Extending Cancer Cell Drug Response Models

Wei Zhou
weiz2@stanford.edu

Bill Evans
bill_evans@post.harvard.edu

Paul Butler
paulgb@stanford.edu

Abstract

This project examines methods for predicting the degree to which various drugs inhibit the growth of a range of cancer cell types. Data consist of over 40,000 features for each of 432 cell lines and cell growth inhibition measurements for 24 drugs. The high dimensionality of this data raises challenges that are addressed through Elastic Net parameter tuning. We also investigate Principal Component Analysis and the use of principal components as features to reduce dimensionality without compromising accuracy. Finally, we explore Support Vector Regression and Random Tree models for this data.

1 Background

In recent years, the study of cancer and cancer genomics has been supported by the development of openly-available research data sets. These include The Cancer Genome Atlas (TCGA), the Wellcome Trust Sanger Institutes Catalogue of Somatic Mutations in Cancer (COSMIC) database, Sangers Genomics of Drug Sensitivity in Cancer (GDSC) [5], and the Broad Institute's Cancer Cell Line Encyclopedia (CCLE) [1].

The CCLE seeks to address the challenges associated with "The systematic translation of cancer genomic data into knowledge of tumour biology and therapeutic possibilities." [1]. It does this by bringing together drug response data, cell line phenotype data, and genomic data.

The feature data contains 40,492 features for each of 432 cell lines. The features are grouped into three categories: 15,702 features representing gene expression, 23,601 features representing copy-number variation, and 1,667 binary features indicating the presense of oncogenes.

The response data consists of a 432x24 matrix in which each data point represents the IC_{50} inhibition dosage[1] of a single drug for a given cell line. This structure lends itself to predictive modeling in which a single column vector (i.e., the drug response data for a single drug across all 432 cell lines) is employed as the response variable for the feature data. Alternatively, the entire response matrix can be modeled simultaneously in a multitask model.

2 Preprocessing of Data

2.1 Normalization and Imputed Values

Continuous valued and discrete non-binary (gene expression and copy-number variation, respectively) valued feature data were scaled; binary valued data (oncogene indicator variables) were not altered.

A portion of the response variable matrix contained missing values. In the CCLE data set, 390 of 10,368 (approximately 3.7%) of response values were missing. A regression-based model, Elastic Net cannot fit missing values. This necessitates imputation of response data. Comparative evaluation of imputation methods allows for more robust estimation [12]. We employed SVD, K-Nearest Neighbor (KNN), regression, and

[1] This is the amount of the compound required to inhibit the cell growth by half.

SVT. Each imputation method was scored using leave-one-out cross validation in which a single known value was removed from the model. Those known values were then imputed using each of the methods under consideration.

RMSE for each method was calculated as the difference between the known, held-out data and the imputed data. Four methods (KNN, LM, mean value, SVD, and SVT) were fully evaluated. The optimal imputation method, SVD, had an RSME of .73 compared with worst-case performance for SVT (RMSE=2.14). SVD-imputed response data were employed in the modeling throughout this project to ensure comparability between results from models which do not handle missing data and those that do. Missing values were not imputed for test data (in order to avoid misstatement of the validity of the model that could result).

2.2 Feature Selection Methods

The dimensionality of cell line genetic data exceeds the number of observations by more than 100-fold. Feature selection is necessary in order to enhance model prediction (by identifying and excluding features that lack significant explanatory power) and to improve the training run-time performance.

Studies of cell line genomic and drug response data include Elastic Net regression [1, 3, 5]. Automatic variable selection accomplished with Elastic Net lends the model naturally to the high dimensionality of genetic data [2].

Pearson correlation of features with response variables has been utilized for feature selection. (i.e., Only features with a Pearson's coefficient above a specified threshold are included in the model.) [1]. While feature reduction is achieved, this approach relies on an arbitrarily established cutoff.

As an alternative, we employed PCA for dimension reduction and then model the principal components as features. We established the relationship between the variance captured by PCA and model error for a given number of principal components. The first 286 principal components capture more than 95% of variance in the original data. This decrease reduces model training latency dramatically with only modest increases in Elastic Net RMSE.

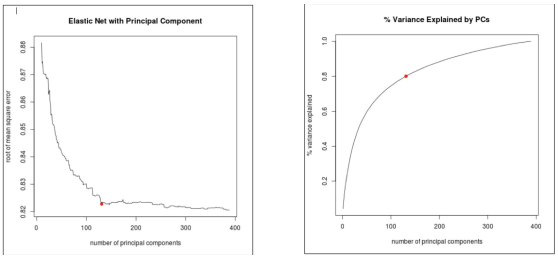

Figure 1: Principal components were used as features of Elastic Net Regression. In the plot on the left, the rate in decrease in RMSE slows sharply near 131 principal components. Similarly, 131 principal components account for approximateley 80% of the variance in the feature data. The red dot marks these points in each graph respectively.

3 Methodology and Results

We first performed Elastic Net regression individually on single-drug response vectors. Later, we compared this result to Multitask Elastic Net, in which the complete response matrix is utilized in a single predictive model. Finally, we explored Support Vector Regression (SVR) and Random Forest models with a view on the tradeoff between model training performance and model error.

3.1 Elastic Net

Elastic Net is a regularized regression model that combines Lasso and Ridge regression L_1 and L_2 norms.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}(\|y - X\theta\|^2 + \lambda_2\|\theta\|^2 + \lambda_1\|\theta\|_1) \tag{1}$$

The Lasso penalty tends to select individually correlated predictors and discards the others while Ridge shrinks them towards each other [2]. The Elastic Net penalty mixes the two while also overcoming a limitation faced by Lasso, namely its inability select more features than there exist observations in the dataset.

In the above equation, λ_1 and λ_2 correspond to the Ridge and Lasso penalties, respectively. Elastic Net has the capacity to select groups of correlated variables[2], ideally suiting it to genetic data in which genes are believed to interact in causative networks.

For each compound we first searched for the pair of parameters λ_1, λ_2 which achieved the lowest R^2 on a holdout set. We experimented with coordinate descent, but found grid search over a small grid to be faster because it could be easily parallelized. However, grid search necessarily discretized the search ranges of λ_1 and λ_2. To address this, once having completed the grid search for each Elastic Net regression on individual drug response data (using Python scikit-learn [6]), we employed the R glmnet package to select λ_2 over a continuous range.

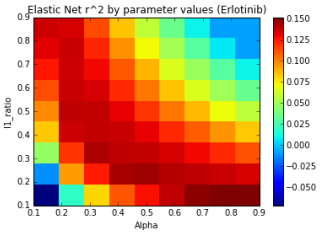

Figure 2: Best parameter search results for one compound, Nutlin3

Figure 2 shows the results of the grid search for one compound, Nutlin3. (Note that axes are labeled using scikit-learn parameter names, in which $l1\_ratio = \lambda_1/(\lambda_1 + \lambda_2)$ and $alpha = (\lambda_1 + \lambda_2)$, the L1 regularization and $alpha$ with λ_2).

In this way cell line response to each drug compound was examined in isolation. However, we observed that several compounds have highly correlated response variables. (See Figure 3.) Modeling compounds in isolation, while reasonable, failed to exploit this characteristic.

In order to take advantage of correlations in compound response we implemented Multitask Elastic Net, which models the interaction as a multivariate Gaussian. RMSE for Multitask Elastic Net using all zero values assigned to missing response variables in the training dataset yielded an RMSE of 0.73 on the test dataset, and an RMSE of 0.68 was achieved on the test data set when using SVD-imputed response training data.

These results from Multitask Elastic Net underperformed the average result of individual Elastic Net regression (which had an average RMSE of .56). Because the multitask model can make use of additional information (i.e., the correlation among different response variables), we anticipated a performance improvement.

A total of 647 features, 1.60% of the original feature set, were active variables in multitask Elastic Net. This warrants additional investigation.

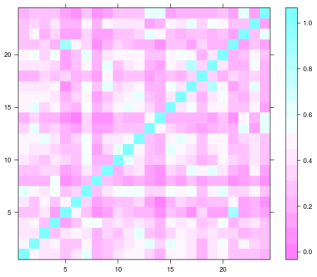

Figure 3: Correlation matrix of compound response data.

3.2 Support Vector Regression

We investigated the Support Vector Regression (SVR) model to explore the comparative runtime performance of the learning algorithm and its accuracy on highly dimensional data.

SVR parameter tuning was performed using Python sklearn GridSearchCV with 8-fold cross-validation. RBF kernels were tuned over a range of values for kernel parameter $\gamma \in \{.001, .01, .1, 1\}$, and the parameter $C \in \{.001, .01, .1, 1\}$ was tuned in both linear and RBF models using RMSE to score model prediction. Given a training set $S = \{(x_i, y_i); i = 1, ..., m\}$, the Lagrange optimization for SVR [4] with slack variables ξ_i and ϵ precision is given by:

$$\text{minimize: } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{2}$$

$$\text{subject to } \begin{cases} y_i - \langle w, x_i\rangle - b \leq \epsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

RBF tuning parameter:

$$K(x, z) = exp(\gamma\|x - z\|_2^2) \tag{3}$$

Kernel parameter selection varied across the respective search domains. However, selection of the optimal kernel was nearly uniform with all but three models performing optimally using a RBF kernel. In three cases, a linear kernel outperformed RBF, indicating a possible difference in structure in the data as well as the underlying biology for these drug-cell line pairs.

3.3 Random Forest

We implemented the random forest model to establish its comparative performance on a wide feature matrix. The random forest model is an ensemble of a set of regression trees. In addition to constructing each tree using a different bootstrap sample of the data, random forests added an additional layer of randomness by splitting each node using the best split among a subset of predictors randomly chosen at that node [11]. The random forest training algorithm complexity is $O(M(mn \log n))$ (with $M = 5$, the number of trees; $m = 389$, the number of observations; n=40,492, the number of original features).

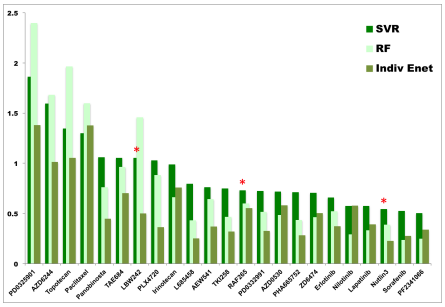

Figure 4: RMSE for SVR, Random Forest, and individual Elastic Net for all 24 drugs. Mean RMSE was 0.89, 0.80, and 0.56 for SVR, RF, and Elastic Net respectively. Red asterisk indicates selection of linear (as opposed to RBF) kernel in SVR regression.

The random forest was built with the randomForest package in R v3.0.2. Due to the complexity of computing a forest on all features, a single random forest was built using the first 286 principal components (instead of the original feature vector). Each random forest was built by 5 regression trees using 1/3 of the features. Figure 4 compares random forest performance to SVR and Elastic Net.

4 Data Extensions: PaDEL Dataset

Other research has investigated the degree to which chemical features can of themselves be used as model features in biological data sets. These experiments rely on software that produces a feature set that describes the physical structure of a chemical agent.

Menden, M. et al. [8] describe a method for modeling predictions on the GDSC cell line data by adding chemical descriptors to the genetic feature data [7]. A neural network and random forest were trained on the resulting feature set. We were not able to find research that makes use of gene data alongside chemical descriptors in an Elastic Net regularization framework.

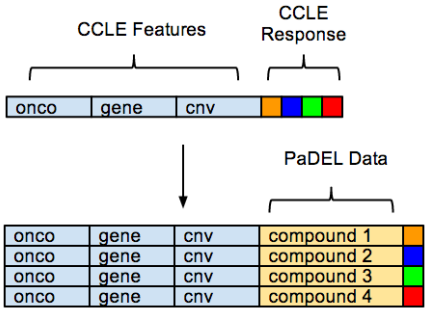

Figure 5: Integration of PaDEL data with CCLE dataset. Prior to adding PaDEL compound data, cell line features appear in a single row vector. After adding PaDEL data for each compound, each cell line's feature vector is repeated (once for each response/compound pair.

We collected SMILES [9] files for 23 of the 24 compounds represented in the CCLE data from PubChem. The structure for one chemical, LBW242, was not in the database and so was omitted from our dataset. These SMILES files were used as input to PaDEL [7], a program which converts molecule representations into a variety of numeric properties describing the underlying chemical structure. The resulting output created 770 features that we added to the CCLE feature data.

Incorporating PaDEL data required a transformation of the feature and response matrices. The CCLE data contains 24 response variables for each feature vector. To add the PaDEL compound data, we exploded the response vector into individual values spread over multiple rows as shown in Figure 5.

The resulting dataset was fed into Vowpal Wabbit [10], a high-performance implementation of gradient descent learning. We used a subset of its features equivalent to Elastic Net regression.

The addition of PaDEL data did not result in improved performance. RMSE was 0.79 with an R^2 of .67, worse performance than Multitask Elastic Net for the base CCLE data. A comparison of included (versus excluded) features did not yield insight into the reduced in performance that resulted from PaDEL data.

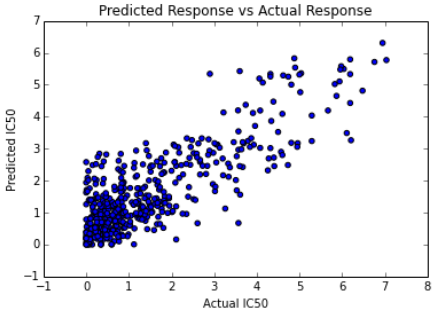

Figure 6: Predicted versus actual IC_50 drug response using PaDEL data in Multitask Elastic Net.

5 Conclusion

We propose to explore the addition of gene network information to the model, whereby the known biological association between genes (amplification, silencing, etc) is utilized. It would likewise be interesting to explore feature selection methods with PaDEL data included, possibly making use of information about the biological relevance of particular physical chemical features to enhance the selection process. Separately, we would like to further investigate the degree to which Elastic Net is able to predict gene associations using this data through a measure of similarity in regression coefficients.

While our models cannot replace lab experiments, we have established a baseline for multitask modeling. This information may be help guide compound/cell-line combinations that merit laboratory investigation.

Acknowledgments

We would like to thank David Knowles for his support and for providing the cell line data used in this project.

References

[1] Barretina, J. et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* **483**:603607.

[2] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *J. Royal. Stat. Soc. B.* **67**(2):301-320.

[3] Garnett, M. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature.* **483**:570575

[4] Smola, A. and Scholkopf, B. (2003) A Tutorial on Support Vector Regression. *Statistics and Computing* **14**:199-222

[5] Yang, W. et al. (2013) Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research* **41**:D955-D961

[6] Pedregosa et al. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**:2825-2830

[7] Yap CW (2011) PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **32**:1466-1474

[8] Menden, M. et al. (2013) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE* **8**(4):e61318 doi:10.1371/journal.pone.0061318

[9] Weininger, David (1988). SMILES, a chemical language and information system. *Journal of Chemical Information and Modeling* **28**(1):316

[10] Langford, L. et al. Vowpal Wabbit http://hunch.net/~vw/

[11] L. Breiman. Random forests (2001). *Machine Learning* **45**(1):532

[12] Hastie, T.et al. Imputing Missing Data for Gene Expression Arrays. *Technical Report, Division of Biostatistics, Stanford*