18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

A comparison of Bayesian classification trees and random forest to identify classifiers for childhood leukaemia

O'Leary, R.A. [1], **R.W. Francis**[1], **K.W. Carter**[1], **M. J. Firth**[1], **U.R. Kees**[1] **and N. H. de Klerk**[1]

[1] *Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, Perth, Australia*
Email: roleary@ichr.uwa.edu.au

Abstract: Recently, microarrays technologies have been extensively used to distinguish gene expression in acute lymphoblastic leukaemia (ALL) (e.g. Pui et al., 2004; Hoffmann et al., 2008). ALL is the most common type of leukaemia diagnosed in children, with an incidence rate of about 4 per 100,000 per year (Pizzo and Poplack, 2001; Milne et al., 2008). There are six main subtypes of leukaemia, one of which is T-cell acute lymphoblastic leukaemia (T-ALL) which generally has lower cure rates than other forms of ALL. Ribonucleic acid (RNA) samples from each patient can be put onto microarrays to provide gene expression levels for around 20 thousand genes (depending on which microarray chip is used). One of the challenges with microarray analysis in leukaemia research is identifying the smallest possible set of genes that predict relapse with the highest predictive performance.

Currently, one approach used to identify important differentially expressed genes is Random Forest (RF) (e.g. Hoffmann, 2006; Díaz-Uriarte and Alvarez de Andrés, 2006). RF is a classifier that consists of an ensemble of classification trees, and yields the average class for each Y observation (each patient). Díaz-Uriarte and Alvarez de Andrés (2006) identified the characteristics that make RF ideal for microarray data, these include: RF can handle more variables than observations (large p small n problems); RF can be applied to binary and multi-class problems; RF has good predictive performance for datasets containing a large number of noise variables and does not overfit; RF can use both categorical and continuous predictors and investigates interactions; the results from RF are unaltered by monotone transformations of the variables; a free R library exists that performs RF; RF provides measures of variable importance and for the most part one does not have to fine-tune parameters to obtain good predictive performance.

This paper describes an alternative approach to identifying a gene classifier for predicting relapse in ALL. Bayesian approaches to classification and regression trees (BCART) were proposed by Chipman et al. (1998), Denison et al. (1998) and Buntine (1992). BCART identifies "good" trees using a stochastic search algorithm that applies a reversible jump Markov chain Monte Carlo method. The set of best trees are selected that have the highest prediction accuracy (O'Leary et al. 2008). Fan and Gray (2005) gave BCART an A+ for interpretability and B+ for prediction. To date, BCART has been largely based on "non-informative", usually conjugate priors. Moreover, there are only a few real-world applications of BCART (Lamon & Stow, 2004; Partridge et al., 2006; Schetinin et al., 2007). This statistical approach has not been applied to large p small n problems (to the author's knowledge).

Here we compare RF and BCART for predicting relapse in three ALL datasets, using gene expression values as the covariates. In all three datasets, the best tree identified from BCART had better accuracy and in particular better prediction of relapse (higher sensitivity) than RF. BCART also had better performance than RF in identifying important genes that predicts whether a patient will relapse.

Keywords: *Bayesian classification and regression trees; expert elicitation informative priors.*

1. INTRODUCTION

Recently, microarrays have been used extensively to distinguish gene expression in acute lymphoblastic leukaemia (ALL) (e.g. Pui et al., 2004; Hoffmann et al., 2008). Despite dramatically increased cure rates, up to 25% of T-cell acute lymphoblastic leukaemia (T-ALL) patients still relapse. Microarrays provide a ‘snap-shot’ view of the expression level of tens of thousands of genes concurrently (Quackenbush, 2002). One statistical approach used to identify important genes relating to whether a patient relapses after therapy is random forest (RF) (e.g. Hoffmann, 2006; Díaz-Uriarte and Alvarez de Andrés, 2006). This statistical method is used to identify the smallest possible set of genes that achieve “good” predictive performance from a large number (typically around 20,000 genes) that could be used for diagnostic purposes in clinical practice.

Classification and regression trees (CART), proposed by Breiman et al. (1984), are a popular statistical methodology because they are easy to interpret and have good predictive power (Breiman, 2001b; De’ath & Fabricius, 2000). Therefore, Breiman (2001b) gave it a rating of *A*+ on interpretability and *B* for prediction. A comparable method with respect to predictive capacity is a RF (Breiman, 2001a), but the interpretability of this method was rated much lower (*F*) by Breiman (2001b). RF results in many trees, and provides the average prediction for each observation (patient) in the response (dependent) variable. Bayesian approaches to CART (BCART) were proposed by Chipman et al. (1998), Denison et al. (1998) and Buntine (1992). BCART has been rated an A+ for interpretability and B+ for prediction (Fan and Gray, 2005). Unlike RF, the output from BCART is a set of good trees and a figure of the best tree, which is achieved through several accuracy measures (O’Leary 2008b). An additional benefit of BCART is the ability to incorporate expert opinion or historical data into the prior model, and combine this information with the observed data to produce the trees (posterior distribution). To date, there have been very few real applications of BCART published (Lamon & Stow, 2004; Partridge et al., 2006; Schetinin et al., 2007). One reason is that BCART is only available in non-user friendly software.

This paper compares RF and BCART techniques for identifying genes that predict relapse in three different ALL datasets. We discuss the similarities and dissimilarities of these two approaches, and determine whether BCART is a suitable approach for variable selection problems (large p small n) and for the identification of genes that predict relapse in ALL patients.

2. CASE STUDY

2.1. Dataset 1 (DS1)

Patient Specimens: Bone marrow specimens from 50 T-ALL patients were obtained at diagnosis from Children’s Oncology Group (CCG/COG); 22 patients went on to relapse (DR) and 28 achieved complete clinical remission (DN) after treatment on the COG-1961 chemotherapy protocol. Patients were diagnosed and treated at the COG institutions. Informed consent was acquired from either parents, patients or both.

Microarray Experiments: For each patient (DR and DN specimens) the gene expression was measured using Affymetrix U133-Plus 2.0 oligonucleotide microarrays. This array provides data for 54,675 probes sets, which represent expression for most of the genes in the human genome (about 20,000 genes). The gene expression values were normalised using Robust Multi-Array (RMA) methodology. For further details on the DS1 microarray experiments see Hoffman et al. (2008).

2.2. Dataset 2 (DS2)

Patient Specimens: We examined the 44 T-ALL patient data of Winter et al. (2008); 14 patients went on to relapse (DR) and 30 achieved complete clinical remission (DN) after treatment. There were also 6 patients who failed to achieve remission (induction failure), these patients were excluded from this analysis.

Microarray Experiments: Gene expression was measured using Affymetrix U133-Plus 2.0 oligonucleotide microarrays, as above. Similar to DS1, the gene expression values were normalised using RMA methodology. .For further details see Winter et al. (2008).

2.3. Dataset 3 (DS3)

Patient Specimens: We accessed 132 ALL patient data from Ross et al. (2003). There were 14 patients who later relapsed (DR) and 63 who achieved remission (DN); 14 T-ALL patients and 41 patients with unknown status were excluded.

Microarray Experiments: Gene expression was measured using Affymetrix U133-A chip oligonucleotide microarrays, which target 22,284 probe sets. Like in DS1, the gene expression values were normalised using RMA methodology. .See Ross et al. (2003) for more details.

3. STATISTICAL METHODOLOGY

3.1. Random Forest (RF)

Random forest (RF) is a supervised decision-tree based algorithm (Breiman, 2001a). It is a classification method that consists of many trees $h(x,\Theta_r)$, where *x* is input vector, Θ_r are independent identically distributed random vectors and each rth tree classifies each observation (where *r* = 1,…, *R*). Final class label for each observation is the average class. Specifically, given a collection of classifiers $h_1(x), h_2(x), \ldots h_R(x)$, and from the distribution of random vector Y,X the training set is randomly drawn, the margin function is $mg(X,Y) = av_r I(h_r(X)=Y) - \underset{s\neq Y}{max}\ av_r I(h_r(X)=s).$

Each tree is constructed as follows:

1. Let the number of observations in the training dataset be N, and the number of variables be M.
2. Sample N at random (bootstrap sampling with replacement), from original data. This subset will be used for growing the tree. The samples not used to construct tree are called *out-of-bag* samples,
3. For each node of tree, randomly select m variables (where m<M), assess best split based on these m variables in subset set.
4. Each tree is fully grown and not pruned.
5. The outcome or final prediction is the average of out-of bag estimators over all Bootstrap samples.

3.2. Bayesian Classification and Regression Trees (BCART)

3.2.1. Notation

We explain the mathematical notation for trees by referring to the tree in Figure 1. The root node (*k*=1) is at the top, and the tree progressively branches (at *K*=4 nodes altogether) to reach the four terminal nodes hanging at the ends of the branches on the tree. This tree shows three variables of the 54,675 probe sets from DS1 (J=54,675) fit in the model.

Consider the first branch at node $S_1=1$. The variable that defines the first split V_1 is RNF17 (ring finger protein 17 (220270_at)). The splitting rule R_1 splits the data into two groups, $y_i \in R_1 \le 5; y_i \notin R_1$ defining the left and right branches respectively. At the right branch, we have a terminal node, in which one non-relapse and five relapse patients are predicted as relapse, with estimated probability of relapse DR_4 = 5/6 = 0.83.

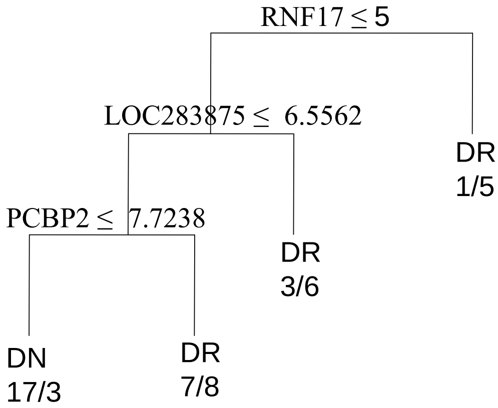

Figure 1. Best BCART tree for DS1 where RNF17 is ring finger protein 17 (220270_at), LOC288375 is hypothetical protein LOC283875 (1564122_at), PCBP2 is poly(rC) binding protein 2 (213263_s_at).

3.2.2. Model

The joint distribution is $p(K,\theta_k,y) = p(K)p(\theta_k \mid K)p(y \mid K,\theta_k)$, where $p(K)$ is the distribution for the tree size (number of terminal nodes K), $p(\theta_k \mid K)$ is the distribution for parameter set $\theta_k = \{R_k, S_k, V_k\}$ given the tree size K, and $p(y \mid K,\theta_k)$ is the likelihood. For classification trees, observations are assumed to have a multinominal distribution, so the likelihood is

$$p(y \mid K,\theta_k) \propto \prod_{k=1}^{K}\prod_{j=1}^{N}\left(p_{kj}\right)^{m_{kj}}$$

Here m_{kj} is the number of data points at terminal node k, which are classified into category j, and p_{kj} is the corresponding probability. A conjugate Dirichlet prior can be adopted for p_{kj}. In the absence of other information, a uniform distribution can be used to define a non-informative prior so that $\pi(p_{k1},\ldots,p_{kJ}) =$ $\mathrm{Dir}_{J\text{-}1}(p_{k1},\ldots,p_{kJ} \mid 1,\ldots 1)$. The prior for the model is $p(\theta_k \mid K)p(K) = p(S_k \mid K)p(V_k \mid S_k,K)p(R_k \mid V_k,S_k,K)p(K)$.

Dirichlet priors may be allocated to several elements of the prior: selecting possible splitting nodes via $p(S_k \mid ,K) = \mathrm{Dir}(S_k \mid \alpha_{S_1},\ldots\alpha_{S_k})$; specifying important variables V_k that determine the split at node S_k via $p(V_k \mid S_k,K) = \mathrm{Dir}(V_k \mid \alpha_{V_1},\ldots\alpha_{V_k})$; defining splitting rules R_k for variable V_k at node S_k via $p(R_k \mid V_k,S_k,K) = \mathrm{Dir}(R_k \mid \alpha_{R_1},\ldots\alpha_{R_k})$. The prior p(K) is assumed to be a truncated Poisson (with parameter λ) $p(K) = \frac{\lambda^k}{(e^{\lambda}-1)k!}$. This prior imposes a left limit of k > 0 because minimum model contains one terminal node. When no information is available for a particular prior, then non-informative uniform priors are used, with $p(V_k \mid S_k,K) = Dir(V_k \mid 1,\ldots,1)$, $p(R_k \mid V_k,S_k,K) = Dir(R_k \mid 1,\ldots,1)$ and $p(S_k \mid K) = Dir(S_k \mid 1,\ldots,1)$. A weakly informative prior for the size of the tree, following Denison et al. (1998), by setting λ=10 in the prior p(K).

Previous computational approaches for BCART adopted stochastic search algorithms to efficiently explore part of the parameter space (Chipman et al. 1998; Denison et al. 1998). We apply the algorithm of O’Leary (2008b), which simulates the joint posterior distribution using reversible jump Markov chain Monte Carlo. The stopping criterion and identification of good trees applied by O’Leary (2008b) is achieved through several accuracy measures (Fielding and Bell, 1997). The accuracy measures chosen are the overall accuracy (number of correct classification of both relapse and non-relapses), the correct classification of relapses (sensitivity) and correct classification of non-relapses (specificity).

Table **1**. Accuracy measures of random forest and BCART for DS1; three measures are overall accuracy (number of correct classification of both relapse and non-relapses), sensitivity (number relapse patients correctly classified) and specificity (number of non-relapses correctly classified).

	Accuracy	Sensitivity	Specificity
Random Forest	0.62	0.32	0.86
BCART	0.72	0.86	0.61

Table **2**. Accuracy measures of random forest and BCART for DS2; accuracy, sensitivity and specificity as defined in Table 1.

	Accuracy	Sensitivity	Specificity
Random Forest	0.64	0	0.93
BCART	0.68	0.93	0.56

Table **3**. Accuracy measures of random forest and BCART for DS3; accuracy, sensitivity and specificity as defined in Table 1.

	Accuracy	Sensitivity	Specificity
Random Forest	0.82	0	1
BCART	0.76	0.88	0.73

4. RESULTS

For all three patient datasets, RF and BCART were applied to predicting relapse, where normalised gene expression values of the probe sets are the predictor variables. The accuracy measures of RF and BCART for DS1, DS2 and DS3 are shown in Tables 1, 2 and 3 respectively, the measures included are overall accuracy, sensitivity (number of relapse patients correctly classified) and specificity (number of non-relapses correctly classified).. For DS1 and DS2, the accuracy of the best tree identified by BCART is higher than RF. For all three datasets, BCART produced higher sensitivity but lower specificity compared to RF. In DS2 and DS3, RF predicted all relapse patients as non-relapse. For these clinical datasets, it is more important that the relapse patients are predicted correctly (i.e. sensitivity is more important than specificity).

Figure 2 shows the percentage of iterations for each tree size sampled by the two approaches for DS1. A similar pattern occurs for the other two datasets (data not shown). For both approaches the sample tree size ranges from one to eight. A similar percent of each tree size is sampled by RF and BCART, except for size three.

The best BCART tree for DS1 is displayed in Figure 1. Of the 22 relapse patients, only three are misclassified, whereas 11 of the 28 non-relapse patients are classified as relapse.

Of the 54675 variables in DS1, four variables/genes were identified as important using both methods. These genes were: DLEU7 (Deleted in lymphocytic leukemia, 7; probe set ID 1566081_at); GSK3A (glycogen synthase kinase 3 alpha; probe set ID 202210_x_at); CDNA FLJ14169 fis, clone NT2RP2002056 (probe set ID 216160_at); LOC202347 (hypothetical protein LOC202347; probe set ID 225654_at).

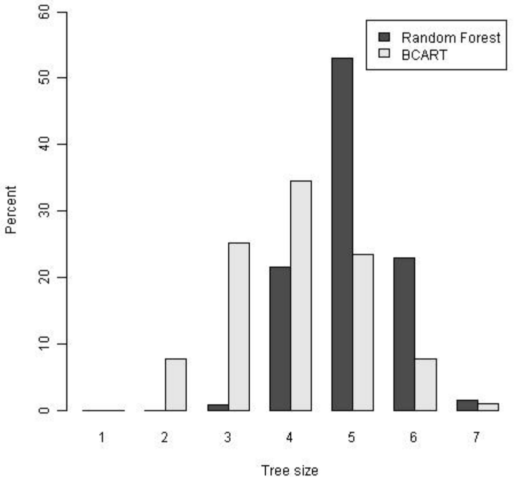

Figure 2. Percent of iterations for each tree size sampled by random forest and BCART for DS1.

5. DISCUSSION AND CONCLUSIONS

This work compared RF and BCART for predicting relapse in the three ALL datasets, using normalised gene expression values as the predictor variables. Both approaches were found to be appropriate for variable selection problems as both have the potential to identify the most important genes (for predicting relapse in cancer patients) from a large number of genes.

For all three datasets tested for this project, we found that the best tree identified from BCART had superior accuracy and in particular, better prediction of relapse, compared to RF. One reason for the substantial difference in the accuracy between the algorithms tested is that RF averages the prediction for each patient over all iterations. In contrast, BCART identifies the best tree using selection criteria appropriate to the particular study, thus the best tree is defined as the tree with highest sensitivity, specificity and accuracy. If there are a large number of noise variables or variables that are not good at predicting relapse versus non-relapse, then RF will result in more incorrect predictions. Therefore BCART may be better at identifying important genes compared with RF. Our results were in contrast to an earlier study by Díaz-Uriarte and Alvarez de Andrés (2006), who suggested that RF has good predictive performance for datasets containing large number of noise variables.

Table **4**. Advantages of RF and BCART.

	RF	BCART
Suitable for variable selection problems (large p small n)	✓	✓
Applies bagging (bootstrapping)	✓	X
Selection criteria for choosing best tree	X	✓
Final class label for each observation is average class	✓	X
Applied to binary and multi-class problems	✓	✓
Good predictive performance for datasets containing large number of noise variables	X	✓
Investigates a wider variety of tree structures with different variables, splitting rules & number of terminal nodes	✓	✓
Can use both categorical and continuous predictors	✓	✓
Investigates interactions	✓	✓
Results unaltered by monotone transformations of the variables	✓	✓
Software freely available	✓	X
Provides measures of variable importance	✓	✓
Ideal for microarray data	✓	✓

RF seems to have a prediction bias towards assignment to the largest group, in all three datasets tested there were a larger number of non-relapse patients compared to relapses. In datasets DS2 and DS3, RF predicted all relapse patients as non-relapse. Predictions produced by many statistical methods can be affected by the unequal sample sizes of binary response variables, especially for methods based on modelling the mean, since the mean will reflect the dominant value. Logistic regression, for example, under an unweighted loss function yields prediction biases towards the larger group (Hosmer and Lemeshow, 1989; Fielding and Bell, 1997). BCART does not model the mean prediction of patients overall iterations, therefore is not biased towards the larger non-relapse group.

Table 4 displays the advantages of RF versus BCART, which is based on Díaz-Uriarte and Alvarez de Andrés (2006) summary. Most importantly, both RF and BCART are ideal for microarray data, because both can handle variable selection problems. Previously BCART has not been applied to microarray data or variable selection problems (large p small n). We note that bootstrapping was not applied to BCART, whilst it was for RF. Therefore the BCART models are not validated. Future work will examine incorporating bootstrapping into BCART.

In conclusion, for the three ALL relapse versus non-relapse datasets, we found BCART is superior to RF. While the applications of these BCART methods are not yet widespread, they may shed new light into a range of research problems where other techniques have failed.

REFERENCES

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics Supplement*. 32, 496-501.

Breiman, L. (2001a). Random forests. *Machine Learning* 45, 5-32.

Breiman, L. (2001b). Statistical modelling: the two cultures. *Statistical Science* 16 , 199-231.

Breiman, L., J. H. Friedman, R. Olshen, and C. J. Stone (1984). Classification and Regression Trees. Wadsworth: Belmont, CA.

Buntine, W. (1992). Learning classification trees. *Statistics and Computing* 2, 63-73.

Chipman, H. A., E. I. George, and R. E. McCulloch (1998). Bayesian CART model search. *Journal of the American Statistical Association* 93, 935-960.

Denison, D., B. Mallick, and A. Smith (1998). A Bayesian CART algorithm. Biometrika 85, 363-377.

Díaz-Uriarte R. and S. Alvarez de Andrés (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7, 3-16.

Fan, G. and J. B. Gray (2005). Regression tree analysis using TARGET. *Journal of Computational and Graphical Statistics* 14, 206-218.

Fielding, A. H. and J. F. Bell (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24, 38-49.

Hoffmann, K., M.J. Firth, A.H. Beesley, N.H. de Klerk and U.R. Kees (2006) Translating microarray data for diagnostic testing in childhood leukaemia. *BMC Cancer* 6**,**229-240.

Hoffmann, K., M.J. Firth, A.H. Beesley, J.R. Freitas, J. Ford, S. Senanayake, N.H. de Klerk, D.L. Baker and U.R. Kees (2008) Prediction of relapse in paediatric pre-B acute lymphoblastic leukaemia using a three-gene risk index. British Journal of Haematology 140, 656-664.

Hosmer, D. W. and S. Lemeshow (1989). Applied Logistic Regression. New York, USA: Wiley.

Lamon E. C. and C. A. Stow, (2004). Bayesian methods for regional-scale eutrophication models. *Water Research* 38, 2764-2774.

Milne, E. C. L. Laurvick, N. de Klerk, L. Robertson, J. R. Thompson and C. Bower (2008). Trends in childhood acute lymphoblastic leukemia in Western Australia, 1960-2006. *International Journal of Cancer*: 122, 1130-1134.

O'Leary, R. A., Murray, J., Low Choy, and Mengersen, K (2008). Expert Elicitation for Bayesian Classification Trees. *Journal of Applied Probability & Statistics* 3, 95-106.

Partridge, D., V. Schetinin, D. Li, T. J. Coats, J. E. Fieldsend,W. J. Krzanowski, R. M. Everson and T. C. Bailey, (2006). Interpretability of Bayesian decision trees induced from trauma data. *Artificial Intelligence and Soft Computing - ICAISC 2006* 4029, 972-981.

Pizzo, P. P. and D.G. Poplack (2001). Principles and practice of paediatric oncology (4th ed.) Philadelphia, USA: Lippincott Williams & Wilkins.

Pui, C.H., Relling, M.V. & Downing, J.R. (2004) Acute lymphoblastic leukaemia. New England Journal of Medicine, 350, 1535-1548.

Ross, M.E., X. Zhou, G. Song, S. A. Shurtleff, K. Girtman, W. K. Williams, H. C. Liu, R. Mahfouz, S. C. Raimondi, N. Lenny, A. Patel, and J. R. Downing (2003). Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood 102, 2951-2959.

Schetinin, V. , J. E. Fieldsend, D. Partridge, T. J. Coats, W. J. Krzanowski, R. M. Everson, T. C. Bailey and A. Hernandez, (2007). Confident interpretation of Bayesian decision tree ensembles for clinical applications, IEEE TITB 11, 312-319.

Winter, S. S. , Z. Jiang, H. M. Khawaja, T. Griffin, M. Devidas, B. L. Asselin and R. S. Larson (2007). Identification of genomic classifiers that distinguish induction failure in T-lineage acute lymphoblastic leukemia: a report from the Children's Oncology Group. *Blood* 110, 1429-1438.