

# Bagging in the Presence of Outliers

P. Hall

B.A. Turlach

ACSys CRC & CMA  
The Australian National University  
Canberra ACT 0200  
Australia

ACSys CRC & CMA  
The Australian National University  
Canberra ACT 0200  
Australia

## Abstract

We investigate the performance of bagging methods in the presence of outliers. The results are best illustrated and intuitively explained for classical classification problems to which we shall restrict our focus in this paper. It is shown that bagging methods can improve the resistance of classification rules to outlier contamination, especially if an  $m$ -out-of- $n$  bagging scheme is used. However, the outlier-reduction property does not improve performance to such an extent that outliers are no longer noticeable. It is also shown that in the absence of contamination by outliers the effects of bagging are negligible. Therefore, when bagging is not really needed, deleterious effects that result from employing it are quite small.

## 1 Introduction

The method of bootstrap aggregation, or bagging, was introduced by Breiman (1996a). Recent accounts of its properties, in the settings of both prediction and classification, include those of Breiman (1996b) and Tibshirani (1996). When used for classification it involves applying a classifier to bootstrap replications of the training sample, and discriminating in favour of that population which is favoured in the majority of replications. This process is depicted in Figure 1.1. Breiman (1996a) gives a simple theoretical argument, involving little more than the Cauchy–Schwarz inequality, explaining why bagging works for prediction. In essence, bagging minimises prediction mean squared error because (a) it provides an empirical, or bootstrap, approximation to the mean value of the predictor in the distribution of training samples; and (b) the mean square of a random quantity is always minimised by centring at its mean.

In the context of classification, Breiman (1996a) interprets bagging as an empirical device correcting for some of the error incurred by not using a classifier based

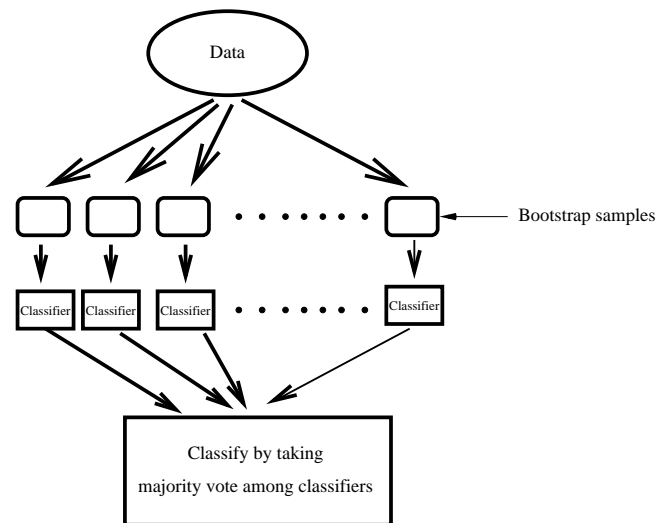


Figure 1.1: Illustration of the bagging idea

strictly on likelihood. In effect, if we use a poor classifier then bagging enables us to recover much of the performance of a likelihood-based rule. However, for both classification and prediction Breiman warns that bagging should be used with caution, since it can negatively affect the performance of a rule that is already close to optimal. In effect, the information that bagging extracts from the data to improve performance might not be directly available for the original purpose of classification or prediction, and in a sense could be wasted.

In this paper we take a somewhat different view of bagging in the context of classification, and come up with results that are partly in contradistinction to those of Breiman (1996a). We consider a classical problem based on likelihoods with unknown parameters, which must be estimated from data. Thus, we are never very far from the classical, optimal setting described in the Neyman–Pearson lemma. However, we do focus on situations where classification decisions are difficult, and in that sense the variability of our classifiers is high.

Our conclusions are twofold. First, if the model is valid then the likelihood-based classifier and its bagged version both perform equivalently, in the sense that they have asymptotically equal tendency to commit errors in favour of either population. Hence, in this example the method used for discrimination was close to optimal even before bagging, but despite this, there is negligible degradation due to bagging. Secondly, if the model is invalid because of contamination by outliers, and if the outliers are such that they skew the decision procedure predominantly in one direction or another, then bagging may produce a non-negligible improvement of error rates. It does not remove the impact of outliers completely, but it may reduce the worst effects of outliers on classification error.

In Section 2 we shall describe the classification problem on which we focus in this paper. We describe on an intuitive level how the presence of outliers effects the classifiers that we consider. The results of a simulation study, given in Section 3, confirm this discussion. Some conclusions are given in Section 4. These conclusions also hold in other settings that are described in a forthcoming manuscript of the authors which will also give a theoretical discussion and quantification of the results discussed here.

## 2 Setting

To study the effect of bagging in classical classification problems, with or without the presence of outliers, we use a setup in which the data come from either of two populations. One population, say  $\Pi_1$ , has a normal  $\mathcal{N}(-1, 1)$  distribution and the other, say  $\Pi_2$ , has an  $\mathcal{N}(1, 1)$  distribution. Our aim is to compare the performance of the likelihood-based classifier, say  $R_1$  (see, e.g. Mardia et al., 1979, pp. 300), and the bagging classifier, say  $R_2$ . The rule  $R_2$  allocates a point according to the following procedure: Resample with replacement from each training set, i.e. take bootstrap samples, and construct the classifier  $R_1$  on each resample. Repeat this resampling process (in our simulations we used 50 resamples) and classify the point according to a majority vote of its classifications on the bootstrap samples.

We are especially interested in how these rules perform on points that are difficult to classify correctly. In our setting, such points are (a)  $x_0 = 0$ , the abscissa at which the density functions of the two populations cross, and, since we are in a parametric setting, (b) points which are of order  $n^{-1/2}$  away from  $x_0$ . In our simulation study we shall use  $x$ -values of the form  $cn^{-1/2}$ ,  $c = -1, -0.9, \dots, 0.9, 1$  and study how these are classified under the different settings by each of the rules  $R_1$  respectively  $R_2$ .

Let us discuss the rule  $R_1$  in order to understand on an intuitive level its behaviour in the presence of outliers. Assume that we have a training set available from each population, and both sets contain the same number of points. Rule  $R_1$  calculates the sample mean  $\hat{\mu}_i$ ,  $i = 1, 2$  for each training set and the over all means  $\hat{\mu} = (\hat{\mu}_1 + \hat{\mu}_2)/2$ . A point is now allocated to population  $\Pi_1$  if it lies on the same side of  $\hat{\mu}$  as  $\hat{\mu}_1$ , otherwise it is allocated to  $\Pi_2$ .

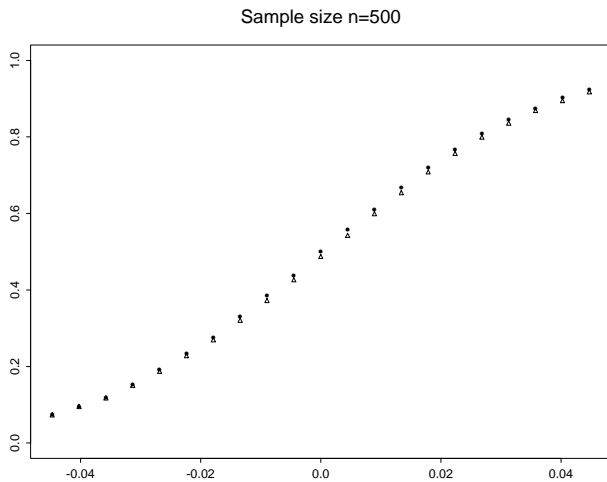
The outliers which we shall use in our simulation study result in an increase in  $\hat{\mu}_1$ . For the sake of simplicity, let us assume that the  $j$ th outlier is of the form  $X_j + w_j$  where  $X_j$  is generated from  $\Pi_1$  and  $w_j > 0$  is constant. Set  $w = \sum_j w_j$ . The resulting values of  $\hat{\mu}_1$  and  $\hat{\mu}$  are shifted to the right by approximately  $n^{-1}w$  and  $\frac{1}{2}n^{-1}w$ , respectively. We can distinguish two cases. First, if  $w$  is below a critical value, say  $w_0$ , such that with high probability  $\hat{\mu}_1 \leq \hat{\mu} \leq \hat{\mu}_2$  holds, the rule  $R_1$  is biased in favour of  $\Pi_1$  for the  $x$ -values that we consider. Secondly, if  $w$  is larger than  $w_0$  such that with high probability  $\hat{\mu}_1 \geq \hat{\mu} \geq \hat{\mu}_2$  holds, the rule  $R_1$  is biased in favour of  $\Pi_2$ . In fact, in this situation all  $x$ -values that we consider will be allocated to  $\Pi_2$ . In our setup, the critical value  $w_0$  is approximately  $2n$  since  $E[\hat{\mu}_1] = -1 + n^{-1}w$  and  $E[\hat{\mu}] = \frac{1}{2}n^{-1}w$ .

## 3 Numerical Results

We conducted the following simulation study of the effect of bagging in the setting described in Section 2. For all choices of parameters, described below in detail, we performed 5000 replications. We generated training samples from populations  $\Pi_1$  and  $\Pi_2$ . Using these training samples we classified points in the neighbourhood of  $x_0$  using the likelihood-based classifier  $R_1$  and the bagged classifier  $R_2$ . In the first part of our simulations study we investigated the relationship between  $R_1$  and  $R_2$  if no outliers were present. We recorded how  $x$ -values of the form  $cn^{-1/2}$ ,  $c = -1, -0.9, \dots, 0.9, 1$ , were classified by  $R_1$  and  $R_2$  using training sets of size  $n = 25, 50, 100, 250, 500, 1000$  and  $2000$  from each population.

There was no significant difference in the performance of these two classifiers. Figure 3.1 gives, on the vertical axis, the empirical probabilities with which each point was classified as belonging to  $\Pi_2$  for  $n = 500$ . The horizontal axis gives values of  $cn^{-1/2}$  for  $-1 \leq c \leq 1$ .

To study the effect of outliers we fixed the size of the training sets at  $n = 500$  and, in the training set sampled from  $\Pi_1$ , replaced the first  $k$  observations (for  $k = 1, 2, 4, 6, 8$  or  $10$ ) by outliers of different magnitude (values between 1 and 2000). We recorded how the  $x$ -values in Figure 3.1 were classified with this kind of contamination present.



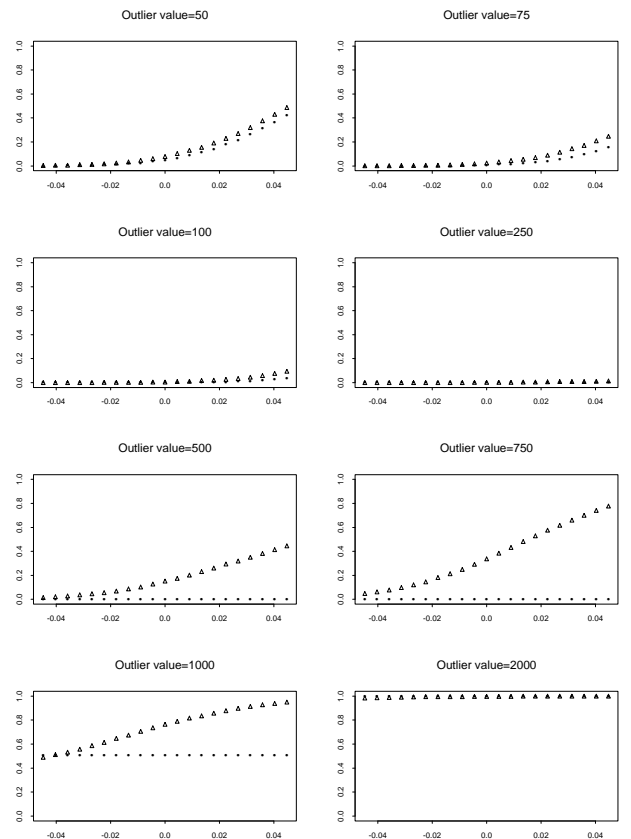
**Figure 3.1:** Comparison of classification rules  $R_1$  (points) and  $R_2$  (triangles) if no outliers are present. The picture shows the empirical probabilities, calculated from 5000 replications, that a point on the  $x$ -axis would be allocated to  $\Pi_2$ . Classification to  $\Pi_2$  is optimal when  $x > 0$ , and otherwise classification to  $\Pi_1$  is optimal.

Our simulation study shows that usually, but not always, the bagging rule  $R_2$  does at least as well as  $R_1$ , and sometimes substantially outperforms  $R_1$ . Let us first consider the case where there is only one outlier, coming from  $\Pi_1$  and taking a positive value  $w$ , say. As a result, the decision rule  $R_1$  is biased in favour of  $\Pi_2$  for large values of  $w$  and biased in favour of  $\Pi_1$  otherwise.

To understand the performance of  $R_2$ , note that the probabilities that a given resample contains 0, 1 or more than 1 replicate of the outlier are approximately  $e^{-1} \approx 0.368$ ,  $e^{-1}$  and  $1 - 2e^{-1} \approx 0.264$ , respectively. If  $w$  is greater than the critical value  $w_0$  the outlier induces a bias in favour of  $\Pi_2$  for discrimination rules based on bootstrap samples with 1 or more replicates of the outlier. Hence, since  $0.632 > 0.5$  and the bagging rule takes the majority vote over the bootstrap samples, we cannot expect  $R_2$  to perform better than  $R_1$  in this situation.

If  $w$  is smaller than the critical value  $w_0$ , we have to investigate its influence on discrimination rules based on resamples that have more than one replicate of the outlier. For the sake of simplicity, we shall neglect in this discussion the case of having more than two replicates of the outlier in a resample and only distinguish the cases  $2w < w_0$  and  $2w > w_0$ .

If  $2w < w_0$ , the bias in favour of  $\Pi_1$  of a discrimination rule based on a bootstrap resample with 0, 1 or more than 1 replicate of the outlier, is less than, equal to, or greater than, respectively, the bias of  $R_1$ . Therefore, since  $0.368 > 0.264$ , we might expect the bagging rule – which is the average of decision rules over resamples –



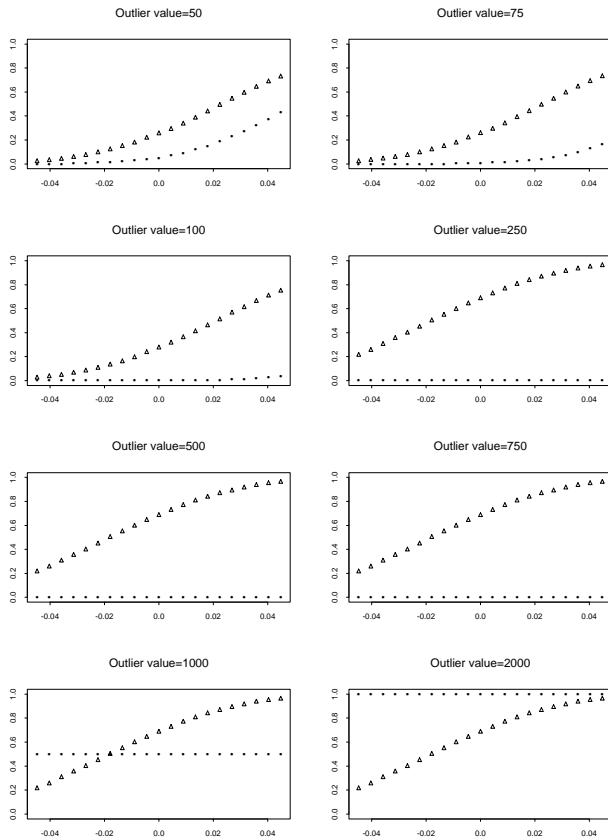
**Figure 3.2:** Comparison of classification rules  $R_1$  (points) and  $R_2$  (triangles) with outliers present. The picture shows the empirical probabilities that points on the  $x$ -axis would be allocated to  $\Pi_2$  when a single outlier was present in samples generated from  $\Pi_1$ . Classification to  $\Pi_2$  is optimal when  $x > 0$ , and otherwise classification to  $\Pi_1$  is optimal.

to be less biased in favour of  $\Pi_1$  in the majority of cases, and therefore perform better overall.

If  $2w > w_0$ , the bias in favour of  $\Pi_1$  of a discrimination rule based on a bootstrap resample with 0 or 1 outlier is less than respectively equal to the bias of  $R_1$ . However, discrimination rules based on bootstrap resamples with more than 1 replicate of the outliers will be biased in favour of  $\Pi_2$ . Hence, we again might expect the bagging rule to be less biased in favour of  $\Pi_1$  in the majority of cases and therefore perform better than  $R_1$ .

This discussion is confirmed by Figure 3.2. Our simulation results show that the biggest improvement of the bagged rule  $R_2$  over the likelihood-based rule  $R_1$  is observed for the case  $w < w_0 < 2w$ .

These results motivate studying an  $m$ -out-of- $n$  bagging scheme as this should reduce the influence of the outlier(s) in the resampling step. Here, each bootstrap sample consists of  $m$  observations drawn with replace-

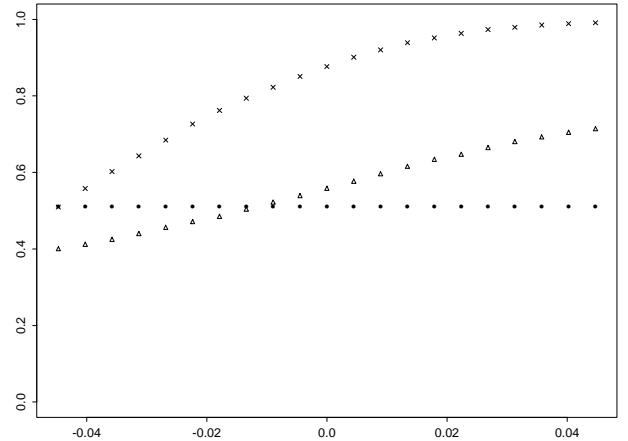


**Figure 3.3:** Comparison of classification rules  $R_1$  (points) and  $R_2$  (triangles) with outliers present. The picture shows the empirical probabilities that points would be allocated to  $\Pi_2$ . One outlier was present in the samples generated from  $\Pi_1$ , and  $m$ -out-of- $n$  bagging was used with  $m = 100$ .

ment from the training sets. We used  $m = 100, 250$  and  $400$ . Intuitively, the  $m$ -out-of- $n$  bootstrapping scheme should improve the rule  $R_2$  since, for appropriate choice of  $m$ , the probability that an outlier is included in a bootstrap resample will be small.

For the settings in which only one outlier was present our simulations demonstrate a significant improvement if an  $m$ -out-of- $n$  bagging scheme is used. Figure 3.3 shows the results for the same settings as Figure 3.2 but with an  $m$ -out-of- $n$  scheme where  $m = 100$ .

The case of  $k \geq 2$  outliers is more complex. In principle, since the probability that a bootstrap resample of size  $m$  contains just  $j$  replicates of the  $k$  outliers equals approximately  $(1/j!)(km/n)^j e^{-km/n}$ , and so decreases with  $m$  if  $j \geq 1$ , then the performance of an  $m$ -out-of- $n$  bagging rule can be improved by decreasing the value of  $m$ . However, this argument fails to take account of the fact that the amount of bias of a rule with  $j$  outliers depends on the value of  $j$  and the sizes of the outliers, not



**Figure 3.4:** Comparison of classification rules  $R_1$  (points),  $R_2$  using “standard” bagging (triangles) and  $R_2$  using 100-out-of- $n$  bagging (crosses) with multiple outliers present. The picture shows the empirical probabilities that the points would be allocated to  $\Pi_2$ . Here, two outliers of magnitude 500 are present in samples ( $n = 500$ ) generated from  $\Pi_1$ .

just on whether  $j > k$  or  $j < k$ . As a result, relatively small  $m$ 's can give poor performance against the standard  $n$ -out-of- $n$  rule. This is illustrated in Figure 3.4, which shows that for  $k = 2$  the 500-out-of-500 form of the rule  $R_2$  can outperform both the 100-out-of-500 form and the likelihood ratio rule  $R_1$

## 4 Conclusions

In summary, our simulation results show that if no outliers are present then the bagging rule  $R_2$  does not improve the likelihood-based rule  $R_1$ . If outliers are present then  $R_2$  may improve on  $R_1$ , especially if an  $m$ -out-of- $n$  bootstrap scheme is used. However, the correct choice of  $m$  depends on the number and sizes of outliers.

## References

- Breiman, L. (1996a). Bagging predictors, *Machine Learning* **26**: 123–140.
- Breiman, L. (1996b). Bias, variance, and arcing classifiers. Unpublished manuscript.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*, Academic Press, Duluth, London.
- Tibshirani, R. (1996). Bias, variance, and prediction error for classification rules. Unpublished manuscript.