BOOSTING AND BAGGING IN KERNEL DENSITY ESTIMATION

Siloko, I. U[1] and Ishiekwene, C. C[2]
[1]Department of Mathematical Sciences, Edwin Clark University, Kiagbodo, Delta State, Nigeria.
[2]Department of Mathematics, University of Benin, Benin City, Edo State, Nigeria.
*Corresponding author. E-mail: cycigar@yahoo.co.uk

ABSTRACT
Boosting is a bias reduction technique while bagging is a variance reduction method. These two methods aim at reducing the asymptotic mean integrated square error (AMISE). This study aims to show that bagging is a boosting algorithm in kernel density estimation since both techniques use large smoothing parameter(s). This relationship was verified by real and simulated data.

Key words: Smoothing parameter, bias, variance, boosting, bagging, asymptotic mean integration squared error (AMISE), weak learners.

INTRODUCTION

Boosting (Freund and Shapire, 1995) and bagging (Breiman, 1996) are two techniques used for combining weak models in order to build better models (Rosset, 2003; Rosset et al., 2004). These algorithms have been discussed by many researchers (Friedman et al., 2000; Buhlmann and Yu, 2003; Mason et al., 1999), and they have come up with different views. The general theoretical and practical conclusion reached is that the weak learners for boosting should be weak while the weak learners for bagging should be strong and in "bias-variance" terms, bagging is a variance reduction technique while boosting is bias reduction operation (Rosset, 2003; Ridgeway 2002).

Boosting and bagging have been shown to be connected because the bootstrap procedure can reduce to boosting procedure, and it implies that the bagging algorithm is a boosting algorithm provided there is an appropriate loss function, thus bagging can be considered as a boosting algorithm which utilizes a very robust linear function as explained by Rosset (2003). Bagging has resulted in excellent performance in classification and regression problems (Breiman, 1998 and Breiman, 2001), leading to taking bagging as a reference point when boosting is been evaluated (Gey and Poggi, 2006).

The boosting model involves the re-weighting of data based on a loss function and in the case of the kernel density estimation; Marzio and Taylor (2004, 2005) obtained such a measure by comparing their first boosting step with the leave-one-out estimate (Silverman, 1986). The multivariate boosting algorithm using the product kernel is a sequential algorithm where at each step $\boldsymbol{m}$ the weak learner is computed as:

$$\hat{\boldsymbol{f}}_{\boldsymbol{m}}(\mathbf{x}) = \left(\prod_{j=1}^{d} \boldsymbol{h}_j\right)^{-1} \sum_{i=1}^{n} \mathbf{W}_{\boldsymbol{m}}(\boldsymbol{i}) \boldsymbol{K}\left(\frac{\mathbf{x}_j - \mathbf{X}_{ij}}{\boldsymbol{h}_j}\right) \quad (1)$$

Where $\boldsymbol{K}$ is a fixed kernel, $\boldsymbol{h}_j$ are the smoothing parameter(s) and $\mathbf{W}_{\boldsymbol{m}}(\boldsymbol{i})$ is the weight of observation $\boldsymbol{i}$ at step $\boldsymbol{m}$. The weight of each observation is updated as:

$$\mathbf{W}_{\boldsymbol{m+1}}(\boldsymbol{i}) = \mathbf{W}_{\boldsymbol{m}}(\boldsymbol{i}) + \log\left(\frac{\hat{\boldsymbol{f}}_{\boldsymbol{m}}(\mathbf{x}_i)}{\hat{\boldsymbol{f}}_{\boldsymbol{m}}^{(-i)}(\mathbf{x}_i)}\right) \quad (2)$$

where $\hat{\boldsymbol{f}}_{\boldsymbol{m}}^{(-i)}(\mathbf{x}_i)$ is the leave-one-out estimator of the multivariate product kernel given by:

$$\hat{\boldsymbol{f}}_{\boldsymbol{m}}^{(-i)}(\mathbf{x}_i) = (\boldsymbol{n}-\mathbf{1})^{-1}\left(\prod_{j=1}^{d} \boldsymbol{h}_j\right)^{-1} \sum_{i=1}^{n} \mathbf{W}_{\boldsymbol{m}}(\boldsymbol{i}) \boldsymbol{K}\left(\frac{\mathbf{x}_j - \mathbf{X}_{ij}}{\boldsymbol{h}_j}\right) \quad (3)$$

Also $\hat{\boldsymbol{f}}_{\boldsymbol{m}}(\mathbf{x}_i)$is of the form given by:

$$\hat{\boldsymbol{f}}_{\boldsymbol{m}}(\mathbf{x}_i) = \left(\prod_{j=1}^{d} \boldsymbol{h}_j\right)^{-1} \sum_{i=1}^{n} \mathbf{W}_{\boldsymbol{m}}(\boldsymbol{i}) \boldsymbol{K}\left(\frac{\mathbf{x}_j - \mathbf{X}_{ij}}{\boldsymbol{h}_j}\right) \quad (4)$$

In boosting in kernel density estimation, the weights will be updated at each step, and the final output is the product of all the density estimates, normalised so that the integrand is unity (Marzio and Taylor, 2004). The algorithm given below is for the multidimensional case in which the product kernel was employed. In product kernel, the axes are restricted to be parallel to the coordinate axis and $\boldsymbol{d}$ inde-pendent smoothing parameters are allowed for each of the coordinate axes (Sain, 2002).

ALGORITHM (Marzio and Taylor Algorithm 2005)

STEP 1. Given
$X_{ij} = (X_{i1}, X_{i2}, \dots, X_{id})^T,\ i = 1,2, \dots, n,\ j = 1,2, \dots, d,$

Initialise $W_1(i) = 1/n$

STEP 2. Select $H = h_1, h_2, \dots, h_d$ the smoothing parameters.

STEP 3. For $m = 1, \dots, M$.

(i) Obtain a weighted kernel estimate.

$$\hat{\boldsymbol{f}}_{\boldsymbol{m}}(\mathbf{x}) = \left(\prod_{\boldsymbol{j=1}}^{\boldsymbol{d}} \boldsymbol{h_j}\right)^{-1} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{W_m(i)K}\left(\frac{\mathbf{x_j} - \boldsymbol{X_{ij}}}{\boldsymbol{h_j}}\right)$$

(ii) Update the weights according to:

$$\mathbf{W_{m+1}}(\boldsymbol{i}) = \mathbf{W_m}(\boldsymbol{i}) + \mathbf{log}\left(\frac{\hat{\boldsymbol{f}}_{\boldsymbol{m}}(\mathbf{x_i})}{\hat{\boldsymbol{f}}_{\boldsymbol{m}}^{(-i)}(\mathbf{x_i})}\right)$$

STEP4. Provide as output

$$\boldsymbol{C}\prod_{\boldsymbol{m=1}}^{\boldsymbol{M}} \hat{\boldsymbol{f}}_{\boldsymbol{m}}(\mathbf{x}),$$

where $\boldsymbol{C}$ is the normalization constant such that $\hat{\boldsymbol{f}}_{\boldsymbol{m}}(\mathbf{x})$ integrates to unity.

Marzio and Taylor (2004, 2005) boosting algorithm for kernel density estimation takes the form of a multistep estimators whose first step is the standard kernel method. Boosting in kernel density estimation is a higher-order bias method that uses the basic kernel density estimator as its *weak learner* while bagging is a variance reduction technique in kernel density estimation with larger smoothing parameter.

Bandwidth selection in boosting and bagging

As generally known, in kernel density estimation methods, the right choice of the smoothing parameter must be made due to its importance in the process of estimation and much research has been done on smoothing parameter selectors. The rules for selecting smoothing parameters are generally based on the simple idea of balancing the asymptotic integrated squared bias and the asymptotic integrated variance globally (Sain, 2002).

Boosting and bagging in kernel density estimation are connected by using larger smoothing parameter because in kernel density estimation both methods are based on the principle of oversmoothing and appropriate "*weak learner"*. A complex learner is characterized by low bias and big variance (Marzio and Taylor, 2005; Ishiekwene, 2008) while a weaker learner is characterized by big bias and low variance. This means that a natural and direct approach for reducing the complexity of whatever kernel method is by *oversmoothing* because larger smoothing parameter increases the bias and reduces the variance (Marzio and Taylor, 2005).

In statistical terms, a strategy was devised by Marzio and Taylor (2005) as "*use very biased and low variance estimates by adopting larger smoothing parameters, then reduce the bias component using several boosting steps".* Bagging in kernel density estimation involves using larger smoothing parameter to reduce the variance term. The smoothing parameter can be described as a major determinant for boosting and bagging because oversmoothing weakens the learner thereby reducing the variance and with several boosting steps the bias will also be reduced which resulted in a reduction in the asymptotic mean integrated squared error (Marzio and Taylor, 2005).

Boosting and bagging aimed at reducing the bias and the variance term that resulted in a reduction in the asymptotic mean integrated squared error (AMISE). This reduction in the AMISE can easily be achieved by using large smoothing parameter(s) to reduce the variance term first, and then carry out some boosting steps to reduce the bias term which means bagging can be considered as a boosting algorithm in kernel

density estimation.

RESULTS

The study aims is to reduce the variance term and then carry out some boosting stepsin order to reduce the bias term. To achieve this, the study used the oversmoothed bandwidth for each of the data set considered because large smoothing parameter is needed for boosting and bagging to be effective and beneficial. The study calculated the variance, bias, and the asymptotic mean integrated squared error (AMISE) as seen in Tables 1 and 2. The first data set examined is the Annual Snowfall in Buffalo Scott (1992). The sample size of this data is 63.

Table 1. Analysis of bagging and boosting steps.

Analysis	Normal	1st boosting	2nd boosting
$Bias^2$	0.237633000	0.023185300	0.000782669
Variance	0.000386261	0.000386261	0.000386261
AMISE	0.238019261	0.023571561	0.001168930

Table 2. Analysis of bagging and boosting steps.

Analysis	Normal	1st boosting	2nd boosting
$Bias^2$	0.071910300	0.000550635	0.0000006147
Variance	0.000148035	0.000148035	0.0001480350
MISE	0.072058335	0.000698670	0.0001486497

The oversmoothed bandwidth for the Snowfall data ($n = 63$) is 11.5924. The same smoothing parameter was used for the various boosting steps, showing the same variance value in Table1 while the bias term is been progressively reduced. This smoothing parameter reduced the variance term (bagging) and with two boosting steps,the bias term was reduced and it resulted in a reduction in the asymptotic mean integrated squared error (AMISE).

Table 1 shows the analysis of the first and second boosting iterations with the result being a reduction in the bias and AMISE at last. The second data setof sample size 100 were simulated without reference to any distribution that is, they were randomly simulated real numbers. The oversmoothed bandwidth values are $h_x = 2.39108$ and$h_y = 2.24819$. The product kernel estimate using these smoothing parameter values is shown in Figure2 while the kernel estimates (surface plots and contour plots) of the "boosted" and "bagged" version of this data are shown in Figure 3 and Figure 4.

The kernel estimate (surface plot and contour plot) of the oversmoothed bandwidth shows clearly that the data are bimodal. This bimodality is obviously noticed even in the first and second boosting steps with the estimates being smoother. Table 2shows the analysis of the first and second boosting iterations with the result being a reduction in the bias and *AMISE* with the variance remaining unchanged in each of the boosting steps because the same smoothing para-

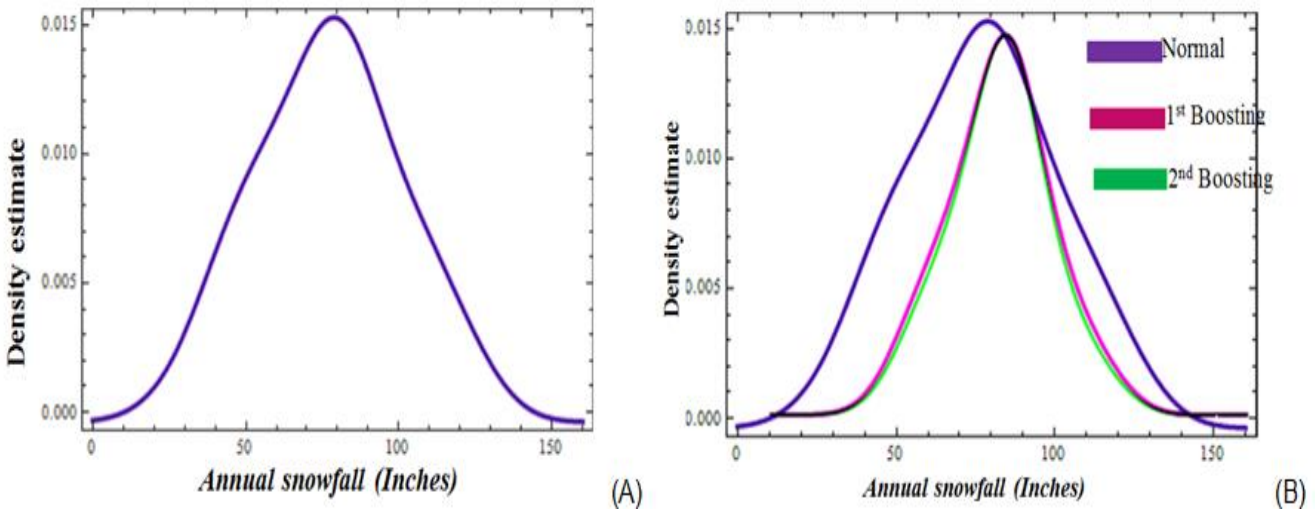

Figure 1. (A) Kernel estimate of oversmoothed bandwidth; (B) Estimates of oversmoothed bandwidth with 1st and 2nd boosting steps.

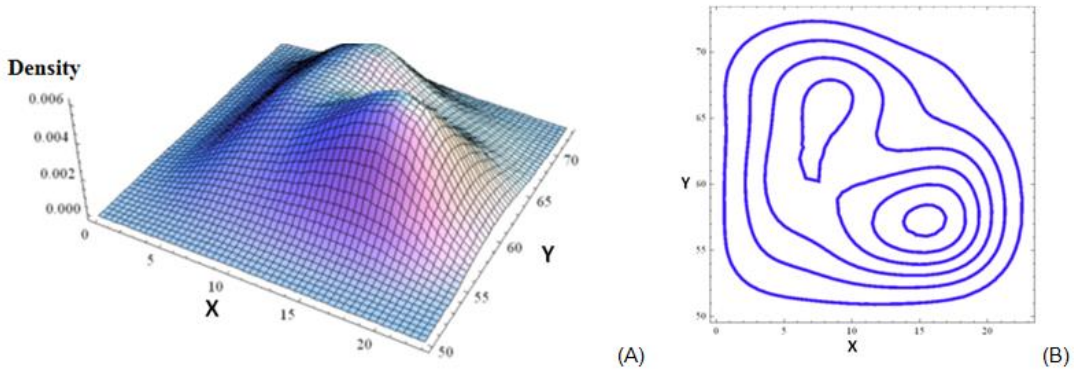

Figure 2. (A)Surface plot of oversmoothed bandwidth; (B) Contour plot of oversmoothed bandwidth.

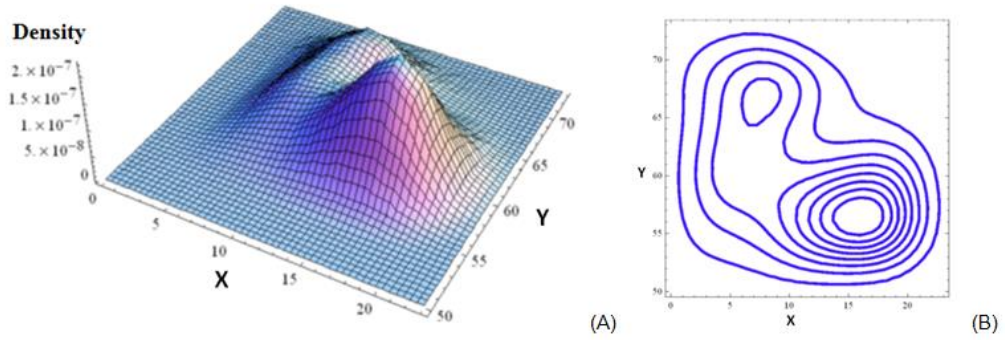

Figure 3. (A) Surface plot of first boosting step; (B) Contour plot of first boosting step.

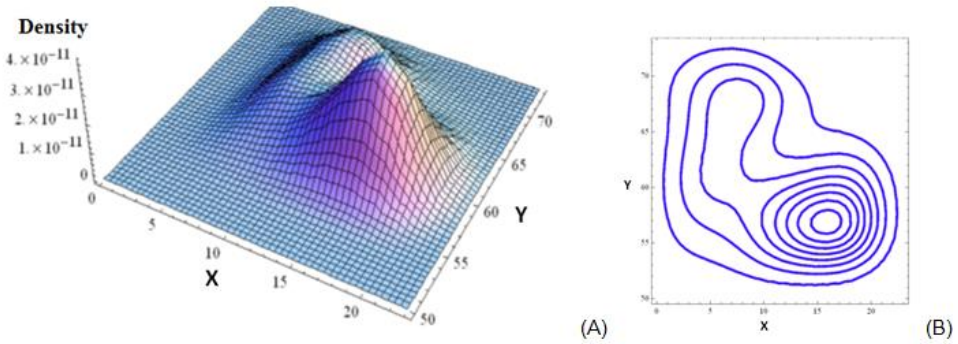

Figure 4. (A) Surface plot of second boosting step; (B) Contour plot of second boosting step.

meter was applied. This also confirms that boosting is a bias reduction while bagging is a variance reduction method. Table 3 shows the simulated data.

Conclusions

Boosting and bagging in kernel density estimation are bias, and variance reduction techniques characterized by using larger smoothing parameter(s) suggested a powerful new tool for addressing the *curse of dimensionality* effects (Marzio and Taylor, 2004). Since large bandwidths reduce the variance term of the AMISE and increases the bias term, the study carried out two boosting steps to demonstrate the reduction of the bias and variance term that translated to a reduction in the AMISE using the oversmoothed bandwidth. As can be seen from Figures 1 to 4 and Tables 1 and 2, bagging can be considered as a boosting algorithm in kernel density estimation since both methods

Table 3. Sets of simulated data N=100.

X	9.250	19.92	14.94	7.440	17.74	17.51	7.574	11.63	19.11	10.56	19.50	19.06	17.95	7.067	6.470	15.75	8.050	9.375	13.39	19.72
Y	58.74	57.18	66.94	63.56	58.41	59.15	60.40	67.59	61.88	67.32	62.39	60.43	56.72	62.65	60.29	68.30	58.54	62.88	67.31	68.61
X	19.74	14.84	11.53	11.36	15.30	7.234	9.002	8.477	18.21	19.72	9.286	5.180	13.12	13.67	15.72	16.04	11.55	6.143	5.650	13.04
Y	60.50	60.23	68.46	66.67	61.83	69.85	65.34	69.97	58.51	69.25	57.36	58.57	56.17	67.11	58.42	61.50	55.11	57.40	57.44	56.73
X	12.90	5.199	13.82	10.22	16.81	17.74	8.014	9.086	6.617	14.39	6.216	10.30	9.835	5.057	14.40	10.06	18.44	15.96	13.82	16.50
Y	56.48	63.43	69.92	60.40	64.44	62.38	65.76	69.33	58.12	60.23	57.29	60.81	67.94	60.39	67.95	58.97	57.81	59.52	56.92	60.84
X	16.78	12.37	8.100	11.38	18.92	6.775	19.17	10.77	13.97	9.803	11.59	5.894	18.23	16.15	18.56	7.572	11.67	10.39	7.013	12.78
Y	59.36	64.19	57.38	63.50	66.57	59.20	66.06	57.30	60.57	64.21	68.26	58.43	62.75	57.23	65.87	66.02	62.57	67.77	60.83	68.60
X	16.38	10.16	9.600	18.63	6.699	17.38	18.86	12.30	5.011	10.09	18.82	19.40	16.75	16.83	14.95	6.522	9.117	10.31	14.24	12.11
Y	58.75	55.75	69.36	58.01	61.17	69.10	64.43	60.10	64.37	63.72	57.79	62.63	63.03	58.65	60.96	59.62	63.43	55.05	67.54	61.67

used larger smoothing parameter aimed at reducing the AMISE. Although boosting and bagging depend on larger smoothing parameter but we have demonstrated that their targets are different in terms of their contribution to the AMISE.

REFERENCES

Breiman, L. (2001). Using Iterated Bagging to Debias Regressions.Machine Learning. **45(3):**261–277.

Breiman, L. (1998). Arcing Classifiers. AS **26**(3)**:**801–849.

Breiman, L. (1996). Bagging Predictors. Machine Learning.**26:**123–140.

Buhlmann, P. and Yu, B. (2003). Boosting With the L_2 Loss: Regression and Classification. Journal of the American Statistical Association. **98:**324–339.

Freund, Y. and Schapire, R. (1995).A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.European Conference on Computational Learning Theory. **Pp.** 23–37.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive Logistics Regression: A statistical View of Boosting (With Discussion). Annals of Statistics. **28(2):**337–374.

Gey, S. and Poggi, J. M. (2006). Boosting and Instability for Regression Trees. Computational Statistics and Data Analysis. **50:**533–550.

Ishiekwene, C. C. (2008). Bias Reduction Techniques in KDE. An Unpublished Ph.D Thesis Submitted to the School of Postgraduate Studies, University of Benin, Benin City, Nigeria.

Marzio, D.M. and Taylor, C.C (2004)."Boosting Kernel Density Estimates: A Bias Reduction Technique? " Biometrika **91:**226–233.

Marzio, D.M. and Taylor, C.C (2005). "On Boosting Kernel Density Methods for Multivariate Data: Density Estimation and Classification", Statistical Methods and Applications. **14:**163–178.

Mason, L., Baxter, J., Bartlett, P. and Frean, M. (1999). Boosting Algorithms as Gradient Descent.Neural Information Processing Systems, Vol. 12.

Ridgeway, G. (2002). Looking for Lumps: Boosting and Bagging for Density Estimation. Comput.Stat. Data Anal.38(4):379–392.

Rosset, S. (2003). Topics in Regularization and Boosting. A Dissertation Submitted to the Department of Statistics and the Committee on Graduate Studies of Stanford University.

Rosset, S., Zhu, J. and Hastie, T. (2004). Boosting as a Regularized Path to a Maximum Margin Classifier. Journal of Machine Learning Research. **5:**941–973.

Sain, R.S. (2002). Multivariate Locally Adaptive Density Estimation. Computational Statistics and Data Analysis. **39:**165–186.

Scott, D.W. (1992). Multivariate Density Estimation. Theory, Practice and Visualisation. Wiley, New York.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.