A Review of Predicting *in vitro* Drug Sensitivity using Random Forests

Samuel Jackson, University of Aberystwyth

***Abstract*—This paper provides a review of Riddick et al. [?] and their work on predicting drug sensitivity using random forests. The purpose of this paper is to identify the main issues associated with drug sensitivity prediction and summarise the methods used by the authors. An alternative machine learning algorithm (BART-BMA [?]) is proposed with the potential benefit of providing credible intervals for drug sensitivity predictions. Finally, justification of the choice of alternate method and future directions of research are discussed.**

I. Introduction

Riddick et al. [?] propose a novel method for predicting drug sensitivity from a panel of human cell lines. They propose a machine learning approach to drug sensitivity prediction using Random Forests (RFs). Their method utilises several important advantages of RFs for feature selection and model training, namely variable importance and proximity matrices.

The authors evaluate their methodology on two different test datasets: firstly the sensitivity of two cancer drugs are predicted from a collection of human cell lines from the NCI-60 gene expression dataset. Secondly, sensitivity values for 40 FDA drugs are predicted using the same NCI-60 gene expression dataset. The authors show that their method correctly identifies the top and bottom 10 most sensitive drugs.

In this paper an alternative ensemble method called BART-BMA is proposed that shares many of the benefits offered by RFs but which could offer additional benefits over the original author's methodology, including quantifiable uncertainty in sensitivity predictions and potentially better accuracy.

The rest of the paper proceeds as follows: section II outlines the methodology used by Riddick et al. in more detail. Section III discusses their approach and outlines strengths and weaknesses. Section IV proposes the alternative methodology and provides a discussion and justification.

II. Summary of Methodology

The authors of [?] propose a method for using the pattern of gene expression for a variety of cell lines as a means to predict the sensitivity of a particular drug. For their training data the authors use the NCI-60 datasets which contains a panel of 60 cell lines derived from cancerous cells across 9 different types of human tissue. The IC_{50} measure, which is defined as the amount of a drug required for 50% inhibition of a biological process [?], was used as a response variable for a single drug.

This dataset is challenging to work with. Firstly, examining the data used by the authors shows that the response variable is missing a number of entries. Secondly and more importantly, the dataset is a classic example of what is referred to as "large p small n". This means that the number of features (p) is much larger than the number of samples (n). Such high dimensionality can cause issues for many machine learning algorithms due to the "curse of dimensionality" [?]. The Hughes effect [?] states that predictive power decreases in proportion to the increase in dimensions, given a fixed number of training samples.

Riddick et al. chose to use RFs as the main component of their methodology. A major strength of this learning algorithm relative to their dataset is that RFs can be used to measure how important a variable is likely to be for prediction. The authors of [?] first train a RF on the full set of gene expression data (consisting of 16,644 features) using a large number of trees (25,000). This first RF would likely exhibit poor generalisation, but it does it allow the RF to measure the importance of each variable as a by product of training. This measure is calculated by taking the average difference between the out-of-bag (OOB) error before and after permuting the values of the j^{th} feature. From this measure of variable importance a subset of the best features can be chosen (Riddick et al. select those which are $2\sigma > \mu$, typically 100-500 probesets).

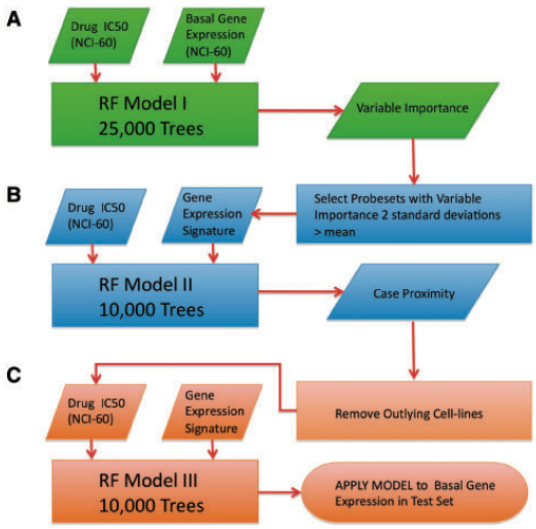

Fig. 1. Diagram from [?] showing the three stages of building their random forests model. A) feature selection using variable importance, B) outlier removal using proximity matrices, C) predicting IC_{50} values.

Using this subset of features, they train another RF with a reduced number of trees (10,000). This second model should be more accurate than the first as well as being faster to train and execute. However, before using this tree for drug

sensitivity regression, outlying cell lines in the model are removed using the proximity matrix of features generated from the second model. The proximity matrix of a RF is defined as the number of times two cases are assigned to the same terminal node. Cell lines which showed high correlation (using Person's correlation coefficient) were retained in the IC_{50} and NCI-60 datasets. They note that they use the Bonferroni correction method [**?**] to correct for multiple hypothesis testing. A regression is obviously sensitive to outliers and with such a low ratio of samples (60) to features it is important to ensure that each case is not simply noise.

Finally, Riddick et al. retrain their second model with the outlying cases removed and use this to infer drug sensitivity from unseen gene expression samples. Figure 1 gives a high level overview of their methodology.

III. Discussion of the Problem

As mentioned in the previously, predicting drug sensitivity from a large number of gene expression signatures is challenging task. The high dimensionality of the dataset immediately rules out many machine learning techniques which will not work well in this situation. A form of dimensionality reduction is almost a prerequisite for data of this size. This problem is exacerbated by the small number of samples (only 60, even less with outliers removed).

Another major challenge in applying machine learning to this problem is the essence of accuracy. In general, the ultimate goal of any application of machine learning is to achieve high generalisation and low error on previously unseen data. However due to the application domain (drug sensitivity) and problem nature (regression) it is essential that a high degree of prediction accuracy is achieved. In this problem the authors are attempting to infer not only if a drug is sensitive but also the degree of sensitivity. Regression values which are falsely indicative of sensitivity must be minimised to reduce the misapplication of a drug. Therefore any estimator applied to this problem should have both an extremely high degree of accuracy and precision.

In a regression problem like this, an important aspect is the presence of outliers in the data. This is especially an issue when there are only a small number of data points from which to generalise from. Even a small number of outliers has the potential to drastically move the regression line from a good fit to a bad one. Here the authors have attempted to remove outliers using their proximity measurement technique. Any machine learning technique applied to this problem should provide some robustness to outliers.

A problem specific consideration is that fact that any approach applied to this problem needs to take into account the interaction effects of combinations of different genes. RFs are able to account for this because they composed of decision trees which are combinations of splits of variables.

One issue that the authors of the paper neglect with their approach and which the alternative approach presented in this paper offers is quantification of the uncertainty in an estimate. While making a prediction about a specific drug response is useful, it would also be useful to have information about confidence we have in the estimate. There is no point in predicting that a drug is highly sensitive when the margins of error are so large as to render the prediction useless.

Additional practical considerations applicable to any machine learning problem are things such as the speed of training, the speed prediction, and the number and complexity of parameters. For the specific problem of drug sensitivity, algorithmic training and prediction speed are probably less important that in other problems, such as online learning problems. It probably does not matter if a hypothetical machine learning approach takes several hours to train or to yield a prediction provided that the prediction is accurate. The number and complexity of parameters is probably more relevant to this problem. RFs generally require little parameter tuning. Having too many "knobs to turn" can easily lead to overfitting and increased tuning times.

One final issue for consideration is the interpretation of predictions. RFs are often said to be black box approaches which give good predictions but offer little explanation for the prediction. Variable importance can be used to gain an insight, but this is still based on the predictive ability of each feature rather than an interpretable explanation.

IV. An Alternative Method

This section discusses an alternative machine learning algorithm that could be applied to the problem described in section II. This method offers some potential advantages over RFs which will be discussed in IV-D as well as outlining some drawbacks. Definitions in the following two subsections are liberally borrowed from [**?**] and [**?**].

A. Outline of BART

Bayesian Additive Regression Trees (BART) [**?**] are a tree based ensemble method which shares many similarities with RFs. BART is a sum of trees model, where each individual tree is a weak learner that is fit to the residuals of the previous trees in the current iteration. Therefore BART is an additive method rather than taking an average as in RFs. BART is a Bayesian method which generates the full posterior distribution model parameters. BART is primarily a regression algorithm, but can be adapted in to a classifier. Because BART offers access to the full posterior distribution it is able to yield credible intervals for a prediction, directly quantifying our uncertainty in the prediction.

The BART model can be defined as follows. Let $x_k \in X$ be the k^{th} observation of p features. The basic BART model is

$$Y_k = \sum_{j=1}^{m} g(x_i; T_j, M_j) + \epsilon_k \tag{1}$$

where g is a Bayesian CART model [**?**], m is the total number of trees, T_j is single decision tree which has terminal node parameters M_j, and ϵ is a Gaussian with zero mean and σ^2 of the residual variance. The BART model is fitted using a back-fitting Gibbs sampler using draws from the joint posterior of all trees, terminal node parameters, and σ^2 given the data.

The priors for BART are formulated so as to regularise the fit to favour small trees.

BART has been shown to be competitive with RFs in terms of prediction accuracy. However, it does come with some drawbacks which make it unsuitable for direct application to the drug sensitivity problem. BART is known to require large amounts of memory when working with higher dimensional data. This is neatly summarised in [**?**] as being for two reasons:

1) Using a uniform prior to choose variable splits results in a high rejection rate when there are a large number of dimensions. Intuitively if there are many dimensions to choose from, it should be possible to do better than random selection.
2) BART becomes memory hungry for large datasets because each tree at each iteration of each MCMC chain must be stored so it can be used for future predictions.

B. Outline of BART-BMA

BART with Bayesian Model Averaging (BART-BMA) [**?**] proposes some modifications to address these issues in an attempt to offer a compromise between both RFs and BART. Like BART, BART-BMA is an ensemble of Bayesian CART models which are summed together to produce a strong learner. The modifications are as follows: Firstly BART-BMA performs a greedy search for the most predictive splits. Secondly, BART-BMA uses Bayesian model averaging rather than MCMC to average over the posterior of many models.

BART-BMA greedily searches for predictive splits using one of two options. The first option uses a change point detection algorithm called Pruned Exact Linear Time (PELT) which attempts to minimise a cost function to find optimal splits. The second option is to use a grid search where each variable is split into n points and each point is used as a potential split point. The best split rules are then chosen based on the residual squared error.

BART-BMA also differs in the way that it calculates the sum of trees likelihood. BART computes the likelihood for each individual tree in the model. In contrast BART-BMA directly specifies the sum of trees in its model. For each tree model T_{hil} (with terminal node i in tree l using split h) the posterior probability of the sum of trees containing T_{hil} can be approximated as ST_{hil} using the Bayesian Information Criterion (BIC)

$$BIC = -2(log(p(Y|X, ST_{hil})) + log(p(ST_{hil}))) + Blog(n) \tag{2}$$

where $p(Y|X, ST_{hil})$ is the likelihood of the sum of trees model containing T_{hil} and $p(ST_{hil})$ is the prior for the sum of trees model. B is the number of model parameters. In a large sample size BART-BMA should produce the model which is a posteriori more probable.

Finally, in order to focus the search on only the most likely trees, a greedy and efficient version of Bayesian Model Averaging (BMA) called Occam's Window is used. Using Occam's Window, only the best models are averaged over according to

$$log(BIC_l) - argmin_l(log(BIC)) \leq log(o) \tag{3}$$

where o is a threshold parameter determining the size of the window. Any model which falls outside of Occam's Window is discarded. This is what allows BART-BMA to keep only the models from which there is a reasonably high support from the data.

Once a sum of trees model has been produced the predicted response is calculated from a weighted average of the response from the models. Each model is weighted proportionally to its BIC_l value according to equation 4.

$$\begin{aligned} w_l &= exp(-0.5BIC_l - v) \\ v &= max_l(-0.5BIC_l) \end{aligned} \tag{4}$$

C. Application of Method to the Problem

Being an ensemble tree based method, BART-BMA shares many similarities with RFs. As such the application of BART-BMA could be achieved in a similar manner to random forests. The variable importance measure offer shows good agreement to the one offered by RFs, so the initial feature selection process could be implemented identically. Likewise, the final prediction stage on the subset of the data could be performed in same way as in [**?**].

The real difficulty with replacing RFs in this methodology with BART-BMA is the middle step which uses the proximity matrices derived from a RF to identify outliers and remove them. One option would be to use the same approach as the authors and train a RF and extract the case proximities before subsequently training a BART-BMA model on the final (outlier removed) data. It might be however that BART-BMA is robust enough to outliers in the data so as to not require the second step in the original methodology. If this was not the case an alternative approach might be to adjust the probability model to account for outliers [**?**].

The original implementation of BART would most likely be able to handle the small subset of features in the third training stage in [**?**]. Therefore, for a fast implementation RFs could be used to implement first two components of the system with either BART or BART-BMA just being used in the final prediction stage. Alternatively BART-BMA could be used for the first stage to counter the high dimensionality of the first input dataset and either BART or BART-BMA could be applied for subsequent stages. Further exploratory analysis is required to comment further on the combination of RFs, BART, and BART-BMA that would give best results. It is difficult to say which combination will work best without first applying the model to some actual data.

D. Discussion of Alternative Method

The first two subsections have outlined BART and BART-BMA. This subsection discusses the implications of this choice of machine learning approach in terms of the problem described in II.

As suggested in section III one major downside to RFs is that they provide no confidence in the estimates produced.

A Bayesian approach such as BART-BMA offers credible intervals for the predictions it makes. This is likely to be useful when making predictions about drug sensitivity. Quantifiable uncertainty can help us make better choices and to have greater confidence in our predictions. It also provides an additional dimension to measure how well the model performs.

The second reason for choosing BART-BMA is that it has been specifically designed for problems with higher dimensional data (HDD). The major weakness with BART and strength of RFs is their performance on HDD. BART-BMA has been developed as an attempt to bridge the gap between the two allowing a Bayesian ensemble algorithm to run on a standard laptop with HDD. Performance in terms of speed shows that RFs would still beat BART-BMA, but the algorithm is much faster and more memory efficient than BART. This would therefore likely make BART-BMA a feasible technique for the drug sensitivity dataset. The authors of BART-BMA also used BART-BMA to make predictions on two different biological datasets with near identical accuracy to RFs. This suggests that BART-BMA may have promising applications to datasets such as the one presented in [**?**].

Furthermore, while BART-BMA is still a black box method, the authors note that BART-BMA tends to choose "shallower and more interpretable" [**?**] trees than RFs. This could have the duel benefit of reducing overfitting while being easier to interpret/visualise. Additionally variable importance scores generated from BART-BMA are in good agreement with those generated from RFs.

However, BART-BMA isn't without a weakness. Firstly, it is a fairly new technique and while promising has yet to be experimented with to the same extent as RFs. Time is still needed to ensure the viability of such a new approach. Secondly, BART-BMA does bring with it an assumption that the most probable models are the ones with the best predictive power. In particular, because Occam's Window necessarily discards some choices of model there is obviously the potential that models which are ignored could have still provided useful contributions to the final prediction.

Finally, it should be noted that there have been recent efforts to derive confidence intervals for the predictions made by RFs [**?**]. This would render the main benefit of this approach fairly redundant. However there's an argument to be made that approaching the problem directly from a Bayesian point for view rather than resorting to frequentist tricks is a more natural approach to quantifying uncertainty.

V. Future Directions and Work

There is an argument to be made that uncertainty in the predictions that a model makes are an essential factor if automated predictions for drug sensitivity are going to be feasible in the wild. Additionally, the interpretability of a model is another important factor. A major downside of both methods outlined in this paper is that, unlike decision trees, they offer a prediction but are not particularly informative about why that prediction was made.

Due to the high dimensionality of the data, ensemble based methods are a good choice for this problem. Another alternative to this methodology which would be worth experimenting might be Gradient Boosting [**?**] which shares many of the same properties with RFs and BART-BMA.

As the authors of the original paper noted outlier removal remains an essential challenge when working with a drug sensitivity dataset. Automated removal is obviously a desirable feature, however using a method that is robust to outliers in the first place would most likely reduce the training time required. Therefore an interesting direction for future research would be to explore modifying the techniques mentioned to improve robustness.

Finally, another direction worth exploring, noted by the authors of BART-BMA would be the modify the algorithm to handle missing data. While this is not directly applicable to this dataset it would be useful for biological data in general.

References

[1] G. Riddick, H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, and H. A. Fine, "Predicting in vitro drug sensitivity using random forests," *Bioinformatics*, vol. 27, no. 2, pp. 220–224, 2011.

[2] B. Hernández, A. E. Raftery, S. R. Pennington, and A. C. Parnell, "Bayesian additive regression trees using bayesian model averaging," *arXiv preprint arXiv:1507.00181*, 2015.

[3] "FDA Website - IC50 versus EC50," http://www.fda.gov/ohrms/dockets/ac/00/slides/3621s1d/sld036.htm, accessed: 2016-03-05.

[4] R. E. Bellman and H. A. Osborn, "Dynamic programming and the variation of green's functions." 1957.

[5] G. P. Hughes, "On the mean accuracy of statistical pattern recognizers," *Information Theory, IEEE Transactions on*, vol. 14, no. 1, pp. 55–63, 1968.

[6] H. Abdi, "The bonferonni and šidák corrections for multiple comparisons," *Encyclopedia of measurement and statistics*, vol. 3, pp. 103–107, 2007.

[7] H. A. Chipman, E. I. George, and R. E. McCulloch, "Bart: Bayesian additive regression trees," *The Annals of Applied Statistics*, pp. 266–298, 2010.

[8] ——, "Bayesian cart model search," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 935–948, 1998.

[9] H. Chipman, E. George, and R. McCulloch, "Bayesian Additive Regression Trees (BART)," http://spectral.kutabiri.com/static/bart_presentation.pdf, accessed: 2016-03-06.

[10] S. Wager, T. Hastie, and B. Efron, "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1625–1651, 2014.

[11] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.