Decomposition Methods for Cell Line Representations in Drug Effect Predictions

Bahattin Can Maral
bahattincanmaral@etu.edu.tr
TOBB University of Economics and Technology
Ankara, Turkey

ABSTRACT

The discovery of new biological interactions, such as interactions between drugs and cell lines, can improve the way drugs are developed. Recently, there has been a surge of interest for predicting interactions between drugs and targets using deep neural networks and specifically, using neural networks to predict drug activity on cellular lines. However, the variety of the cell lines, the size of the cell line vertices, and the dozens of different drug representations don't always work in the favor of researchers. In this research, we focused on the cell line representation of the GDSC (Genomics of Drug Sensitivity in Cancer) dataset to improve on the previous attempts on the same focus. In the end, reducing the 20 000 length cell line vertices to 256 length ones with PCA (Principal Component Analysis) proved to be superior when compared to other decomposition types we tried, with a 0.91 R2 score.

KEYWORDS

datasets, neural networks, drug predictions, cell line representation

ACM Reference Format:
Bahattin Can Maral. 2021. Decomposition Methods for Cell Line Representations in Drug Effect Predictions. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Over the past two decades, substantial improvements in high-throughput profiling technologies and systems approaches have increased expectations that personalized or precision medicine will become the paradigm of future medical science [18][12][4]. In contrast to the one-size-fits-all approach that has dominated cytotoxic chemotherapy, personalized medicine exploits tumor response and vulnerability based on identified molecular traits to overcome some of the limitations associated with conventional symptoms-oriented disease diagnoses and therapies.The most important step in implementing personalized medicine will be the identification of biomarkers useful for predicting the drug response of a given patient [5][19]. However, the development of predictive biomarkers would require substantial efforts and is often prohibitively expensive in human or animal models. Therefore, many studies conduct large-scale drug screenings on cultured human cell line panels to identify predictive biomarkers [11] of the earliest such attempts is the NCI-60 study [16][15], which included a set of 60 human cell line sand their responses to more than 100,000 chemical compounds. Drug response results for the NCI-60 dataset [9] [13] revealed that different types of cancers have different drug response signatures, and that different tumors derived from the same type of cancer may have distinct molecular patterns [1].

One of the most prominent recent datasets, Genomics of Drug Sensitivity in Cancer (GDSC)[20], systematically addressed the issue of predictive biomarker identification by collectively analyzing around 850 clinically relevant human cell lines and their pharmacological profiles for 178 cancer drugs. In the study, an ElasticNet[21] model is used to fit the data to predict the drug responses. However, the linear model approach has been left in the dust with the recent powerful machine learning models and the drug representation techniques emerging every day. In one of these approaches[10], an auto-encoder model is fit to summarize the cell line vertices to usable sizes while for the drug representation, the state-of-the-art ECFP[14] fingerprints are used. Then these 2 representations fed into a deep neural network to predict the IC50 values.

Inspired by aforementioned research, we decided to replace the auto-encoder with different decomposition models to see if it could be replaced with a different, more available approach.

2 REPRESENTATIONS

2.1 Cell Line Representations

In the original GDSC data the cell lines are represented with 20 000 length vertices, to avoid the curse of dimensionality, we used various feature projection models to summarize the cell line data.

2.1.1 Principal Component Analysis. Principal component analysis (PCA)[17], the most common linear technique for dimensionality reduction, performs a linear mapping of the data to a lower-dimensional space with the aim of maximizing the variance of the data in the low-dimensional representation. In practice, the data's covariance (and sometimes correlation) matrix is built, and the eigenvectors are computed on this matrix.

2.1.2 Random Projection. Random projection[2] is an easy and computationally effective method for reducing data dimensionality by exchanging a controlled amount of error for shorter processing times and smaller model sizes. The dimensions and distribution of random projection matrices are carefully managed to keep the pairwise distances of any two dataset samples as close as possible.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.1.3 FastICA. FastICA[7] (Fast Independent Component Analysis), an accessible and widely used algorithm for independent component analysis. FastICA, like most ICA algorithms, uses a fixed-point iteration scheme to find an orthogonal rotation of prewhitened data that maximizes a metric of non-Gaussianity of the rotated components.

2.1.4 Sparse PCA. The analysis of Sparse prinpipal components (PCA sparse)[6] is used in special technologies used in statistical analysis, especially multivariate data sets. A particular disadvantage of normal PCA is usually that the main component is a linear combination of all input variables. Spare PCA exceeds this disadvantage when finding a linear combination that contains several input variables.

2.1.5 Encoding. We also tried to encode the cell line names using One Hot Encoding and Multi Label Binary Encoding methods, to see if the 20 000 length vertices were actually useful, which they were. These approaches yielded significantly worse scores compared to their counterparts.

2.2 Drug Representations

2.2.1 Morgan Fingerprints. This family of fingerprints, better known as circular fingerprints, is built by applying the Morgan algorithm to a set of user-supplied atom invariants. In our research we transformed our drugs in the SMILES format to various sized and different featured morgan fingerprints using the RDkit libraries. The fingerprints acquired with this method corresponds to ECFP2, ECFP4, FCFP2, and FCFP4.

2.2.2 MACCS Fingerprints. Another widely used fingerprint type is the 166-bit length MACCS fingerprints[3]. This fingerprint type was also acquired using RDkit[8].

3 EXPERIMENTS

To fit the newly acquired representations a deep neural network model was created. After hyper optimizing the network parameters for each possible combination of drug and cell line representation combination, a 5 layer deep neural network (Layers with 2048, 1024, 512, 128, 64 neurons) model was created with batch normalization and dropout layers with 0.4 strength in between. The final model was trained with 256-length PCA'd cell line representations and 512-bit length FCFP4 drug fingerprints with 10-fold cross validation.

During the experiments, while we thought would be the most hopeful feature projection type of the bunch, Sparse PCA was proved to be not possible to aquire after terminating the first fold's runtime on the 24th hour mark, probably due to hardware limitations.

In the end, among all the combination of representations and the network types, aforementioned model has acquired 0.91 R2 score with a respectable 0.8 RMSE.

Compared to DeepDCS's scores there is a visible improvement of 0.13 on the R2 score, there is a trade-off of 0.3 RMSE. Also there is a visible difference between the scores DeepDSC compared their scores to and our's.

4 CONCLUSION

In the end, even though previously mentioned scores couldn't compete with it's contemporaries, it was a valiant effort to compete in-explainable models such as autoencoders with explainable models. In the future, we would like to see if the SparsePCA method would work or not in an addition with data that could be added from the Cancer Cell Line Encyclopedia (CCLE). Also other cell line representation methods such as omics profiles could be tried in training as extra or replacement features for the cell line data.

REFERENCES

[1] You Han Bae. 2009. Drug targeting and tumor heterogeneity. *Journal of controlled release: official journal of the Controlled Release Society* 133, 1 (2009), 2.

[2] Ella Bingham and Heikki Mannila. 2001. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.* 245–250.

[3] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. Molecular fingerprint similarity search in virtual screening. *Methods* 71 (2015), 58–63.

[4] Rui Chen and Michael Snyder. 2013. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 5, 1 (2013), 73–82.

[5] Juan Cui, Yunbo Chen, Wen-Chi Chou, Liankun Sun, Li Chen, Jian Suo, Zhaohui Ni, Ming Zhang, Xiaoxia Kong, Lisabeth L Hoffman, et al. 2011. An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer. *Nucleic acids research* 39, 4 (2011), 1197–1207.

[6] Alexandre d'Aspremont, Laurent E Ghaoui, Michael I Jordan, and Gert R Lanckriet. 2005. A direct formulation for sparse PCA using semidefinite programming. *Advances in neural information processing systems* 17 (2005), 41–48.

[7] Zbynek Koldovsky, Petr Tichavsky, and Erkki Oja. 2006. Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Transactions on neural networks* 17, 5 (2006), 1265–1277.

[8] Greg Landrum. 2013. Rdkit documentation. *Release* 1, 1-79 (2013), 4.

[9] Jae K Lee, Dmytro M Havaleshko, HyungJun Cho, John N Weinstein, Eric P Kaldjian, John Karpovich, Andrew Grimshaw, and Dan Theodorescu. 2007. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proceedings of the National Academy of Sciences* 104, 32 (2007), 13086–13091.

[10] Min Li, Yake Wang, Ruiqing Zheng, Xinghua Shi, Fangxiang Wu, Jianxin Wang, et al. 2019. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM transactions on computational biology and bioinformatics* (2019).

[11] Ultan McDermott, Sreenath V Sharma, Lori Dowell, Patricia Greninger, Clara Montagut, Jennifer Lamb, Heidi Archibald, Raul Raudales, Angela Tam, Diana Lee, et al. 2007. Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proceedings of the National Academy of Sciences* 104, 50 (2007), 19936–19941.

[12] Reza Mirnezami, Jeremy Nicholson, and Ara Darzi. 2012. Preparing for precision medicine. *N Engl J Med* 366, 6 (2012), 489–491.

[13] Gregory Riddick, Hua Song, Susie Ahn, Jennifer Walling, Diego Borges-Rivera, Wei Zhang, and Howard A Fine. 2011. Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics* 27, 2 (2011), 220–224.

[14] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50, 5 (2010), 742–754.

[15] Robert H Shoemaker. 2006. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer* 6, 10 (2006), 813–823.

[16] John N Weinstein, Timothy G Myers, Patrick M O'Connor, Stephen H Friend, Albert J Fornace, Kurt W Kohn, Tito Fojo, Susan E Bates, Lawrence V Rubinstein, N Leigh Anderson, et al. 1997. An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 5298 (1997), 343–349.

[17] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.

[18] Paul Workman, Paul A Clarke, and Bissan Al-Lazikani. 2012. Personalized medicine: patient-predictive panel power. *Cancer cell* 21, 4 (2012), 455–458.

[19] Yang Xie, Guanghua Xiao, Kevin R Coombes, Carmen Behrens, Luisa M Solis, Gabriela Raso, Luc Girard, Heidi S Erickson, Jack Roth, John V Heymach, et al. 2011. Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non–small-cell lung cancer patients. *Clinical Cancer Research* 17, 17 (2011), 5705–5714.

[20] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson,

et al. 2012. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research* 41, D1 (2012), D955–D961.

[21] Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* 67, 2 (2005), 301–320.