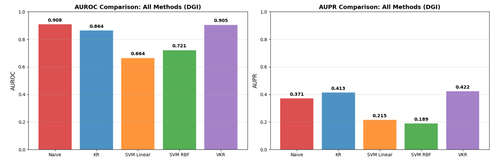

Figure 4: Early Performance of ADR Prediction Methods

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]

Acknowldgement

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] Yezhao Zhong, Cathal Seoighe, Haixuan Yang, Non-Negative matrix factorization combined with kernel regression for the prediction of adverse drug reaction profiles, Bioinformatics Advances, Volume 4, Issue 1, 2024, vbae009.
[2] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016 Jan 4;44(D1):D1075-9.
[3] Y. -D. Kim and S. Choi, "Weighted nonnegative matrix factorization," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 2009, pp. 1541-1544.

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