Research on Machine Learning for Drug Discovery and Precison Medicine

ZHAO, Kai

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
School of Biomedical Sciences

Supervised by

Prof. So Hon-cheong

The Chinese University of Hong Kong
July 2020

Thesis Assessment Committee

Professor Cheng Sze Lok Alfred (Chair)

Professor So Hon-cheong (Thesis Supervisor)

Professor Chen Yangchao (Committee Member)

Professor Sham Pak Chung (External Examiner)

Abstract of thesis entitled:

Research on Machine Learning for Drug Discovery and Precison Medicine

Submitted by ZHAO, Kai

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in July 2020

There have been a rapid advance in artificial intelligence (AI) and rise in the availability of biomedical data. This offers a great opportunity to integrate them to benefit our healthcare. Additionally, drugs and therapies play crucial roles in healthcare. However, the number of drug approved by FDA is much slower than anticipated. Meanwhile, doctors still practice the one-drug-fits-all model in disease treatment.

In this thesis, we seek to ease the issues by application of AI on biomedical data, especially "omics" data. In this study, we propose a computational drug repurposing approach based machine learning methods using drug expression profiles, in which outcome variable is defined as whether drugs can treat the disease. Systematic approaches have been employed to validate our results. In practice, drug repurposing is not always available. Moreover, the traditional drug development suffers from a high failure rate, due to lack of efficacy of compounds under investigation. Thus, in order to relief this problem we propose a computational framework to identify promising drug targets for further study, independent of all kinds of known target information. We verified our results by examining whether our top candidates are more likely to enrich known drug targets.

Precision medicine has been advocated in the past years. The aim of precision medicine is to estimate individualized treatment effects (ITE). In order to examine the ITE of a risk factor or treatment, we employed forests based methods given subjects' clinical and genetic information, and proposed a "weighted mean" approach to incorporate time-to-event data. Since the true ITE cannot be directly observe, there are no well-established methods to validate the model fitting. We proposed several statistical methods to address this issue. In order to validate our framework of es-

timation of ITE, we carried out a simulation study with inclusion of different kinds of relationships in simulation, and simulation results show our proposed statistical methods maintain a strong power in detecting the heterogeneity of ITE. We also applied our approach to study individualized effects in GAWS data with clinical variables considered as treatments given genetic and other clinical variables as covariates. Our findings from real data analysis are also supported by previous studies.

we hope that this study will open a new avenue for drug development and estimation of ITE, and hence benefit patients eventually.

Acknowledgement

I would like to thank my supervisor Prof. So Hon-cheong for offering me the opportunity to work with him. He teaches me how to conduct research works and gave me generous help in my daily life. He is humorous and highly empathic. Truth to be told, it is a wonderful journey to study and work with him, and I am lucky to have the experience.

I also would like to thank my wife. It's nearly ten years since the first meet, and she has become the most important person in my life. Thanks for supporting my decision to pursue a higher degree and bearing all burden from family to free me from distractions. This is not easy to her. You are a brave girl and wonderful mother for the kids. My any achievement is impossible without you.

Thanks my father and mother for never saying NO to the decision of further my education and for taking care of my kids in the past years. I know the hardness for you to bear the burden. I highly appreciate the support.

Thanks my kids for tolerating my absence of parental responsibility for the past year. You let me be your father and share tremendous joy with me. I promise my love to you will alway be the same.

Thanks my lab mates for having some awesome years with you. You are a part of my daily life in those years. The times we spent together will be a precious memory.

Thanks everybody who helped me for their kindness!

Dedicated to those who risked their life to fight against COVID-19.

Contents

List of Figures

List of Tables

Symbols and Acronyms

In general, we denote a scalar by an italic lower case letter, a vector by a roman lower case bold letter, and a matrix by a roman upper case bold letter respectively, e.g., $a \in \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{M} \in \mathbb{R}^{p \times q}$, with any exceptions to be mentioned in the context case by case.

An identity matrix is written as $\mathbf{I}$. Specifically, an $n \times n$ identity matrix is written as $\mathbf{I}_n$. A zero matrix or vector is written as $\mathbf{0}$. Specifically, an $m \times n$ zero matrix is written as $\mathbf{0}_{m \times n}$.

Specialized symbols and major acronyms are defined as follows:

$\mu_0(\mathbf{x})$	a function for control group $f(\mathbf{x}, w = 0)$
$\mu_1(\mathbf{x})$	a function for treatment group $f(\mathbf{x}, w = 1)$
$p(\cdot)$	the probability density function (PDF)
$\Pr(\cdot)$	the probability value
$\mathbb{E}(\cdot)$	the expectation
$\mathbf{\Sigma}$	a covariance matrix
$\mathbf{N}(\mu, \mathbf{\Sigma})$	a normal distribution with mean μ and covariance $\mathbf{\Sigma}$
ε	a noise vector
$\mathbf{e}(\cdot)$	an error/residual function
$\mathbf{H}$	Hessian matrix
$\mathrm{tr}(\cdot)$	trace of a matrix
$\det(\cdot)$	determinant of a matrix

DNN	deep neuron network
GBM	gradient boosting machine
SVM	support vector machine
CF	causal forests
RF	risk factors
ML	machine learning
EN	elastic net
ITE	individualized treatment effects
TE	treatment effects
tx	treatment
CM	cardiometabolic
GWAS	genome-wide association studies
CNV	copy number variation
SML	supervised machine learning
MCMC	Markov chain Monte Carlo
GRF	general causal forests
CF	causal forests
CV	cross validation
MSE	mean squared error
CI	confidence interval

Chapter 1

Introduction

Machine learning (ML) is one of the fastest growing fields in science in the past decade. At the same time, the rapid accumulation of 'omics' and other forms of biomedical data have revolutionized the field. One of the most pressing questions to date is how to make use of modern ML methodologies and ever-growing data from biomedical sciences to improve our understanding of human diseases and develop better treatments. In this thesis, we will develop and apply several ML approaches to drug discovery/repositioning and predicting individualized effects of risk factors and treatments.

1.1 Drug Discovery Today

New drug development is a lengthy and costly process, and a recent study reported an average cost of $\sim$ 2558 million US dollars in developing a new drug [50]. Part of the reason of the high cost is due to the high failure rate of preclinical drug candidates, which is largely caused by lack of efficacy of these candidates; this indicates that the wrong target is pursued [189]. Computational drug repositioning and target discovery may serve as a new way to shorten the process of drug development, due to the lower cost and more established safety profile of existing drugs [51]. A number of

in silico approaches have been developed for drug repositioning and target discovery and are reviewed elsewhere [95, 219, 107]. With the rapid rise of machine learning (ML) technologies in the past decade, there has been a rising interest in applying ML methods in drug repositioning or target discovery.

Machine learning refers to a vast number of methods for computers to ”learn” and gain insight into data without human interference. These methods are classified into two categories: supervised and unsupervised. Supervised machine learning methods are models for prediction or estimation based on one or more inputs. They are called supervised methods, as their learning is ”supervised” by known output values. On the other hand, unsupervised machine learning methods can be used to detect relationship or patterns underlying ”unlabeled” data. Here we focus on supervised learning methods for classification, since in most cases studies related to drug discovery using ML are the application of classification algorithms. One approach to drug repurposing is to employ drug expression profiles as predictors (i.e., features) to predict a drug's treatment potential. The outcome variable can be the drug category (e.g., whether it is a cardiovascular or anticancer agent) or whether the drug is indicated for a particular disorder (e.g., whether the drug is indicated for diabetes). In the former case, drugs that are classified into categories other than its own indications may be considered for repositioning. In the latter case, drugs with high predicted probabilities but not indicated for the disorder may serve as candidates for repositioning. Additionally, we may also interested in whether the expression profile induced by genetic perturbation (e.g. over-expression or knock down) showed similar pattern to expression profiles of drugs considered as treatments for specific disease, since in this case the perturbed gene can be considered as promising target for the disease. Note that indications for drugs can easily be obtained from publicly available resources such as the Anatomical Therapeutic Chemical (ATC) Classification System [230]. An important advantage is that ML algorithms are abundant and in rapid development, and any existing or new algorithms can be applied without much modification.

With the growth of availability of biomedical data, especially "omics", computational methods can offer a fast, cost-efficient and systematic way to priority promising drug target and drug repurposing candidates for various diseases. The approach has several advantages. Specifically, finding new indications for existing drugs, an approach known as drug repositioning or repurposing, can serve as a useful strategy to shorten the development cycle. Repurposed drugs can be brought to the market in a much shorter time-frame and at lower costs. Meanwhile, computational drug target identification also can speed up the drug development by prioritizing the most promising drug target candidates in a short time, greatly reducing the time in seeking for potential drug targets.

There has been increasing interest in computational drug repositioning recently, in view of the rising cost of new drug development. Hodos et al. provided a comprehensive and updated review on drug repositioning [95], and G. Kandoi et. al. briefly reviewed applications of machine learning and system biology on discovery of target proteins [107]. For the purpose of repositioning, similarity-based methods [77, 161, 134, 140, 155, 126] usually were employed to explore repositioning opportunities, but as noted by Hodos et al., the dependence on data in the "nearby pharmacological space" might limit the ability to find medications with novel mechanisms of actions. Another related methodology is the network-based approach [136], which typically requires data on the relationship between drugs, genes and diseases as well as connections within each category (e.g. drug-drug similarities). It still constraints by the focus on a nearby pharmacological space and the choice of tuning parameters in network construction or inference is often ad hoc [59]. The present work is different in that we apply a broad framework for repositioning and we do not focus on one but many different kinds of learning methods. There is comparatively less reliance on known drug mechanisms or the "nearby pharmacological space" as we let the different algorithms "learn" the relationship between drugs, genes and disease in their own ways. We note that kernel-based ML methods such as support vector machine

(SVM) are also based on some sort of similarity measures. A related work [155] have also examined SVM as a promising ML approach for drug repositioning and identified several interesting candidates. However, here our focus is different in that we considered a variety of other approaches and SVM is one of the methods which falls under the broader framework of ML for repositioning.

Although computational drug repositioning has attracted increased attention recently, few studies focus on psychiatric disorders, compared to other areas like oncology. Psychiatric disorders are leading causes of disability worldwide [223], however there have been limited advances in the development of new pharmacological agents in the last two decades or so [100]. Development of new therapies is also limited by the difficulty of animal models to fully mimic human psychiatric conditions [157]. Investment by drug companies has in general been declining [100] , and new approaches for drug discoveries are very much needed in this field. We will explore repositioning opportunities for schizophrenia along with depression and anxiety disorders. Here depression and anxiety disorders are analyzed together as they are highly clinically comorbid [110, 164], show significant genetic correlations[164], and share similar pharmacological treatments [17].

Meanwhile, we also have witnessed a rise in the interest of computational target discovery in recent years. G. Kandoi et. al. briefly reviewed applications of machine learning and system biology on discovery of target proteins [107]. These studies explored different biological properties by machine learning methods to identify druggable targets [16, 57, 127]. Biological features of human proteins like amino acid composition and amino acid property group composition were studied by a sequence-based prediction method to identify drug target proteins, and a comprehensive comparison of several machine learning methods was conducted [116]. In another study, eight key properties of human drug target were extracted, and learned by support vector machine (SVM) to discover new targets; similar studies extracted simple physicochemical properties from known drug targets and explored the predictive power of

these properties [127, 55]. Topological features of human protein–protein interaction network also were utilized by network based methods to identify potential drug targets [128]. In a recent study, gene-disease association data from Open Targets was explored by four different machine learning methods, including deep neuron networks, to find novel targets, and a large proportion of new targets identified were supported by previous literatures [59]. Dorothea Emig et. al. proposed an integrated network-based method to predict drug targets based on disease gene expression profiles and a high-quality interaction network, and some novel drug targets for scleroderma and other types of cancer were presented [55]. A most recent study proposed pairwise learning and joint learning methods constructed on chemically and genetically perturbed gene expression profiles, and outcome variable was defined as highly correlated pair given by the direct correlation calculation [183]. These studies aim to discover new targets by making use of structural attributes of proteins or properties of known targets, so targets with similar properties usually are identified, but drug targets with novel mechanisms are difficult to identify using this kind of approaches. Network based methods for target discovery, as mentioned previous, rely on known nearby targets to inference potential relationship, so they suffers the same drawback.

Even though repositioned drugs can be brought into market in a much shorter time, drug repositioning may not always be feasible (for example due to side-effects of existing drugs), and drug repositioning and target discovery can complement each other in drug development and pharmacological research. Additionally, in traditional drug development majority of drugs fail to complete the development process due to lack of efficacy, indicating that the wrong target is pursued [189].Usually, a drug target is selected via analyzing how its function influences the disease. However, this process is time-consuming, because investigating a large number of potential targets is usually necessary for finding an ideal one [172]. Computational methods can be utilized to hasten this process by prioritizing promising drug targets.