COMPARING THE PERFORMANCE OF MATRIX COMPLETION METHODS FOR PREDICTING THE EFFECTS OF DRUG COMBINATIONS

● LE VAN TUAN - LE DUC HAU

ABSTRACT:

This study applies some matrix completion methods for predicting the effects of drug combinations. The prediction of drug combinations will allow pharmacologists to find new therapies for treating many complex diseases like cancer. The study's results show that the Nuclear Norm Optimization method with low-rank assumption outperforms other methods, but the running time is quite expensive.

Keywords: performance, matrix completion, drug combination, cancer.

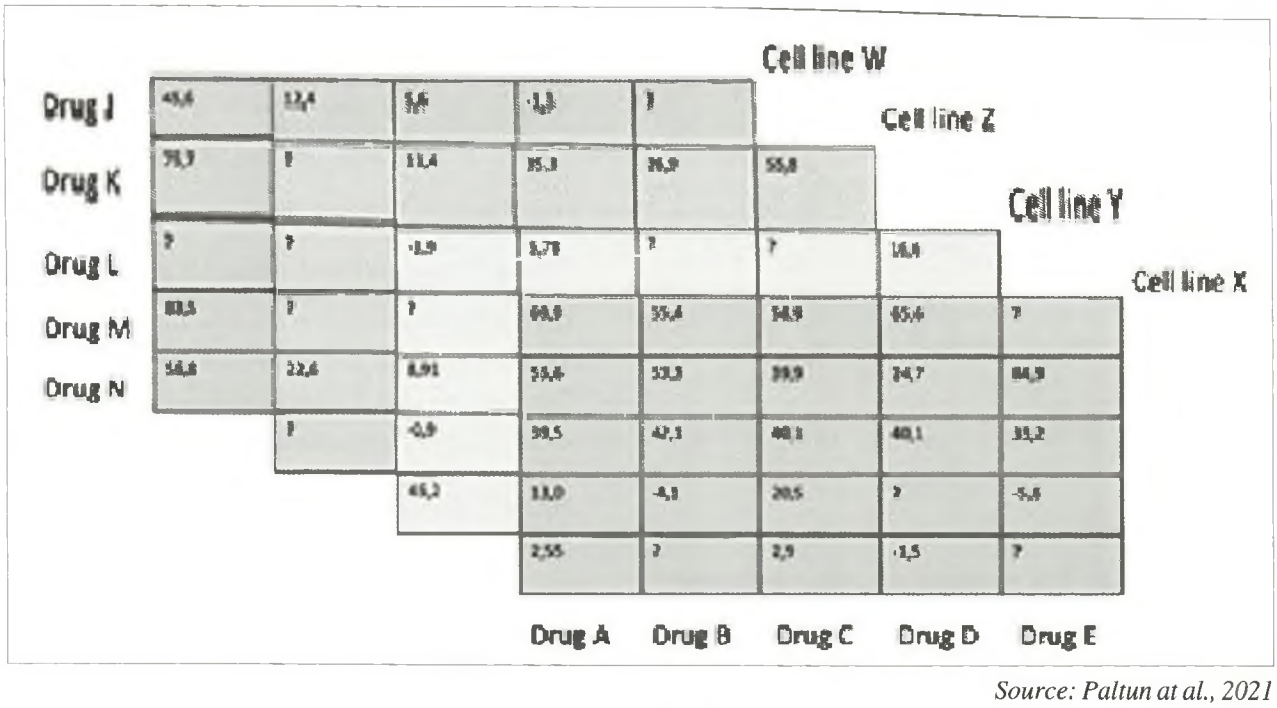
1. Introduction

In the pharmacology area, drug combination therapy is a promising strategy to treat complex diseases such as cancer. The role of this therapy is that it can increase therapeutic efficacy, reduce toxicity and overcome drug resistance. However, there is limited information about effective drug combinations since screening all possible drug combinations is challenging and expensive. Thus, the computational method for predicting the effects (efficacy and synergy) of a new drug combination is important to provide more sustainable treatment for patients.

In recent years, with the appearance of many database sources for drug combinations, such as DCDB and FDA (for multiple diseases) and NCI-ALMANAC, DREAM-AZ and ONEIL (for cancer), there have been a lot of studies about the methods in this topic. These methods include machine learning/deep learning approach and network-based approach; and usually incorporate many kinds of information such as transcriptomic or proteomic data, compound chemical structures, drug targets (Paltun at al., 2021; Weikaixin et al., 2022).

One simple approach for predicting the effects of drug combinations is the methods in the Matrix Completion problem. These methods only use the effects of drug combinations. Mathematically, the purpose of these methods is to fill in the missing

Figure 1: Drug combination effects prediction



Source: Paltun at al., 2021

values of a symmetric matrix containing the effects of drug combinations. Each row and each column of the matrix corresponds to a drug, and the matrix elements are the effects of the drug combinations (see Figure 1). In (Nafshi at al., 2021), they use Probabilistic Matrix Factorization method (in Matrix Completion topic) for predicting the effects of drug combinations. They find that Probabilistic Matrix Factorization is able to predict drug combination efficacy with high accuracy from a limited set of combinations and is robust to changes in the individual training data. In our study, we used many other methods in the Matrix Completion topic for predicting drug combination effects. We also compare the performance of these methods, including the accuracy and running time.

2. Methods

In this section we review some methods for the Matrix Completion problem. These methods were implemented in R-package filling (Kisung, 2021). See (Davenport, 2016) for a comprehensive overview of methods in the Matrix Completion topic.

2.1. Generalized Spectral Regularization

It is known that the LASSO type shrinkage estimator overestimates the number of non-zero coefficients if the assumed underlying model has sufficiently many zeros. The goal of this method is to overcome such difficulty via low-rank assumption and hard thresholding idea, well-known concept in conventional regression analysis.

2.2. Weighted K-nearest Neighbors

One of the simplest ideas to predict missing entry is to use a part of the data that has most similar characteristics across all covariates. The process for imputation in Weighted K-nearest Neighbors method follows such reasoning in that it finds K-nearest neighbors based on observed variables and uses weighted average of nearest elements to fill in the missing entry. Note that when there are many missing values, it's possible that there are no surrogates to be calculated upon. Therefore, if there exists an entire row or column with full missing entries, the algorithm will finish.

2.3. Nuclear Norm Optimization

In many situations, it is appropriate to assume

that there exists an underlying low-rank structure. The assumption of low-rank property makes matrix completion problem become an optimization problem:

Minimize rank (X) s.t $X_{ij} = A_{ij}$ for $A_{ij} \in E$

where $A_{ij} \in E$ means the (i, j)-th entry of data matrix A is not missing. The objective function can be further relaxed by nuclear norm:

$$\| X \|_* = \sum \sigma_i(X)$$

where $\sigma_i(X)$ is i-th singular value of the matrix X.

2.4. OptSpace

Let M be an m × n matrix of rank r << n, and assume that observed E is a uniformly random subset E of its entries. OptSpace is an efficient algorithm that recovers M from |E| = O(rn) observed entries with relative root mean square error

$$RMSE \leq C(\alpha)\sqrt{nr/|E|}$$

2.5. Simple Rules

One of the simplest ways to impute the missing entries is to apply any simple rule for each variable. The options include "mean", "median", and "random". The assumption is that every column has at least one non-missing entries. For each column, the rule is applied from the subset of non-missing values.

2.6. Spectral Regularization

The method carries out convex relaxation techniques to create a sequence of regularized low-rank solutions for matrix completion problems. For the nuclear norm optimization problem, it uses soft thresholding technique iteratively. It leads that the algorithm returns several matrices in accordance with the provided vector of regularization parameters λ.

2.7. Singular Value Thresholding

The method is a repeatedly updating scheme for Nuclear Norm Minimization problem. The objective function is

$$minimize \frac{1}{2}\|P_\Omega(X - A)\|^2_F + \lambda\| X \|$$

where $P_\Omega(X) = X_{ij}$ if it is observed, or 0 otherwise. It performs repeatedly shrinkage on newly estimated singular values.

2.8. Universal Singular Value Thresholding

The method is suitable for low-rank structure. The goal of this method is that it exploits the idea of thresholding the singular values to minimize the mean-squared error, defined as

$$MSE(\hat{A}) = E\left\{\frac{1}{np}\sum_{i=1\text{-}n;\, j=1\text{-}p}(\hat{a}_{ij} - a_{ij})^2\right\}$$

where A is an (n × p) matrix with some missing entries and is an estimate.

3. Results

In this study, we use NCI ALMANAC database - one of the pioneers in the characterization of drugs in vitro. This database is a collection of pairwise combinations of 104 FDA approved anticancer drugs against the NCI-60, a set of 60 common human tumor cancer cell lines collected by the National Cancer Institute. A total of 5,232 drug-drug pairs were evaluated in each of the cell lines; 304,549 experiments were performed to test each drug at either 9 or 15 combination dose points, for a total of 2,809,671 dose combinations. At each dose combination, the percent cell growth after 2 days was measured and recorded, and the efficacy of the combination calculated as the percent of growth inhibition. For each cell line, the combination efficacies are arranged into a symmetric matrix, M104x104, where each row and column represent a drug, and each element represents the efficacy of a unique drug-drug combination on that cell line. The synergy matrix is generated similarly. Note that, in the NCI ALMANAC database, the synergy of each combination is reported as a "ComboScore" that measures the difference between the recorded growth rate after testing and the growth rate expected by Bliss Independence. A positive

Table 1. Summary of effects of drug combinations of 786-0 celline

	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
Efficacy	-88.44	29.86	57.80	54.61	85.41	150.20
Synergy	-59.333	-3.778	-1.111	-1.523	1.000	28.556

Source: NCI ALMANAC database

ComboScore indicates a synergistic combination, whereas a negative ComboScore indicates an antagonistic combination. To illustrate the data, see Table 1 for a summary of effects of drug combinations of 786-0 celline.

The goal of this study is to compare the ability of Matrix Completion methods in section 2 to recover hidden elements in the drug combination efficacy/ synergy matrix. For each cell line, we randomly hid 5% of the combination efficacy/ synergy matrix, creating non-overlapping "training" and "validation" sets. For each method, we predict the hidden values and complete the matrix. The Root Mean Square Error (RMSE) is calculated between the actual values and the predicted values. Then, the average RMSE is estimated from RMSE values of 60 cellines. The results are shown in the second and fourth columns of Table 2.

The third and fifth columns in Table 2 shows the run time of the methods. The computer for running is the HP Laptop, Intel(R) Core(TM) i5-8265U: CPU @ 1.60GHz, 1.80GHz, RAM 8GB.

We can see that the with efficacy matrices, Nuclear Norm Optimization method outperforms other methods. The RMSE avg. of this method is 19.188. However, with synergy matrices, the accuracy of this method is similar to Weighted K-nearest Neighbors and Singular Value Thresholding method. In two cases, the OptSpace method performs the poorest accuracy campared to other methods. Besides, the time run of Nuclear Norm Optimization method is significantly long. It costs about 1.5 hours and 1 hour for efficacy matrices and synergy matrices, respectively; while other methods take a few dozen seconds. Note that, the method requires low-rank assumption and this assumption is also suitable

Table 2. The performance of methods

Method	RMSE avg. of efficacy	Run time (secs) of efficacy	RMSE avg. of synergy	Run time (secs) of synergy
Generalized Spectral Regularization	62.279	24	6.405	31
Weighted K-nearest Neighbors (k = 25)	25.182	36	5.875	52
Nuclear Norm Optimization	**19.188**	**1.37 (hours)**	**5.85**	**1 (hour)**
OptSpace	119.666	28	25.18	24
Simple Rules	48.351	12	8.88	30
Spectral Regularization	62.765	14	7.53	31
Singular Value Thresholding	30.537	47	5.83	113
Universal Singular Value Thresholding	51.709	12	18.474	28

Source: NCI ALMANAC database

with the drug combination matrix (Yang, 2019). The disadvantage of this method is that computational efficiency may not be guaranteed for large data matrix.

To illustrate, we apply Nuclear Norm Optimization method for predicting the effects of one pair of drugs (drug id 752 and 3088) with 786-0 celline. The predicted values are approximate of the actual values (see Table 3)

Table 3. A prediction of nuclear norm optimization method

	Actual value	Predicted value
Efficacy	17.41	18.67
Synergy	-2.67	-2.23

Source: NCI ALMANAC database

4. Conclusion

Our results show that it is possible to use only information on the effects of drug combinations to predict the effects of novel combinations. A strength of the Matrix Completion approach is that it does not require any outside knowledge of chemical structures, target profiles, or OMICS data. Moreover, not relying on additional information endows the approach with flexibility: Instead of predicting the effects of combinations of drugs, it can be used to predict the effects of combinations of combinations. Many Matrix Completion methods require low-rank assumption and this assumption is also suitable with the drug combination matrix. In all Matrix Completion methods introduced in Section 2, the Nuclear Norm Optimization method gets the best performance with the lowest RMSE; however, this method also requires the longest consuming time ■

REFERENCES:

1. Betül Güvenç Paltun, Samuel Kaski, Hiroshi Mamitsuka. (2021). Machine learning approaches for drug combination therapies. *Briefings in Bioinformatics,* 22(6). https://doi.org/10.1093/bib/bbab293

2. M. A. Davenport and J. Romberg. (2016). An Overview of Low-Rank Matrix Recovery From Incomplete Observations. *IEEE Journal of Selected Topics in Signal Processing,* 10(4), 608-622. doi: 10.1109/JSTSP. 2016.2539100.

3. Kisung You. (2021). Filling: Matrix Completion, Imputation, and Inpainting Methods. Retrieved from: *https://cran.r-project.org/web/packages/filling/filling.pdf.*

4. Nafshi R and Lezon TR. (2021). Predicting the Effects of Drug Combinations Using Probabilistic Matrix Factorization. *Frontiers in Bioinformatics,* 1, 708815. https://doi: 10.3389/fbinf.2021.708815.

5. Yang M, Luo H, Li Y, Wu F-X, Wang J. (2019). Overlap matrix completion for predicting drug-associated indications. *PLoS Computational Biology,* 15(12), e1007541. https://doi.org/10.1371/journal.pcbi.1007541.

6. Weikaixin Kong et al. (2022). Systematic review of computational methods for drug combination prediction. *Computational and Structural Biotechnology Journal,* 20. https://doi.org/10.1016/j.csbj.2022.05.055.

Received date: June 3, 2022
Reviewed date: June 15, 2022
Accepted date: June 29, 2022

Author's information:

1. LE VAN TUAN[1]

2. LE DUC HAU[2]

[1]Thuongmai University

[2]Thuy Loi University

SO SÁNH HIỆU NĂNG CỦA MỘT SỐ PHƯƠNG PHÁP HOÀN THIỆN MA TRẬN TRONG DỰ ĐOÁN ĐỘ HIỆU QUẢ CỦA KẾT HỢP THUỐC

● **LÊ VĂN TUẤN[1]**

● **LÊ ĐỨC HẬU[2]**

[1]Trường Đại học Thương mại

[2]Trường Đại học Thủy Lợi

TÓM TẮT:

Trong nghiên cứu này, một số phương pháp hoàn thiện ma trận được áp dụng để dự đoán độ hiệu quả của việc kết hợp thuốc. Kết quả dự đoán sẽ hỗ trợ các nhà dược học trong việc tìm ra các liệu pháp mới trong điều trị nhiều bệnh phức tạp như là ung thư. Kết quả nghiên cứu cho thấy phương pháp Nuclear Norm Optimization với giả thiết low-rank có độ chính xác tốt nhất trong các phương pháp được nghiên cứu. Tuy nhiên, phương pháp này cũng có thời gian chạy lâu nhất.

Từ khóa: hiệu năng, hoàn thiện ma trận, kết hợp thuốc, ung thư.