Your Paper

You

April 29, 2024

1 Introduction

Drug response data supports making therapy decisions and clinical use of drugs. These data explores change in important biomarkers after treatment with drug, and to draw further conclusion about whether the drug being responsive or not. Traditional acquisition of these data are expensive, due to the cost incurred during drug response tracking, data manifesting and management, needs for professional operators and equipment involved during the process. If this approach were to be revolutionized by artificial intelligence, to predict drug response before medical experiment take place in reality, this could significantly reduce the temporal and financial cost to collect these data, and therefore benefit therapeutics decisions and development of new drug.

The detailed of method used for **D**rug **R**esponse **P**rediction(DRP) varies, and could be loosely classified into two categories: machine learning approaches and deep neuron network approaches. Machine learning approaches usually features quick training ,low requirement for amount of data and highly interpret-able result. While deep neuron network approaches are usually good at excavating deeply hidden pattern in data, if given abundant amount of data.

One of the machine learning approach is Multi-output Gaussian process(MOGP), an extended model from Gaussian Process(GP). In GPs, data are considered to be sampled from a latent distribution, which is assumed to be Gaussian distribution. Prediction using GP involves finding the best likelihood of distribution parameters with given data, and conditioning it with the knowns for prediction. MOGP assumes the latent distribution to be a linear combination of either identical or different GPs. With this assumption, MOGP is expected to discover more complex patterns and make more precise predictions.

Existing research that applied the model to DRP field indicates traditional MOGP setup don't propagate its knowledge well from one drug to another.[2] This research will combine effort of previous researches, to improve the performance of generalization in the field of DRP.

In the following parts of paper, section 2 will introduce the background knowledge about this project, elaborated with relevant literature. Section 3 will cover the preliminary methodology of the project. Section 4 will cover ethics considerations. Potential risks and backup plan will be introduced in section 5. Expected outcome and evaluation will be discussed in section 6. Finally, section 7 will describe some of the technical milestone so far expected with the time-frame of workload.

2 Background

In this section, some of the relevant background knowledge will be introduced, elaborated with some literature.

2.1 Dataset

Genomics of Drug Sensitivity in Cancer (GDSC) is a resource developed by the Wellcome Trust Sanger Institute and the Massachusetts General Hospital Cancer Center, aims to understand how genomic feature influence drug response on cancer cells.

The dataset provides extensive drug sensitivity data from approximately 75000 experiment, encompassing 138 anti-cancer compounds tested across nearly 700 cancer cell lines. These data are labeled with genomic information, including somatic mutations, gene amplifications, deletions, tissue types, and transcriptional profiles.

GSDC1 and GSDC2 are two different versions provided by GSDC. GSDC1 is collected between 2009 and 2015, and GSDC2 is screened from 2015 until now. There are also discrepancies in screening of these data, where different methods are used in the part of compoud storage, cancer cell seeding, and cell viability measuring.

This research plan to use GSDC1 and GSDC2 dataset, as it is used by many in drug response prediction research.[1, 2, 3]

2.2 GP

2.3 MOGP

2.4 Kernal function

2.5 Transferable kernal

3 Research Methodology

4 Ethics and professional considerations

This project does involve human-related data about drug response. However collection of GSDC1 and GSDC2 are done by Wellcome Trust Sanger Institute cooperating with Massachusetts General Hospital Cancer Center and therefore independent of this research. According to section 3.7 in POLICY ON THE ETHICAL INVOLVEMENT OF HUMAN, UNIVERSITY OF MANCHESTER, to use secondary data is considered exempted from ethical review as long as 3 criterion is met:

The data are completely anonymous. GSDC data is obtained by transfering experimented compound to seeded cancer cells in a plate, and then measuring cell viability. No

The researcher has explicit consent from the data controller to access the data. GSDC1 and GSDC2 are both open source dataset that is open to general public. As stated in the source of data, ”Users have a non-exclusive, non-transferable right to use data files for internal proprietary research and educational purposes, including target, biomarker and drug discovery.”

The researcher is able to prove that the data will be used for a purpose which falls within the remit of the original consent provided by data subjects. The GSDC1 and GSDC2 aim to support analysis of biomarkers to improve cancer treatment. This research aims to introduce transfer kernal to MOGP model and improve its performance across different drugs. Therefore, this criteria can be considered satisfied.

Should any more dataset is needed in the future, these criterion will be taken into consideration to make sure it goes accord with ethic considerations.

This project does not involve cooperation with company, therefore free from professional considerations.

5 Risk consideration

The risk of the research is considered minor, as both transferable kernal and MOGP model are very matured design, and no convincing evidence has been found to conclude the lack of viability of the research.

If the research is later proved to be not feasible, then the focus can be shifted to:

1. feature embedding of the datasets Current MOGP solution feeds raw data directly to the model. Research can investigate possible embedding methods for features and experiment on whether they can lead to performance improvement.

2. different likelihood function for regression. Likelihood function is involved when doing regression on kernal parameters. It is an assumed probability density function that the kernal parameters are expected to follow. This assumption is strong and therefore a different likelihood function may influence performance of the model.

6 Project evaluation

The project can be considered success, if the model with transferable kernal reaches similar performance on each of the domain of the previous model, and proved to generalize better accross different domains.

7 Planning

References

[1] Kexin Qiu, JoongHo Lee, HanByeol Kim, Seokhyun Yoon, and Keunsoo Kang. Machine learning based anti-cancer drug response prediction and search for predictor genes using cancer cell line gene expression. *Genomics & informatics*, 19(1), 2021.

[2] Dennis Wang, Juan-José Giraldo Gutierrez, Evelyn Lau, Subhashini Dharmapalan, Melody Parker, Yurui Chen, and Mauricio Alvarez. Multi-output prediction of dose-response curves enable drug repositioning and biomarker discovery. 2023.

[3] Xinping Xie, Fengting Wang, Guanfu Wang, Weiwei Zhu, Xiaodong Du, and Hongqiang Wang. Learning the cellular activity representation based on gene regulatory networks for prediction of tumor response to drugs. *Artificial Intelligence in Medicine*, page 102864, 2024.