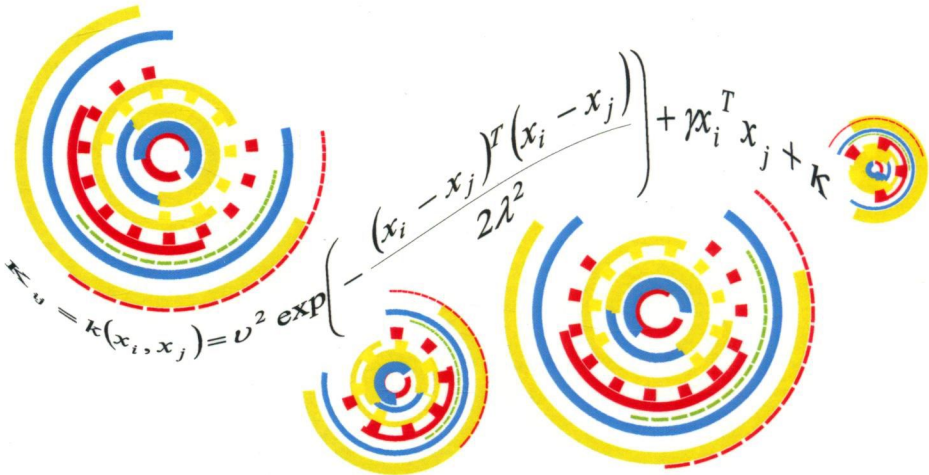
Roger Inge • Jan Leif
Editors

Machine Learning

Advances in Research and Applications



Computer Science, Technology and Applications

NOVA

Machine Learning

Advances in Research and Applications

Roger Inge • Jan Leif
Editors

Contributors

Lei Jia
Hua Gao
Oscar Claveria
Enric Monte
Salvador Torra
Bojan Ploj
Germano Resconi
Ali Yaghoubi
Loris Nanni
Nicolò Zaffonato
Christian Salvatore
Isabella Castiglioni
Alzheimer's Disease Neuroimaging Initiative
F. Dornaika
I. Kamal Aldine
Christos Chrysoulas
Grigorios Kalliatakis
Georgios Stamatiadis

nova
science publishers
www.novapublishers.com

ISBN 978-1-53612-570-2
9 781536 125702

Machine Learning: Advances in Research and Applications • Inge • Leif NOVA

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

MACHINE LEARNING

ADVANCES IN RESEARCH AND APPLICATIONS

ROGER INGE
AND
JAN LEIF
EDITORS

nova
science publishers
New York

NOTICE TO THE READER

Library of Congress Cataloging-in-Publication Data

ISBN: 978-1-53612-570-2

Published by Nova Science Publishers, Inc. † New York

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

MACHINE LEARNING

ADVANCES IN RESEARCH AND APPLICATIONS

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

Additional books in this series can be found on Nova's website under the Series tab.

Additional e-books in this series can be found on Nova's website under the eBooks tab.

PREFACE

In chapter one, Lei Jia, PhD and Hua Gao, PhD analyze machine learning applications in small molecule and macromolecule drug discovery and development while comparing the similarities and differences between the two. They also examine their advantages and limitations with the intent to encourage further creative machine learning applications in drug discovery and development. During chapter two, Oscar Claveria, Enric Monte, and Salvador Torra present a study on the extrapolative performance of several machine learning models in a multiple-input multiple-output setting that permits cross-correlation between the inputs. Bojan Ploj, Germano Resconi, and Ali Yaghoubi parallel the solution of a system by logic gates and by a neural network, in which considerations are computed by the designated one step method during chapter three. In chapter four, Loris Nannia, Nicolò Zaffonatoa, Christian Salvatoreb, Isabella Castiglionib, and the Alzheimer's Disease Neuroimaging Initiative propose a method that could aid in the early diagnosis of Alzheimer's disease. Afterwards, F. Dornaika and I. Kamal Aldine present and experimentally assess two non-linear data self-representativeness coding schemes based on Hilbert space and column generation. Lastly, Christos Chrysoulas, Grigorios Kalliatakis, and Georgios Stamatiadis give an overview of Apache Hadoop, an open-source software framework used to

distribute storage and process big data using the MapReduce programming model.

Chapter 1 - Drug discovery and development is a process to discover and optimize small molecular chemical compounds or macromolecular substances such as proteins for therapeutics use. The whole process engages complex structural and property optimizations. Machine learning tools have been applied in small molecule drug development for many years. The applications are known as quantitative structure-activity relationship (QSAR). In recent years, there is an increasing trend to apply similar methods in protein design and engineering for biologics therapeutics development. In last decade, various machine learning or QSAR methods have been successfully integrated into small molecule drug discovery and development process, especially in the modeling of absorption, distribution, metabolism, excretion, and toxicity (ADMET) as well as other biopharmaceutical properties of drug molecules. ADMET profile of a bioactive compound can impact its efficacy and safety. Moreover, efficacy and safety are considered some of the major causes of clinical attrition in the development of new chemical entities. Recent advances have been made in the collection of data and the development of various *in silico* models to assess and predict ADMET and other biopharmaceutical properties of bioactive compounds in the early stages of drug discovery and development process. Recently, machine learning has been making strong impact in protein design and engineering as well, due to the increase of protein therapeutics especially monoclonal antibodies developments. Applying machine learnings in predicting protein stability, biophysical properties, and chemical hotspots have been integrated as very important steps in biologics drug development. In this chapter, the authors review machine learning applications in both small molecule and macromolecule drug discovery and development, and compare the similarity and difference between the two application domains. In addition, the authors discuss the advantages and limitations as well as how to assess the performance of machine learning models. Finally, the authors hope that this chapter provides insights, facilitates and promotes more creative machine learning applications in drug discovery and development.

Chapter 2 - Machine learning (ML) methods are being increasingly used with forecasting purposes. This study assesses the predictive performance of several ML models in a multiple-input multiple-output (MIMO) setting that allows incorporating the cross-correlations between the inputs. The authors compare the forecast accuracy of a Gaussian process regression (GPR) model to that of different neural network architectures in a multi-step-ahead time series prediction experiment. The authors find that the radial basis function (RBF) network outperforms the GPR model, especially for long-term forecast horizons. As the memory of the models increases, the forecasting performance of the GPR improves, suggesting the convenience of designing a model selection criteria in order to estimate the optimal number of lags used for concatenation.

Chapter 3 - In this chapter the authors compare the solution of a system by logic gates and by a neural network, in which parameters are computed by the denoted one step method. The new method uses vector algebra and a projection operator by which the authors project the output value of the Boolean vector into the space spanned by the input vectors, one for any input variable. The components of the projection into to input space are the weights of the neurons and the weighted average is the neurons output. The authors see that not all Boolean vectors or Boolean functions can be solved by one neuron. In fact the authors have Boolean functions that are in contradiction with the neuronal architecture. To solve the new function the authors extend the number of inputs and the input space where they project the Boolean function. To avoid an exponential explosion of the new inputs, the authors create a special method by which the number of new inputs cannot be more that the number of the old inputs. In conclusion the authors can ascertain how big the number is of the new input associated to hidden neurons, which can be solved by a neural network of the designed Boolean function. At the same time the authors can compute the neuronal parameters weights and thresholds without using the recursion method. In this way the authors avoid the local minimal problem and for the designed function the authors know if the given neural network is in agreement with the Boolean function. If the authors have contradiction the authors have a method to introduce new inputs or hidden neurons to solve the

contradiction itself, by using a new extended neural network. With an extension of the border pairs method the authors give a visual image of the neural or brain contradiction. Such a neural network is associated to a specific lattice or set partial order. When the output function breaks this order the authors have contradiction and they must expand the neural network by new hidden neurons to solve the contradiction itself. Border pairs can be used to understand how and where contradictions grow in the neural network or brain.

Chapter 4 - In this chapter, a method for performing an early diagnosis of AD is proposed, and it combines different feature selection approaches on brain MRI studies. Each selected set is used to train a separate Support Vector Machine (SVM); the results of the ensemble are then combined by a weighted sum rule. Moreover, a novel approach for considering the feature vector as an image is proposed that allows different state-of-the-art texture descriptors to be extracted. The authors report the performance obtained by a histogram of the gradient descriptor. The superior performance of the proposed system is obtained without any ad hoc parameter optimization; in other words, the same ensemble of classifiers and the same parameter settings are used for all datasets. The code to reproduce the experiments will be available at https://www.dropbox.com/s/bguw035yrqz0pwp/ElencoCode.docx?dl=0.

Chapter 5 - Sparse Modeling Representative Selection (SMRS) has been recently introduced for selecting the most relevant instances in datasets. SMRS utilizes data self-representativeness coding in order to infer a coding matrix with block sparsity constraint. The relevance scores of any instance is then set to the `2 norm of the corresponding row in the coding matrix. Since SMRS is based on a linear model for data self-representation, it cannot always provide good relevant representative instances. Besides, most of its selected instances can be found in dense areas in the input space. In this chapter, the authors propose to overcome the SMRS method's shortcomings that are related to the coding matrix estimation. The authors introduce two non-linear data self-representativeness coding schemes that are based on Hilbert space and column generation. Experimental evaluation is carried out on summarizing

a video movie and on summarizing training image datasets used for classification tasks. These experiments demonstrated that the proposed non-linear methods can outperform state-of-the art selection methods including the SMRS method.

Chapter 6 - Organizations are flooded with data. Not only that, but in an era of incredibly cheap storage where everyone and everything are interconnected, the nature of the data the authors are collecting is also changing. For many businesses, their critical data used to be limited to their transactional databases and data warehouses. Nowadays, the variety of data that is available to organizations is tremendous. The challenge here is that traditional tools are poorly equipped to deal with the scale and complexity of much of this data. This is where Hadoop comes in. Hadoop is built to deal with all sorts of messiness. Apache Hadoop is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model. It consists of computer clusters built from commodity hardware. Above all, the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework. This chapter is a welcome to the magical world of Hadoop.

CONTENTS

In: Machine Learning
Editors: Roger Inge and Jan Leif
ISBN: 978-1-53612-570-2

Chapter 1

MACHINE LEARNING IN DRUG DISCOVERY AND DEVELOPMENT

Lei Jia[1,*], PhD and Hua Gao[2,*], PhD
Department of Therapeutic Discovery, Amgen Inc,
[1]Thousand Oaks, CA, US
[2]Cambridge, MA, US

ABSTRACT

Drug discovery and development is a process to discover and optimize small molecular chemical compounds or macromolecular substances such as proteins for therapeutics use. The whole process engages complex structural and property optimizations. Machine learning tools have been applied in small molecule drug development for many years. The applications are known as quantitative structure-activity relationship (QSAR). In recent years, there is an increasing trend to apply similar methods in protein design and engineering for biologics therapeutics development.

* Corresponding Authors Email: leijiachem@gmail.com; hgao@amgen.com.

In last decade, various machine learning or QSAR methods have been successfully integrated into small molecule drug discovery and development process, especially in the modeling of absorption, distribution, metabolism, excretion, and toxicity (ADMET) as well as other biopharmaceutical properties of drug molecules. ADMET profile of a bioactive compound can impact its efficacy and safety. Moreover, efficacy and safety are considered some of the major causes of clinical attrition in the development of new chemical entities. Recent advances have been made in the collection of data and the development of various *in silico* models to assess and predict ADMET and other biopharmaceutical properties of bioactive compounds in the early stages of drug discovery and development process. Recently, machine learning has been making strong impact in protein design and engineering as well, due to the increase of protein therapeutics especially monoclonal antibodies developments. Applying machine learnings in predicting protein stability, biophysical properties, and chemical hotspots have been integrated as very important steps in biologics drug development.

In this chapter, we review machine learning applications in both small molecule and macromolecule drug discovery and development, and compare the similarity and difference between the two application domains. In addition, we discuss the advantages and limitations as well as how to assess the performance of machine learning models. Finally, we hope that this chapter provides insights, facilitates and promotes more creative machine learning applications in drug discovery and development.

Keywords: QSAR, Cubist, data set, descriptor, multi-parameter optimization, protein design, thermostability, deamidation, application domain, prediction confidence

ABBREVIATIONS

ADMET	(*Absorption, Distribution, Metabolism, Excretion, and Toxicity);*
ADME	(*Absorption, Distribution, Metabolism, and Excretion);*
QSAR	*(Quantitative Structure-Activity Relationship);*
SVM	(Support Vector Machine);
RF	(Random Forests);
NBC	(Naïve Bayes Classifier);