Hong Kong Baptist University
Faculty of Science
Department of Computer Science

GCAP 3055 GE Capstone Interdisciplinary Independent Study (COMP)
Spring 2023

A benchmark study of comparing representation learning-based methods with machine learning-based methods and data mining method on drug synergy prediction

Ruiqi ZHUANG

A benchmark study of comparing representation learning-based methods with machine learning-based methods and data mining method on drug synergy prediction

Ruiqi ZHUANG

May 11, 2023

Abstract

Drug synergy and sensitivity prediction is an important field of research in the treatment of human diseases. Recently computational-based approaches have been proposed by researchers to address this issue. However, researchers mostly compare the performance of their models solely with representation learning-based methods. In this work, we compared several machine learning-based methods including XGBoost, linear regression, SVM, logistic regression, and KNN, a classic data mining method with representation learning-based methods. Across all assessment measures, including BCE, AUROC, AUPRC, and AUC, our findings demonstrated that representation learning-based approaches outperformed machine learning-based methods. These findings imply that representation learning algorithms have the potential to enhance drug synergy and sensitivity prediction accuracy. However, other factors such as data quality and the algorithms utilized may have an impact on the outcomes. Finally, we implemented CP decomposition method and combine it with other methods and find it does make improvements on the original methods.

1 Introduction

Monotherapy is a type of therapy in which patients are assigned to a specific medicine or treatment that targets a certain condition. This kind of therapy is relatively effective in dealing with the treatment of human illness. However, the performance of this conventional technique may be restricted by the complexity of human diseases (Sun et al., 2020), patient heterogeneity (patients of different ages and gender may have varied sensitivities to the same medicine), and drug resistance. Among all of these restrictions, the main limitation of monotherapy is drug resistance, which can be caused by a variety of variables associated to the cancer cell line's system. These variables may result in drugs resistance, therefore they may impair the effectiveness of monotherapy (Torkamannia et al., 2022).

Thus, Combinatorial drug therapy is employed to address this issue. Using the drug synergistic effect, this therapy investigates more than one medicine for a particular illness simultaneously and showed a significant enhancement in curative efficacy (Sun et al., 2020).

The identification of drug synergistic effects is at the core of combinatorial drug therapy. However, detecting drug synergistic effects involves considerable studies that are time-consuming and

might be expensive (Kim et al., 2021). As a result, recent research has attempted to focus on applying computational approaches for making predictions on drug synergistic effects (Kim et al., 2021). As input, the computational methods utilize a tensor including information such as drug and cell line information. The input tensor will then be processed through a series of computations, such as feature extraction and fully-connected layer. Finally, the model estimates a drug synergy score or a binary prediction of drug synergy as the final output.

Researchers are investigating different computational models such as random forest, neural networks, XGBoost, and deep learning-based approaches (Torkamannia et al., 2022). Furthermore, several researchers are concentrating on the extraction of biological features from input data when designing model structure.

2 Literature Review

In recent years, scholars have tried various methods to predict the synergy effect. One of the more common methods is representation learning, and an important section in representation learning is feature extraction.

Previous approaches may have relied more on prior knowledge. To pick characteristics, Preuer et al. (2018) and Chen et al. (2016) use human knowledge-based approaches. Preuer et al. (2018) manually select features and utilize them to construct predictions for the synergy effect. Chen et al., (2016) used generated features based on human assumptions.

Sun et al., (2020) and Guan et al., (2019) used matrix/tensor factorization to extract features. Canonical Polyadic Decomposition (CPD) is the traditional method of tensor factorization, which takes a tensor with drug and cell line information as input. Similarly, Sun et al., (2020) take a three-dimension tensor consisting of drug A, drug B, and cell line as input. However, due to the fact that CPD approach fails to handle missing data, Sun et al., (2020) proposed a novel method called CP-WOPT to address the missing values issue. CP-WOPT generates a non-negative weighting tensor with the missing value set to 0 and the known value set to 1, then conducts elementwise multiplication with the initial objective error function. Then, CP-WOPT obtains a weighted objective error function for the tensors with missing values. The LBFGS-B method was then used by CP-WOPT to optimize. Finally, three rank-one tensors A, B, and C are then produced as the decomposition result of the original tensors, each of the three rank-one tensors holds the latent information of drug A, drug B, and cell line in the form of row vectors as features. Guan et al., (2019) proposed a model that uses the drug response dataset as the initial input, which includes drug chemical structure similarity, cell line expression similarity, and drug response in cell line. Sparsification technique is then applied to sparsify drug and cell line similarity by using matrices to represent the p-nearest neighbor graphs. Low-rank approximation (LRA) with graph regularization terms is therefore used with the help of sparsified similarity matrices to partition the drug response matrix into low rank matrices as features. Finally, by making the neighbor drug nearest in the latent space, this method is able to eliminate noisy elements that impair the accuracy of synergy prediction.

Kumar Shukla et al., (2020), Kuru et al., (2021), Kim et al., (2021), and Li et al., (2018) employ a network structure to extract the feature. By running them through a 1D convolutional neural network (CNN), Kumar Shukla et al., (2020) managed to perform features extraction for high-dimension input matrices that encode the information of drug A, drug B, and drug synergy. Kuru et al., (2021) proposed a drug specific network (DSN) and the features are the representation of gene expression learned by aggregrating two DSNs. Li et al. (2018) filter molecular features such as gene expression, copy-number variation (CNV), methylation, and mutation based on target information, and then simulate features based on a previously known gene-gene interaction network.

Kim et al., (2021) proposed a big model including a drug encoder, a cell line encoder, and a merging layer. The drug encoder and cell line encoder are used to learn the drug and cell line em-

bedding representations. The feed-forward layer or transformer encoder will process the original drug input characteristics such as drug ID, MACCS fingerprints, canonical SMILES, and target genes in the drug encoder. The result was then concatenated to feed-forward layers to produce a representation of each drug. The original input features of cell lines, such as cell line ID, tissue, cancer kind, and gene expressions, will go through feed-forward layers respectively and concatenated with the result of different feed-forward layers in the cell line encoder to produce the representation of each cell line. The model then combines the output of the drug encoder with the cell line encoder to provide predictions on synergy and sensitivity using feed-forward layers. Furthermore, to mitigate the impact of data scarcity in data-poor tissues Kim et al., (2021) employed transfer learning to improve the performance of data-poor tissues by transferring information from data-rich tissues.

TreeSHAP, a feature attribution method described by Janizek et al. (2018), is another alternate approach. The technique uses the physical and chemical properties of drugs, as well as cell line expression as input features. Then, for each input feature, SHAP values are calculated, and the most essential features, commonly the top 1,000 or 2,000 features, are selected for retraining models.

3 Limitation

Previous research has been focused on representation learning and has not included a benchmark study that uses machine learning-based approaches. Usually, they compared their representation learning methods to other methods that used representation learning.

Researchers may miss out on possibilities to find the most successful ways for a certain task or application if they focus solely on representation learning and exclude machine learning methods from a benchmark study. Furthermore, researchers may apply representation learning techniques by default without properly examining their relevance for a certain task or application. A benchmark study using machine learning methods might be able to help researchers to investigate a broader range of methodologies.

A benchmark study that uses machine learning methods may be beneficial since it is able to set the baseline of performance and allow researchers to compare the performance of other models or algorithms to machine learning-based methods. This would allow researchers to obtain a better understanding of the relative strengths and shortcomings of different machine learning methods. Additionally, this would help researchers identify areas for making improvement.

4 Method

The model proposed by Kim et al. (2021) is purely based on representation learning, furthermore, the model adopted Transformer, the state-of-the-art model, to encode the chemical information of drugs. Therefore, the high performance makes it a satisfying baseline to be compared with machine learning-based methods.

We first did the hyperparameter tuning to the model of Kim et al. (2021) in order to select the best hyperparameter set for our dataset. After obtaining a relatively satisfying hyperparameter set, we compare the performance with machine learning-based methods.

Previous researchers have investigated XGBoost (Torkamannia et al., 2022), so we select XGBoost and other traditional machine learning algorithms to make comparisons. For the classification task, we implemented logistic regression, XGBoost regression, and support vector machine (SVM). We implemented linear regression and XGBoost regression for predicting the synergy score and sensitivity score. Additionally, we implement, KNN, a classic data mining method to complete regression task. Finally, we implement Canonical polyadic decomposition (CP decomposition), a classic method of tensor decomposition, and try to combine CP decomposition with other methods.

For the regression task, since the work of Kim et al. (2021) used BCELoss for comparison, we compare the BCELoss of synergy predictions and sensitivity predictions as well. BCELoss is a commonly used loss function in representation learning, which measures the error of reconstruction. Our predictions will be mapped into the interval between 0 and 1 via a sigmoid function, then the mapped predictions will be fed into BCELoss to calculate the loss. For the classification task, we compare the Area Under the Receiver Operating Curve (AUROC), the Area Under Curve (AUC), and the Area Under the Precision-Recall Curve (AUPRC) for measuring the ability to identify the positive samples.

5 Dataset and Evaluation Matrices

5.1 Dataset

The dataset of this study integrates the following databases DrugComb database, TTD (Therapeutic Targets Database), NIH-LINC (Library of Integrated Network-Based Cellular Signatures), CCLE (Cancer Cell Line Encyclopedia), COSMIC (Catalogue of Somatic Mutations in Cancer).

DrugComb gathers, analyzes, and disseminates drugs combination screen results on cancer cell lines. This database contains 8397 drugs, 2320 cell lines, 33 tissues, and 739,964 pharmacological combinations. Experiments are conducted on the skin, lungs, ovary, and other organs. TTD contains information on drug targets, such as biological functions, illnesses linked with them, and medications that target them. NIH-LINC is a database of gene expression data from various cellular perturbations, such as pharmacological treatments. Based on the effects on gene expression, it can be used to identify potentially effective drug combinations. CCLE contains genetic and pharmacological information on a considerable number of cancer cell lines. It may be utilized for identifying potential drug targets and screening the effectiveness of drugs. COSMIC is a database of cancer somatic mutations, including those that may affect drug effectiveness. It may be used to identify potential drug targets and to test medications for effectiveness in specific cancer types.

5.2 Data cleaning and processing

So far we have extracted the data of Drug ID, MACCS fingerprints, SMILES, Cell line ID, and cancer type from TTD, NIH-LINC, CCLE, and COSMIC. When extracting data from DrugComb database, we fail to obtain one important missing document, so we managed to utilize the data from existing data extracted from DrugComb website to generate the missing document.

To clean the data we extract, we first count the number of NA values in the dataset and the result shows that the rows containing NA values are relatively small. Therefore we drop the rows containing NA values. Then we find that some strings are mixed in a column that should only contain the number, so we converted these entries into numbers. Finally, when processing the cell line data, we discovered that the disease name column is missing, so we used disease ID to compensate for the cancer type.

5.3 Evaluation

This model used the Loewe synergy score and RI for synergy prediction and sensitivity prediction. Loewe synergy score is to measure the excess over the expected response if the two drugs are the same compound. RI is relative inhibition which measures the ability of a drug in inhibiting the growth of cells.

This study involved regression tasks and classification tasks. Since the objective of drug screening is to predict whether the combination will be synergistic, so we used BCE loss to evaluate the

performance regression models. For classification, the model used cross entropy as the loss function. To evaluate the performance of the classification task, the model applied the Area Under the Receiver Operating Curve (AUROC), the Area Under Curve (AUC), and the Area Under the Precision-Recall Curve (AUPRC) in order to evaluate the ability to identify the positive samples.

6 Results and Discussion

6.1 Hyperparameter tuning

The result of the hyperparameter tuning is shown as followed. We found the hyperparameter set of learning rate equals 0.05 and weight decay equal 0 has a relatively satisfying performance of {0.2970 Synergy BCE, 0.1014 Sensitivity BCE, 0.8840 Synergy AUROC, 0.6676 Synergy AUPRC, 0.7698 Synergy AUC}.

Hyperparameter tuning result					
Hyperparameter set	Syn_BCE	Sen_BCE	Syn_AUROC	Sen_AUPRC	Sen_AUC
lr=1e-2 wd=0	0.2998	0.1005	0.8766	0.6446	0.7771
lr=5e-3 wd=0	0.3079	0.1010	0.8665	0.6237	0.7564
lr=1e-3 wd=0	0.3509	0.1010	0.8945	0.5069	0.7358
lr=5e-2 wd=0	0.2970	0.1014	0.8840	0.6676	0.7698
lr=5e-2 wd=5e-5	0.3125	0.1016	0.8673	0.6214	0.7678
lr=1e-1 wd=5e-4	0.4253	0.1035	0.8313	0.5303	0.7460
lr=1e-1 wd=5e-5	0.3612	0.1061	0.8170	0.5338	0.7697

Table 1: lr: learning rate, wd: weight decay. Among all these hyperparameter sets, learning rate equals 0.05 and weight decay equals 0 has the best performance.

6.2 Classification Model Comparsion

The machine learning methods of XGBoost, SVM, and logistic regression models appear to have lower performance than the baseline model regarding AUROC, AUPRC, and AUC. Especially, in the measurement of AUPRC machine learning methods show a significantly lower performance. This may be due to the fact that AUPRC is for imbalanced data, while the data for this dataset is balanced, and some distributions may not be suitable for AUPRC, so it can not evaluate the performance very well.

6.3 Regression Model Comparison

Compared to machine learning-based methods, the representation learning-based baseline shows higher performance. XGBoost and linear regression achieve similar synergy BCE while linear regression achieves lower sensitivity BCE, and the data mining approach (i.e. KNN) has a higher synergy BCE and sensitivity BCE than linear regression. Thus, Linear regression has the best performance.

Model Comparison			
Model	Syn_AUROC	Sen_AUPRC	Sen_AUC
Kim et al., (2021)	0.8840	0.6676	0.7698
XGBoost classification	0.5106	0.0162	0.7221
Support Vector Machine	0.5212	0.0091	0.5202
Logistic regression	0.4999	0.0086	0.5239

Table 2: Performance comparison of baseline methods and machine learning-based methods on the classification task. XGBoost classifier achieves a relatively good performance regarding synergy AUROC, sensitivity AUPRC, and sensitivity AUC.

6.4 The impact of combining tensor decomposition method

Solely using CP Decomposition achieves a similar performance as the machine learning-based methods in synergy prediction and has a better performance in sensitivity prediction. Combining CP Decomposition with XGBoost regressor and KNN mode is able to enhance the performance of XGBoost regressor and KNN.

Model Compassion		
Model	Syn_BCE	Sen_BCE
Kim et al., (2021)	0.2970	0.1014
XGBoost regressor	0.6956	1.3512
Linear regression	0.6980	0.5723
KNN	0.7203	0.5859
CP Decomposition	0.7159	0.4670
Linear regression+CP Decomposition	0.8563	0.6011
XGBoost regressor+CP Decomposition	0.6551	0.5983
KNN+CP Decomposition	0.6913	0.5779

Table 3: Performance comparison of baseline methods and machine learning-based methods on the regression task. Linear regression achieves a relatively good performance regarding synergy BCE and sensitivity BCE among all machine learning methods. Methods that combine CP Decomposition KNN and XGBoost have a lower BCE loss compared to the original method.

7 Conclusion

According to our benchmark study, the representation learning-based method appears to perform better than the machine learning-based methods and data mining-based method on both synergy prediction and sensitivity prediction, as evaluated by the BCE AUROC AUPRC and AUC metrics. This shows that in the future, representation learning might be a beneficial method for enhancing the performance of these kinds of predictions. However, other factors such as data quality and training and prediction methods may also play a role in these outcomes. Additionally, tensor decomposition has the potential to enhance the performance of machine learning-based methods and data mining-based methods.

References

[1] Berenbaum, M. C. (1989). What is synergy?. *Pharmacological reviews*, 41(2), 93-141.

[2] Bliss, C. I. (1939). The toxicity of poisons applied jointly 1. *Annals of applied biology*, 26(3), 585-615.

[3] Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., & Yan, G. (2016). NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS computational biology*, 12(7), e1004975.

[4] Guan, N. N., Zhao, Y., Wang, C. C., Li, J. Q., Chen, X., & Piao, X. (2019). Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization. *Molecular therapy-nucleic acids*, 17, 164-174.

[5] Janizek, J. D., Celik, S., & Lee, S. I. (2018). Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine. *BioRxiv*, 331769.

[6] Kim, Y., Zheng, S., Tang, J., Jim Zheng, W., Li, Z., & Jiang, X. (2021). Anticancer drug synergy prediction in understudied tissues using transfer learning. *Journal of the American Medical Informatics Association*, 28(1), 42-51.

[7] Kumar Shukla, P., Kumar Shukla, P., Sharma, P., Rawat, P., Samar, J., Moriwal, R., & Kaur, M. (2020). Efficient prediction of drug–drug interaction using deep learning models. *IET Systems Biology*, 14(4), 211-216.

[8] Kuru, H. I., Tastan, O., & Cicek, A. E. (2021). MatchMaker: a deep learning framework for drug synergy prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(4), 2334-2344.

[9] Li, H., Li, T., Quang, D., & Guan, Y. (2018). Network propagation predicts drug synergy in cancerspredict drug synergy with network propagation. *Cancer research*, 78(18), 5446-5457.

[10] Loewe, S. (1953). The problem of synergism and antagonism of combined drugs. *Arzneimittelforschung*, 3, 285-290.

[11] Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C., & Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*, 34(9), 1538-1546.

[12] Sun, Z., Huang, S., Jiang, P., & Hu, P. (2020). DTF: deep tensor factorization for predicting anticancer drug synergy. *Bioinformatics*, 36(16), 4483-4489.

[13] Torkamannia, A., Omidi, Y., & Ferdousi, R. (2022). A review of machine learning approaches for drug synergy prediction in cancer. *Briefings in Bioinformatics*, 23(3).

[14] Yadav, B., Wennerberg, K., Aittokallio, T., & Tang, J. (2015). Searching for drug synergy in complex dose–response landscapes using an interaction potency model. *Computational and structural biotechnology journal*, 13, 504-513.