Kernelized Rank Learning for Personalized Drug Recommendation

Xiao He[1,2], Lukas Folkman[1,2], and Karsten Borgwardt[1]

[1]Machine Learning and Computational Biology Lab and Swiss Institute of Bioinformatics, Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2]These authors contributed equally to this work.

Abstract

Large-scale screenings of cancer cell lines with detailed genomic profiles against libraries of pharmacological compounds are currently being performed in order to gain a better understanding of the genetic component of drug response and to enhance our ability to predict drug sensitivity from genetic profiles. These screens differ from the clinical setting in which (1) medical records only contain the response of a patient to very few drugs, and in which (2) selecting the most promising out of all therapies is more important than accurately predicting the sensitivity to the given drug. Current regression models for drug sensitivity prediction fail to account for these two properties. We present a machine learning approach, named Kernelized Rank Learning (KRL), that ranks drugs based on their predicted effect per patient, circumventing the difficult problem of precisely predicting the sensitivity to the given drug. Our approach outperforms several state-of-the-art predictors in drug recommendation, particularly in a clinically-relevant case where few training data are available.

1 Introduction

One of the key goals of precision medicine is the ability to suggest personalized therapies to patients based on their genomic profiles. For the development of targeted cancer treatments, a large collection of patients with recorded clinical outcomes and genomically characterized tumor samples is needed, making this approach currently prohibitive. Therefore, pre-clinical biological models, such as cultured human cancer cell lines, are a relatively inexpensive alternative approach for finding biomarkers. Recently, several large-scale drug sensitivity screens of genomically profiled cell lines have been established [1]. One of the current challenges lies in building accurate predictive models and translating these models into the clinic.

A number of regression models [2] have been proposed to predict drug sensitivity measured by the half maximal inhibitory concentration (IC_{50}). However, there are two potential problems with these approaches. Firstly, the methods are not directly optimized for the prediction of the clinically-relevant case in which a clinician needs

to know a few most-suited drugs for a given patient. Secondly, patient knowledge banks are much sparser than cell line panels, meaning that only a few therapies are recorded for each patient compared to a comprehensive screen of tens to hundreds of drugs for each cell line.

Here, we phrase the personalized drug recommendation as a ranking problem and propose a method, named Kernelized Rank Learning (KRL), that is directly optimized for the clinically-relevant scenario of personalized drug recommendation.

2 Methods

We stated the personalized drug recommendation as a ranking problem in which we are given $X \in \mathbf{R}^{n \times p}$, a set of n cancer patients where each of them is represented with a genomic profile (e.g., gene expression) with p variables. Furthermore, we are given $Y \in \mathbf{R}^{n \times m}$, a set of drug response measurements to m distinct drugs that have been prescribed to these patients. The majority of values in Y are typically missing since each patient undergoes only a few therapies. Our goal is to recommend the most promising therapies to new patients, not present in X, based on their genomic profiles. This goal is thus different from predicting the exact value of the drug response.

To evaluate goodness of our recommendations, we adopted the Normalized Discounted Cumulative Gain (NDCG@k) [3] that considers the ranking among the top k recommendations. Given the true drug response $y \in \mathbf{R}^m$ for a patient, sorted in a decreasing order, and the predicted ranking vector $f \in \mathbf{R}^m$, our goal is to evaluate if the order of the top recommendations in f agree with the order of the most effective drugs in y:

$$\text{DCG@}k(f, y) = \sum_{i=1}^{m} \delta(r(f_i) \leq k) \frac{2^{y_i} - 1}{log_2(1 + r(f_i))} \tag{1}$$

where $\delta(x) = 1$ if x is true and 0 otherwise, and $r(f_i)$ returns the rank of f_i in f. NDCG@k is the normalized DCG@k in the range of $[0, 1]$.

The proposed KRL method is based on the idea of optimizing a convex upper bound of NDCG@k loss ($1 - \text{NDCG@}k$) proposed in [4]:

$$\text{loss}(f, y) = \max_{\pi}(c^T f[\pi] - a[\pi]^T b) - c^T f + a^T b \tag{2}$$

where π is a permutation, $a_i = \frac{\delta(i \leq k)}{log(i+1)}$ and $b_i = 2^{y_i} - 1$.

Finally, the kernelized objective function of KRL is defined as

$$\min_{\beta} \sum_{i}^{n} \text{loss}(K_i \beta, Y_i) + \lambda trace(\beta^T K \beta) \tag{3}$$

where K is a kernel matrix on X. Eq. 3 is convex but non-smooth. Therefore, we use the Bundle Method [5] for optimization.

We employed the Genomics of Drug Sensitivity in Cancer (GDSC) dataset (release 6) with drug sensitivity measurements for 265 compounds across 1,001 cancer cell lines ranging 30 cancer types [1]. We encoded each cell line with the basal expression of 17,737 genes.

3 Results

First, we used the full GDSC dataset to compare the performance of KRL with related work. KRL yielded more accurate or comparable drug recommendations than any of the four compared methods: Elastic Net (EN), Kernel Ridge Regression (KRR), Random Forest (RF), and Kernelized Bayesian Multi-Task Learning (KBMTL) [6]. Figure 1A shows the comparison for NDCG@k with $k \in \{1, 3, 5, 10, 15, 20\}$.

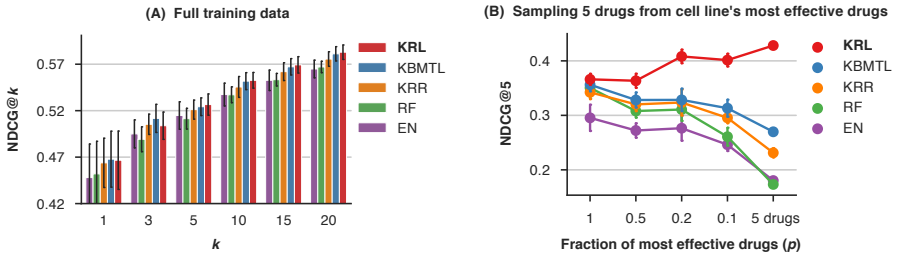

Figure 1: **(A)** Comparison of KRL with related work using the the full training dataset for different values of the evaluation parameter k. **(B)** Comparison using the subsampled training dataset, keeping only five drugs for each cell line.

Next, to evaluate the compared methods in a more clinically-relevant scenario, in which only a few therapies are recorded for each patient, we subsampled the training dataset so that it comprised only five therapies per cell line. Furthermore, to simulate the clinical case where a doctor prescribes these therapies based on their expert judgment, we implemented the following subsampling strategy: first, a predefined proportion (e.g., $p = 20\%$) of the most effective drugs per cell line was selected, second, five drugs were sampled randomly from this selection. Figure 1B shows NDCG@5 as a function of the fraction of the most effective drugs used for the subsampling. As expected, the performance of all compared methods decreased compared to the results using the full training dataset. Nonetheless, the NDCG@5 improvement of KRL compared to the second best method, KBMTL, was gradually increasing from 1% (for sampling from the full dataset) to 13% (for sampling the five most effective drugs).

To analyze the performance of the compared methods under this clinically-relevant scenario in more detail, we looked at different values of the evaluation parameter k. Figure 2A shows that the improvement of KRL was stable for a wide range of k (12–9% for $k \in \{1, 3, 5, 10\}$ and 6–3% for $k \in \{15, 20\}$). This overall trend was present regardless of the subsampling strategy (the proportion of the most effective drugs used for training).

Finally, we compared how close the single top recommendations of KRL and related work were to the most effective drug for the given cell line. In this evaluation, a method scores 'rank one' if its top recommendation was the closest to the most effective drug out of the five compared methods. Figure 2B shows that KRL greatly outperformed related work when trained using only five drugs from the 10 or 20% of the most effective drugs, with around 60% of recommendations ranked in the first place.

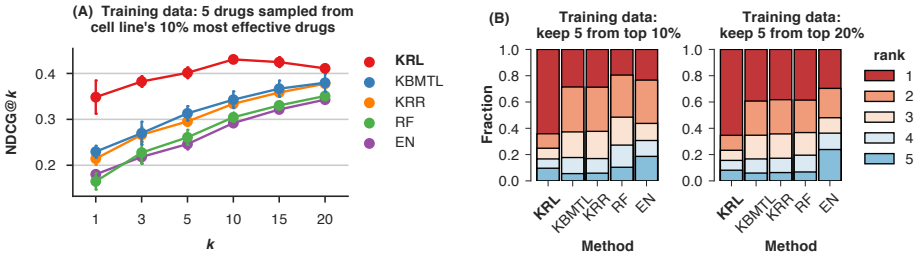

Figure 2: **(A)** Comparison of different values of the evaluation parameter k using the subsampled training dataset, keeping only five from the cell line's 10% most effective drugs. **(B)** Comparison of KRL with related work based on how close their single top recommendations were to the most effective drug ('rank one' means that the method's top recommendation was the closest to the most effective drug).

4 Conclusions

This work phrased personalized drug recommendation as a ranking problem of choosing the most promising therapies. This follows from the observation that in a clinical setting, we are interested in providing a recommendation of the most effective drugs rather than predicting the exact response measurements of all available drugs. To this end, we proposed the KRL method which directly optimizes the ranking loss function. In our empirical evaluation, KRL outperforms state-of-the-art predictors in drug recommendation. Most notably, KRL's improvements were largest when we subsampled the training dataset to simulate a clinically-relevant scenario in which only a few therapies are known for each patient.

References

[1] Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., et al. *Cell* **166**(3), 740–754 (2016).

[2] Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Hintsanen, P., Khan, S. A., Mpindi, J.-P., et al. *Nature biotechnology* **32**(12), 1202–1212 (2014).

[3] Järvelin, K. and Kekäläinen, J. *ACM Trans. Inf. Syst.* **20**(4), 422–446 (2002).

[4] Le, Q. V. and Smola, A. J. *CoRR* **abs/0704.3359** (2007).

[5] Teo, C. H., Vishwanathan, S. V. N., Smola, A. J., and Le, Q. V. *Journal of Machine Learning Research* **11**, 311–365 (2010).

[6] Gönen, M. and Margolin, A. A. *Bioinformatics* **30**(17), i556–i563 (2014).