Using Random Forest to Learn Imbalanced Data

Chao Chen, chenchao@stat.berkeley.edu
Department of Statistics,UC Berkeley
Andy Liaw, andy_liaw@merck.com
Biometrics Research,Merck Research Labs
Leo Breiman, leo@stat.berkeley.edu
Department of Statistics,UC Berkeley

Abstract

In this paper we propose two ways to deal with the imbalanced data classification problem using random forest. One is based on cost sensitive learning, and the other is based on a sampling technique. Performance metrics such as precision and recall, false positive rate and false negative rate, F-measure and weighted accuracy are computed. Both methods are shown to improve the prediction accuracy of the minority class, and have favorable performance compared to the existing algorithms.

1 Introduction

Many practical classification problems are *imbalanced*; i.e., at least one of the classes constitutes only a very small minority of the data. For such problems, the interest usually leans towards correct classification of the "rare" class (which we will refer to as the "positive" class). Examples of such problems include fraud detection, network intrusion, rare disease diagnosing, etc. However, the most commonly used classification algorithms do not work well for such problems because they aim to minimize the overall error rate, rather than paying special attention to the positive class. Several researchers have tried to address the problem in many applications such as fraudulent telephone call detection (Fawcett & Provost, 1997), information retrieval and filtering (Lewis & Catlett, 1994), diagnosis of rare thyroid deceases (Murphy & Aha, 1994) and detection of oil spills from satellite images (Kubat et al., 1998).

There are two common approaches to tackle the problem of extremely imbalanced data. One is based on cost sensitive learning: assigning a high cost to misclassification of the minority class, and trying to minimize the overall cost. Domingos (1999) and Pazzani et al. (1994) are among these. The other approach is to use a sampling technique: Either down-sampling the majority class or over-sampling the minority class, or both. Most research has been focused on this approach. Kubat et al. (1997) develop a system, SHRINK, for imbalanced classification. SHRINK labels a mixed region as positive (minority class) regardless of whether the positive examples prevail in the region or not. Then it searches for the best positive region. They made comparisons to C4.5 and 1-NN, and show that SHRINK has improvement in most cases. Kubat

& Matwin (1997) uses the one-sided sampling technique to selectively down sample the majority class. Ling & Li (1998) over-sample the minority class by replicating the minority samples so that they attain the same size as the majority class. Over-sampling does not increase information; however by replication it raises the weight of the minority samples. Chawla et al. (2002) combine over-sampling and down-sampling to achieve better classification performance than simply down-sampling the majority class. Rather than over-sampling with replacement, they create synthetic minority class examples to boost the minority class (SMOTE). They compared SMOTE plus the down-sampling technique with simple down-sampling, one-sided sampling and SHRINK, and showed favorable improvement. Chawla et al. (2003) apply the boosting procedure to SMOTE to further improve the prediction performance on the minority class and the overall F-measure.

We propose two ways to deal with the problem of extreme imbalance, both based on the random Forest (RF) algorithm (Breiman, 2001). One incorporates class weights into the RF classifier, thus making it cost sensitive, and it penalizes misclassifying the minority class. The other combines the sampling technique and the ensemble idea. It down-samples the majority class and grows each tree on a more balanced data set. A majority vote is taken as usual for prediction. We compared the prediction performance with one-sided sampling, SHRINK, SMOTE, and SMOTEBoost on the data sets that the authors of those techniques studied. We show that both of our methods have favorable prediction performance.

2 Methodology

2.1 Random Forest

Random forest (Breiman, 2001) is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Random forest generally exhibits a substantial performance improvement over the single tree classifier such as CART and C4.5. It yields generalization error rate that compares favorably to Adaboost, yet is more robust to noise. However, similar to most classifiers, RF can also suffer from the curse of learning from an extremely imbalanced training data set. As it is constructed to minimize the overall error rate, it will tend to focus more on the prediction accuracy of the majority class, which often results in poor accuracy for the minority class. To alleviate the problem, we propose two solutions: balanced random forest (BRF) and weighted random forest (WRF).

2.2 Balanced Random Forest

As proposed in Breiman (2001), random forest induces each constituent tree from a bootstrap sample of the training data. In learning extremely imbalanced data, there is a significant probability that a bootstrap sample contains few or even none of the minority class, resulting in a tree with poor performance for predicting the minority class. A naïve way of fixing this problem is to use a stratified bootstrap; i.e., sample with

replacement from within each class. This still does not solve the imbalance problem entirely. As recent research shows (e.g., Ling & Li (1998),Kubat & Matwin (1997),Drummond & Holte (2003)), for the tree classifier, artificially making class priors equal either by down-sampling the majority class or over-sampling the minority class is usually more effective with respect to a given performance measurement, and that down-sampling seems to have an edge over over-sampling. However, down-sampling the majority class may result in loss of information, as a large part of the majority class is not used. Random forest inspired us to ensemble trees induced from balanced down-sampled data. The Balanced Random Forest (BRF) algorithm is shown below:

1. For each iteration in random forest, draw a bootstrap sample from the minority class. Randomly draw the same number of cases, with replacement, from the majority class.
2. Induce a classification tree from the data to maximum size, without pruning. The tree is induced with the CART algorithm, with the following modification: At each node, instead of searching through all variables for the optimal split, only search through a set of m_{try} randomly selected variables.
3. Repeat the two steps above for the number of times desired. Aggregate the predictions of the ensemble and make the final prediction.

2.3 Weighted Random Forest

Another approach to make random forest more suitable for learning from extremely imbalanced data follows the idea of cost sensitive learning. Since the RF classifier tends to be biased towards the majority class, we shall place a heavier penalty on misclassifying the minority class. We assign a weight to each class, with the minority class given larger weight (i.e., higher misclassification cost). The class weights are incorporated into the RF algorithm in two places. In the tree induction procedure, class weights are used to weight the Gini criterion for finding splits. In the terminal nodes of each tree, class weights are again taken into consideration. The class prediction of each terminal node is determined by "weighted majority vote"; i.e., the weighted vote of a class is the weight for that class times the number of cases for that class at the terminal node. The final class prediction for RF is then determined by aggregatting the weighted vote from each individual tree, where the weights are average weights in the terminal nodes. Class weights are an essential tuning parameter to achieve desired performance. The out-of-bag estimate of the accuracy from RF can be used to select weights. This method, Weighted Random Forest (WRF), is incorporated in the present version of the software.

3 Experiments

3.1 Data set

We experimented with 6 data sets, and they are summarized in table 1. These data sets are highly imbalanced and have been studied before by various researchers with different methods. We try to compare our proposed

methods with those existing methods, and we will also compare the performance of WRF and BRF. Here is the description of the data.

Dataset	Number of variables	Number of cases	% Minority class
Oil	50	937	4.4
Mammography	6	11183	2.3
Satimage	36	6435	9.7
Hypothyroid	24	2520	4.8
Euthyroid	24	2640	9.1
KDD thrombin	100	2543	7.6

Table 1: Data Set summary

1. The oil data set was first studied by Kubat & Matwin (1997) with their method, one-sided sampling. Kubat et al. (1998) further studied this dataset and provides a new method, SHRINK. Chawla et al. (2002) compared their methods SMOTE with One-sided sampling and SHRINK on the same dataset. This dataset has 41 oil slick samples and 896 non-slick samples.

2. The mammography data set from Woods et al. (1993) has 10,923 negative samples and only 260 positive samples. This dataset was studied with the methods SMOTE and SMOTEboost in Chawla et al. (2002) and Chawla et al. (2003), respectively.

3. The Hypothyroid and Euthyroid data sets (Blake & Merz, 1998) are studied by Kubat et al. (1997) with SHRINK, C4.5 and 1-NN. We follow Kubat et al. (1997) and deleted all cases with missing age and sex and removed the attribute TBG_measured. From Euthyroid, we randomly selected 240 positives and 2400 negatives; from hypothyroid, we select 120 positive and 2400 negatives.

4. The satimage data set (Blake & Merz, 1998) originally has six classes. Chawla et al. (2003) chose the smallest class as the minority class and collapsed the rest of the classes into one, and use the modified dataset to evaluate the performance of SMOTE and SMOTEBoost.

5. The KDD Cup 2001 thrombin data set was originally split into training and test components. We combine the training set and test set together, and we have 2543 negative samples and 190 positive samples. The original data set has 139,351 binary features, and we use maximum entropy to select 100 features for our analysis.

3.2 Performance Measurement

In learning extremely imbalanced data, the overall classification accuracy is often not an appropriate measure of performance. A trivial classifier that predicts every case as the majority class can still achieve very high

accuracy. We use metrics such as true negative rate, true positive rate, weighted accuracy, G-mean, precision, recall, and F-measure to evaluate the performance of learning algorithms on imbalanced data. These metrics have been widely used for comparison. All the metrics are functions of the confusion matrix as shown in Table 2. The rows of the matrix are actual classes, and the columns are the predicted classes. Based on Table 2, the performance metrics are defined as:

$$\begin{aligned}
\text{True Negative Rate } (Acc^-) &= \frac{TN}{TN+FP} \\
\text{True Positive Rate } (Acc^+) &= \frac{TP}{TP+FN} \\
G\text{-mean} &= (Acc^- \times Acc^+)^{1/2} \\
\text{Weighted Accuracy} &= \beta Acc^+ + (1-\beta) Acc^- \\
\text{Precision} &= \frac{TP}{TP+FP} \\
\text{Recall} &= \frac{TP}{TP+FN} = Acc^+ \\
F\text{-measure} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

Keep in mind. The 'predicated positive' is the record you would pay attention to. That's why it is super important!

	Predicted Positive Class	Predicted Negative Class
Actual Positive class	TP (True Positive)	FN (False Negative)
Actual Negative class	FP (False Positive)	TN (True Negative)

Table 2: Confusion matrix.

usually care more about true positive rate(recall rate) - how many positive samples could be incorported while maintaining a reasonable False positive rate.

For any classifier, there is always a trade off between true positive rate and true negative rate; and the same applies for recall and precision. In the case of learning extremely imbalanced data, quite often the rare class is of great interest. In many applications such as drug discovery and disease diagnosis, it is desirable to have a classifier that gives high prediction accuracy over the minority class (Acc^+), while maintaining reasonable accuracy for the majority class (Acc^-). Weighted Accuracy is often used in such situations. Weights can be adjusted to suit the application. Here we use equal weights for both true positive rate and true negative rate; i.e., β equals 0.5. The Geometric Mean (G-mean) is used in Kubat et al. (1997) to assess the performance of their methods. Precision, recall and F-measure are commonly used in the information retrieval area as performance measures. We will adopt all these measurements to compare our methods with published results. Ten-fold cross-validations were carried out to obtain all the performance metrics.

We also use the ROC curve to compare the performance of BRF and WRF. The ROC curve is a graphical representation of the trade off between the false negative and false positive rates for every possible cut off. For BRF, we can change the votes cutoff for final prediction: as we raise the cutoff for the minority class, we can achieve a lower true positive rate and a higher true negative rate, thus yielding a set of points on the ROC diagram. For WRF, we can tune the class weight for final prediction: as we raise the minority class

Methods	Acc^+ (Recall)	Acc^-	Precision	F-measure	G-mean	Wt. Accuracy
One-sided Sampling	76.0	86.6	20.53	32.33	81.13	81.3
SHRINK	82.5	60.9	8.85	15.99	70.88	71.7
SMOTE 500% + Down 50%	89.5	78.9	16.37	27.68	84.03	84.2
SMOTE 500% + Down 100%	78.3	68.7	10.26	18.14	73.34	73.5
BRF cutoff=.5	73.2	91.6	28.57	**41.10**	81.19	82.4
BRF cutoff=.6	85.4	84.0	19.66	31.96	84.70	84.7
BRF cutoff=.7	92.7	72.2	13.24	23.17	81.18	82.5
WRF weight=41:896	92.7	82.4	19.39	32.07	**87.40**	**87.6**

Table 3: Performance comparison on oil spill data set

weight, the misclassification cost of the minority class goes up, and we get a higher true positive rate and a lower true negative rate, thus yielding a set of points on the ROC diagram.

3.3 Performance Comparison

Table 3 compares the performance of different algorithms on the oil spill data. Chawla et al. (2002) provides a comparison of One-side Sampling, SHRINK and SMOTE with down-sampling based on precision, recall and F-measure. We furthermore compute the other metrics from their original numbers. As stated in Chawla et al. (2002), SMOTE with down-sampling achieves results comparable with SHRINK's, and in some cases better results, but shows no clear improvement over One-sided selection. From table 3, apparently both BRF and WRF show great improvement over SHRINK based on all metrics (G-mean, weighted accuracy and F-measure). BRF with a cutoff of 0.5 has comparable performance with One-sided Sampling based on G-mean and weighted accuracy, and a better result in F-measure. WRF using a weight equal to the class proportion has a comparable result in F-measure, but favorable results in G-mean and weighted accuracy. BRF with a cutoff of 0.6 and WRF both achieve better performance than SMOTE 500% with 50% downsampling, which is best among the results for SMOTE. We can conclude that for the oil spill data both BRF and WRF with proper parameters outperform the published results.

Table 4 compares the performance of BRF and WRF with SMOTE and SMOTEboost on the mammography data. WRF shows improvement over both SMOTE and SMOTEBoost based on the F-measure, G-mean and weighted accuracy. BRF has better perfomance than SMOTE, while comparing with SMOTEBoost, it has a better G-mean and weighted accuracy, but worse F-measure.

Table 5 compares the performance of BRF and WRF with SMOTE and SMOTEboost on the satimage data. Both BRF and WRF are superior to SMOTE and Standard RIPPER in F-measure and G-mean. BRF and WRF are better than SMOTEBoost in G-mean, but worse in F-measure. Note that, compared to the

Method	Acc^+ (Recall)	Acc^-	Precision	F-measure	G-mean	Wt. Accuracy
Standard RIPPER	48.12	99.61	74.68	58.11	69.23	73.87
SMOTE 100	58.04	99.26	64.96	61.31	75.90	78.65
SMOTE 200	62.16	99.04	60.53	60.45	78.46	80.58
SMOTE-Boost 100	61.73	99.54	76.59	68.36	78.39	80.63
SMOTE-Boost 200	62.63	99.50	74.54	68.07	78.94	81.07
BRF cutoff=.2	70.00	98.98	62.12	65.83	83.24	84.49
BRF cutoff=.3	76.54	98.21	50.51	60.86	**86.70**	**87.38**
WRF weight=2:1	65.38	99.57	78.34	**71.28**	80.68	82.48
WRF weight=3:1	72.69	99.25	69.74	**71.18**	**84.94**	**85.97**

Table 4: Performance comparison on mammography data set

other methods, BRF and WRF tend to focus more on the accuracy of the minority class while trading off accuracy in the majority class.

Method	Acc^+ (Recall)	Acc^-	Precision	F-measure	G-mean	Wt. Accuracy
Standard RIPPER	47.43	97.59	67.92	55.50	68.03	72.51
SMOTE 100	65.17	94.46	55.88	59.97	78.46	79.82
SMOTE 200	74.89	91.29	48.08	58.26	82.68	83.09
SMOTE-Boost 100	63.88	98.02	77.71	70.12	79.13	80.95
SMOTE-Boost 300	67.87	97.25	72.68	**70.19**	81.24	82.56
BRF cutoff=.3	67.09	95.97	64.22	65.62	80.24	81.53
BRF cutoff=.4	77.00	93.56	56.31	65.05	**84.88**	**85.28**
WRF weight=	69.33	96.71	69.44	69.38	81.88	83.02
WRF weight=	77.48	94.56	60.55	67.98	**85.60**	**86.02**

Table 5: Performance comparison on Satimage data set

Tables 6 and 7 compare the performance of BRF and WRF with SHRINK on the hypothyroid and euthyroid data sets from the UCI repository. Kubat et al. (1997) provides a comparison among C4.5, 1-NN and SHRINK; they use G-mean to evaluate the performance. Clearly based on G-mean, both BRF and WRF outperform SHRINK, C4.5 and 1-NN. We can hardly tell the difference between BRF and WRF. BRF is slightly better than WRF in G-mean and weighted accuracy, but worse in F-measure.

In the tables above, we have shown that WRF and BRF have favorable improvement over existing methods. However, between WRF and BRF, we can not tell clearly which is superior. We will use ROC analysis to further investigate these two methods.

Method	Acc^+ (Recall)	Acc^-	Precision	F-measure	G-mean	Wt. Accuracy
C4.5	.	.	.	.	93.6	.
1-NN	.	.	.	.	88.9	.
SHRINK	.	.	.	.	95.0	.
BRF cutoff=.5	95.0	98.6	63.3	76.0	**96.8**	**96.8**
WRF weight=1:5	93.3	99.0	83.6	**88.2**	96.1	96.2

Table 6: Performance comparison on Hypothyroid data set

Method	Acc^+ (Recall)	Acc^-	Precision	F-measure	G-mean	Wt. Accuracy
C4.5	.	.	.	.	88.2	.
1-NN	.	.	.	.	60.8	.
SHRINK	.	.	.	.	74.0	.
BRF cutoff=.5	91.3	97.1	75.8	82.8	**94.1**	**94.2**
WRF weight=1:5	90.0	98.0	81.5	**85.5**	93.9	94.0

Table 7: Performance comparison on Euthyroid data set

Figures 1–5 compare WRF and BRF using ROC curves on the data sets. From the figures, we can see that the ROC curve of both WRF and BRF are very close. WRF seems to be slightly superior than BRF on the oil spill and mammography dat a sets and vice versa on the euthyroid and Thrombin data sets. From several other data sets we have experimented with, we do not see a clear winner.

4 Conclusion

We presented two ways of learning imbalanced data based on random forest. Weighted Random Forest put more weights on the minority class, thus penalizing more heavily on misclassifying the minority class. Balanced Random Forest combines the down sampling majority class technique and the ensemble learning idea, artificially altering the class distribution so that classes are represented equally in each tree.

From the experiments on various data sets, we can conclude that both Weighted RF and Balanced RF have performance superior to most of the existing techniques that we studied. Between WRF and BRF, however, there is no clear winner. By the construction of BRF and WRF, we found that BRF is computationally more efficient with large imbalanced data, since each tree only uses a small portion of the training set to grow, while WRF needs to use the entire training set. WRF assigns a weight to the minority class, possibly making it more vulnerable to noise (mis-labeled class) than BRF. A majority case that is mislabled as belonging to the minority class may have a larger effect on the prediction accuracy of the majority class in WRF than in BRF. Further study may be carried out to see how these two methods perform under label noise.

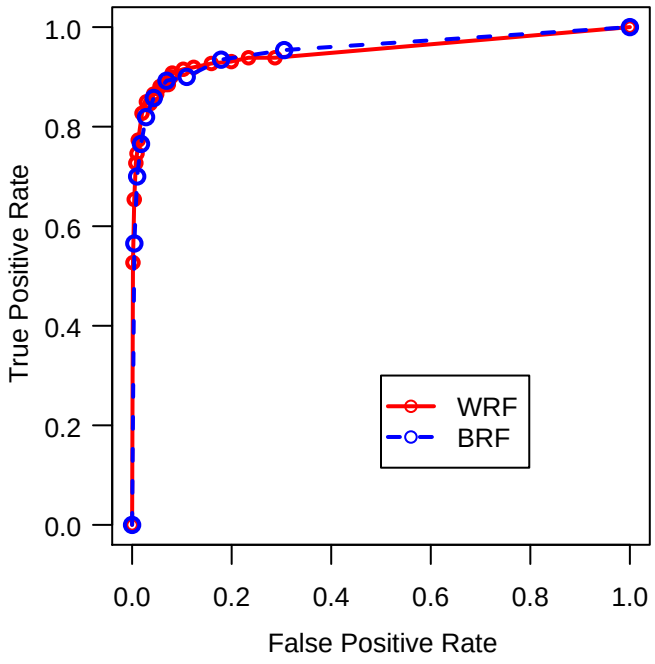

Figure 1: Mamo ROC

5 Acknowledgements

This work is partially a result of collaboration on QSAR modeling with the Biometrics Research Department at Merck Research Labs. We thank Vladimir Svetnik for his support and discussion on ensemble learning ideas. We also thank Nitesh V. Chawla for providing the mammography data set and Robert Holte for providing the oil spill data set used in their papers, and DuPont Pharmaceuticals Research Laboratories for providing KDD Cup 2001 thrombin data.

References

Blake, C. & Merz, C. (1998). Uci repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/~MLRepository.html`.

Breiman, L. (2001). Random forest. *Machine Learning*, *45*, 5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEboost: Improving prediction of the minority class in boosting. In *7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, (pp. 107–119).

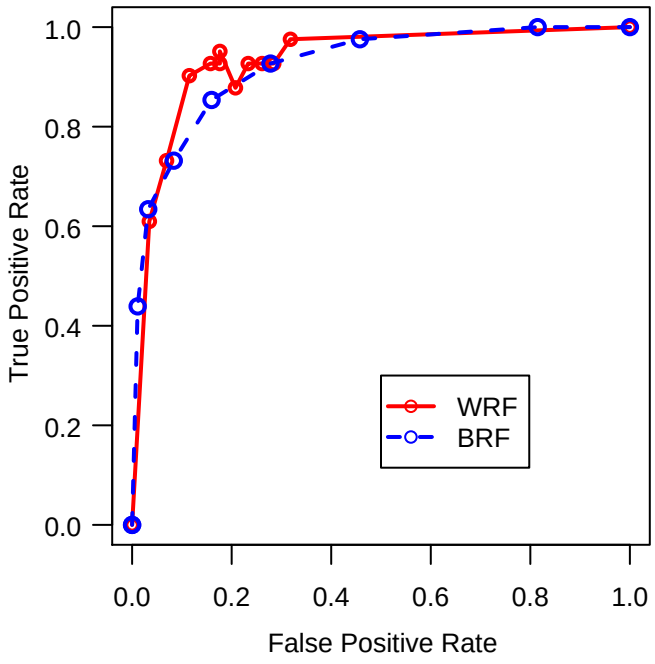

Figure 2: Oil Spill ROC

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 155–164)., San Diego. ACM Press.

Drummond, C. & Holte, R. (2003). C4.5,class imbalance, and cost sensitivity:why undersampling beats over-sampling. In *Proceedings of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II*.

Fawcett, T. E. & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, *1*, 291–316.

Kubat, M., Holte, R., & Matwin, S. (1997). Learning when negative examples abound. In *ECML-97*, (pp. 146–153).

Kubat, M., Holte, R., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, *30*, 195–215.

Kubat, M. & Matwin, S. (1997). Addressing the curse of imbalanced data sets: One-sided sampling. In *Proceedings of the 14th International conference on Machine Learning*, (pp. 179–186). Morgan Kaufmann.

Lewis, D. D. & Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In Cohen, W. W. & Hirsh, H. (Eds.), *Proceedings of ICML-94, 11th International Conference on Machine Learning*, (pp. 148–156)., San Francisco. Morgan Kaufmann.

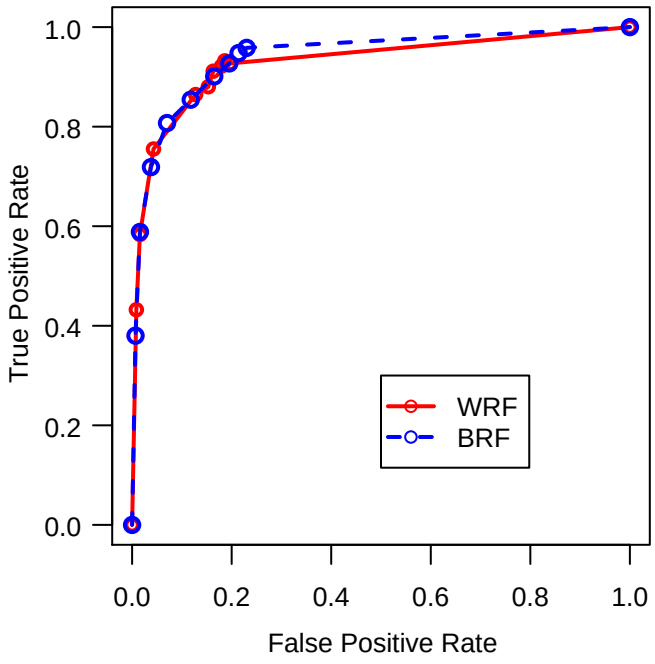

Figure 3: KDD Thrombin ROC

Ling, C. & Li, C. (1998). Data mining for direct marketing problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, New York.

Murphy, P. M. & Aha, D. W. (1994). UCI repository of machine learning databases. University of California-Irvine, Department of Information and Computer Science. `http://www1.ics.uci.edu/mlearn/MLRepository.html`.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing misclassification costs. In *Proceedings of the 11th International Conference on Machine Learning*, San Francisco. Morgan Kaufmann.

Woods, K., Doss, C., Bowyer, K., Solka, J., Preibe, C., & Kegelmyer, P. (1993). Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*, 1417–1436.

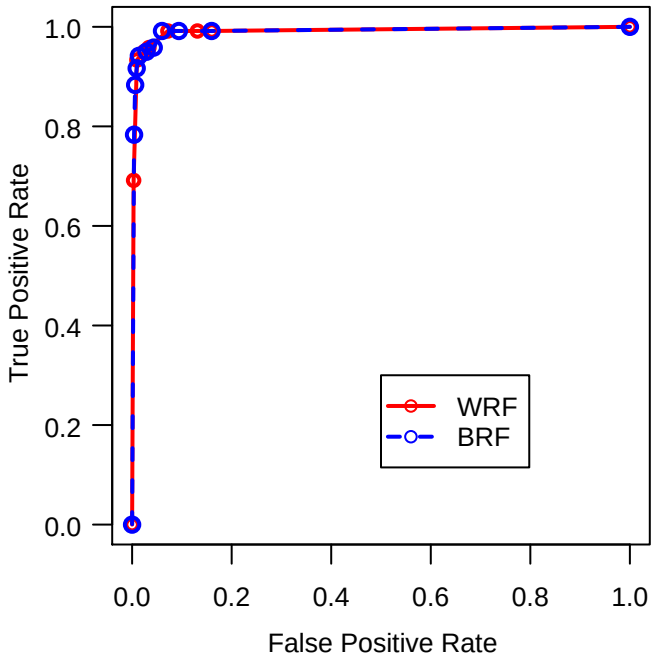

Figure 4: Hypothyroid ROC

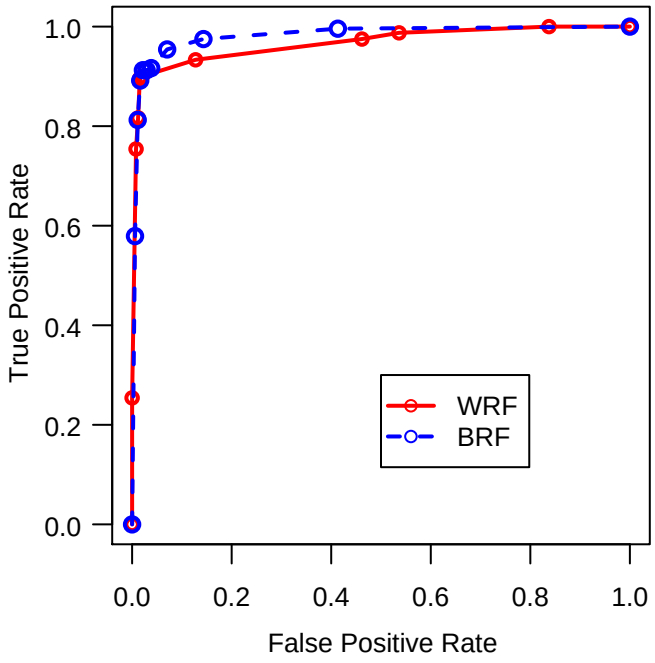

Figure 5: Euthyroid ROC