Predicting Drug Induced Liver Injury Through Combined Genomics Indicator and Ensemble Machine Learning Approaches

Zhixiu Lu[1], Miyuraj Harishchandra Hikkaduwa Withanage[2] and Erliang Zeng[1,2,*]
[1]**Department of Computer Science,** [2]**Department of Biology, University of South Dakota, Vermillion, SD, USA**
[*] **E-mail: Erliang.Zeng@usd.edu**

1 Introduction

Conventional toxicity assessment is usually conducted using indicators such as pathology and clinical chemistry data, which could only detect around 60% of drug-induced liver injury (DILI) cases in the preclinical studies. To improve the drug safety assessment during the preclinical and/or early clinical stages, modern "omics" techniques including high-throughput microarray and next-generation sequencing have been used to identify alternative genomic biomarkers for risk assessment. The underlying hypothesis is that genomic biomarkers will be more sensitive than conventional markers in detecting toxicity signals in early drug development stages.

In this paper, we designed a computational framework to predict DILI of drug compound from cell-based Connectivity Map (CMap) gene expression responses of two different cancer cell lines (MCF7 and PC3) [1], a CAMDA 2018 challenge. The computational framework takes advantage of ensemble feature selection methods to first identify candidate discriminative genomic indicators, and then evaluates the performance of those genomic indicators by ensemble classifiers. Finally, the inherent connections among genomic indicators identified were explored using network analysis. The network analysis is able to discover the redundancy among genomic indicators and benefits the identification of a set of optimum non-redundant genomic indicators, so that improves the DILI prediction. We use the ROC (receiver operating characteristic) curve and AUC (area under the curve) to comprehensively evaluate our methods. The cross-validation results show that our method can achieve higher AUCs, indicating the effectiveness of our ensemble machine learning methods.

2 Materials and Methods

We explored the possibility of predicting the DILI potential in humans using the CMap data [1], which provided gene expression dataset generated from two different cell lines (MCF7 and PC3). The MCF7 is a breast cancer cell line and the PC3 is a prostate cancer cell line. Data from each cell line is divided into a training set and a validation set. The training set contains clinical DILI results as training labels for 190 drugs. The clinical DILI labels are binary("1" indicates "DILI " and "0" means "No DILI").

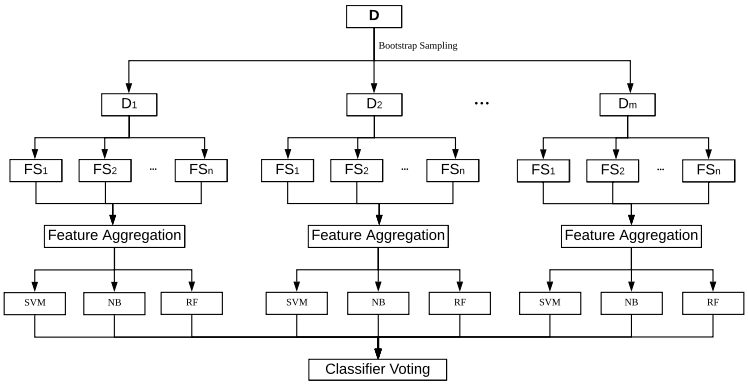

Figure 1. Flowchart of the computational framework.

All CEL files were preprocessed using R package *mas*5 to obtain raw gene expression values. All raw gene expressions were then log transformed and normalized using quantile normalization.

2.1 Feature selection and ensembles

We used five base feature selection methods to select candidate informative biomarkers, including Information Gain [2], Information Gain Ratio [2], Relief [3], Symmetrical Uncertainty [2], and mRMR [4]. 1) Information Gain: Given a training set Z, the Information Gain method attempts to find a feature set X such that the information gain of X with respect to class C in the training set Z is maximized; 2) Information Gain Ratio: Given a feature set X, Information Gain Ratio attempts to mitigate Information Gains bias by considering the weight of intrinsic information; 3) Symmetrical Uncertainty: The Symmetrical Uncertainty (SU) method attempts to find a feature set containing features that are highly correlated with the class, in the meantime, are uncorrelated to each other; 4) Relief: The key idea of Relief is to estimate the quality of features according to how well their values distinguish between instances that are close to each other. 5) mRMR (Maximal Relevance and Minimal Redundancy): The mRMR attempts to find a feature set that maximizes statistical dependency (maximal relevance) between feature set and target class C, in the meantime, minimizes the redundancy among selected features. After selecting five feature sets, each by a single base feature selection method, different ensemble feature sets were obtained by aggregating five base individual feature sets into a series of ensemble feature sets, including 1) union of all five base individual feature sets (Union), 2) feature set containing features that are shared by at least two of five individual base feature sets (At Least 2), 3) feature set containing features that are shared by at least three of five individual base feature sets (At Least 3), 4) feature set containing features that are shared by at least four of five individual base feature sets (At Least 4), and 5) feature set containing features that are shared by all five individual base feature

sets (At Least 5). Finally, the ensemble feature sets were evaluated using the performance of ensemble classifiers measured by the area under the curve (AUC) of a 5-fold cross validation. The AUC is calculated from receiver operating characteristic (ROC) curve, which is a plot of the true positive rate (TPR) against false positive rate (FPR).

2.2 Classification and ensembles

In order to evaluate the effectiveness of an ensemble feature set, we used three different classifiers: Naive Bayes [5], Random Forest [6], and SVM [7], on training data set by 5-fold cross validation. The ensemble result from all classifiers was chosen for predicting the DILI class of each compound. 1) Naive Bayes Classifier: Naive Bayes (NB) is a probabilistic classifier. It is obtained by using Bayes' rule with a strong independence assumption, i. e., features are independent of each other given target classes. Despite the independence assumption, Naive Bayes has been shown to have very good classification performance in practice; 2) Random Forest (RF) Classifier: The algorithm considers an ensemble of unpruned classification trees which offers low bias as well as low variance. Random forest algorithm performs as a good classifier even when the number of variables outnumber the number of observations. In Random forest algorithm, overfitting is minimized. For this project, we set estimators parameter to 100. 3) Support Vector Machine (SVM) Classifier: SVM is a margin classifier that constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification. A good separation defines a boundary that maximizes the margin between data samples in different classes, therefore leading to good generalization properties. A key factor in SVM is to use kernels to construct a decision boundary. We used a linear kernel in this project.

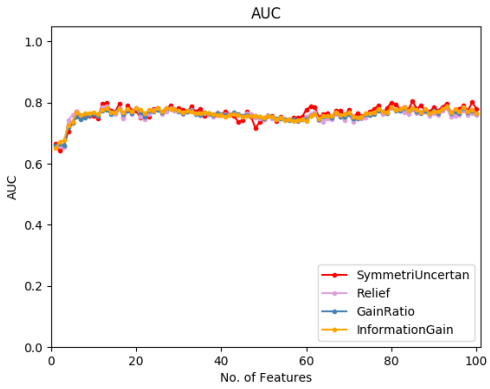

Figure 2. AUCs with respect to different number of selected features.

The final prediction is an ensemble of all classification results on each of the ensemble feature sets as obtained in Section 2.1. The overall flowchart of our computational framework is shown in Figure 1. We have implemented our pipeline as a web server. The alpha version is available at: `http://m2roc.biocoms.org/`. The final version of this online tool will be available no later than the CAMDA meeting time.

3 Results

We first explored the effectiveness of our ensemble framework on each dataset of the two cell lines respectively. that is, prediction of human clinical DILI results from

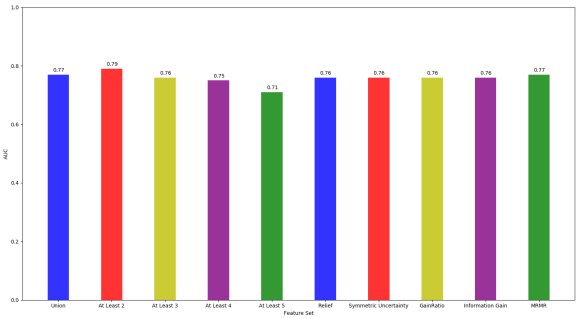

Figure 3. AUCs with respect to different selected feature sets.

each cell-line responses independently. We then attempted to identify and interpret differences in cell-line responses across cell-line types, that is, using the biomarkers identified in one cell line to predict the DILI label in the other cell line. For either task, we need first determine how many features should be used to achieve the best performance. Figure 2 demonstrates ROC curves for the classification on different number of features obtained using different feature selection methods. The ROC curves are averaged results using 5-fold cross validation. Based on the information as shown in Figure 2, the AUCs achieve the first peak when 19 features are used and become flat after it, then begin to decline when feature number is 28. In order to keep a balance between the feature diversity and the feature stability, we set 28 as a cutoff for selecting an appropriate feature set.

3.1 Classification using different feature set

Due to the space limit, in this extended abstract, we only included the results using MCF7 cell line data. The results of the cell line PC3 and the cross cell line are quite similar to those presented here, and will be included in the full paper.

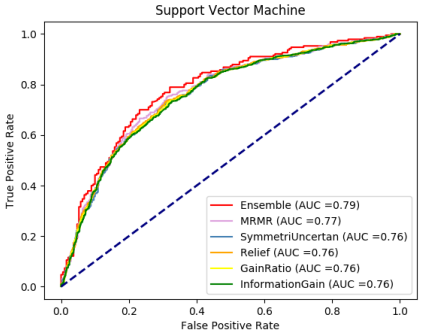

Figure 4. ROCs with respect to selected feature sets.

For MCF7 cell line CMap data, we first compared the ensemble classification performance using the different ensemble feature sets as described in Section 2.1, including "Union", "At Least 2", "At Least 3", "At Least 4", and "At Least 5", along with features selected by each single base feature selection method, that is, features without ensemble. The bar chart graph representing the AUCs of all feature sets is shown in Figure 3. As can be seen from Figure 3, the common features shared by all five base feature selection methods has the worst perfor-

mance (AUC = 0.71), and the ensemble feature set "At Least 2" has the best performance (AUC = 0.79), which is also better than any single base feature selection method. The detailed ROC of "At Least 2" ensemble feature set compared to all five base feature selection methods is shown in Figure 4.

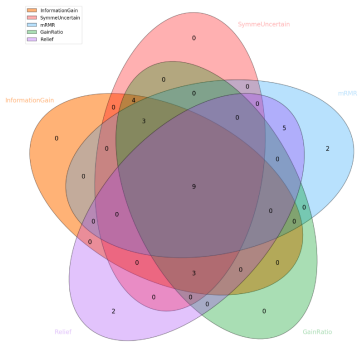

Figure 5. Venn diagram showing the relationship among different feature sets.

Although it is still in alpha version, the online tool implemented our computational framework already has functionalities that can be used to generated above mentioned figures. Scientists in drug development may be more interested in obtaining detailed information about different feature sets. Our online tool can generate a Venn diagram to show the relationship among different feature sets (Figure 5 is an example) and such relationship can be downloaded as a text file. Another component to visualize the network of features is an ongoing task and will be integrated into the online tool before the CAMDA meeting.

References

1. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, et al. (2006) The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. Science 313: 1929–1935.

2. Elomaa T, Rousu J (1999) General and efficient multisplitting of numerical attributes. Mach Learn 36: 201–244.

3. Robnik-Sikonja M, Kononenko I (2003) Theoretical and empirical analysis of relieff and rrelieff. Machine Learning 53: 23–69.

4. Ding CHQ, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. J Bioinformatics and Computational Biology 3: 185–206.

5. Rish I (2001) An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. IBM New York, volume 3, pp. 41–46.

6. Breiman L (2001) Random forests. Machine learning 45: 5–32.

7. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Machine Learning 46: 389–422.