Optimizing Random Forest from the pondering of regression tree leaves

Caio Ponte[1], **Carlos Caminha**[2], **Vasco Furtado**[1]

[1] Programa de Pós Graduação em Informática Aplicada, Universidade de Fortaleza
Fortaleza – CE – Brazil

[2]Laboratório de Ciência de Dados e Inteligência Artificial (LCDIA), Universidade de Fortaleza
Fortaleza – CE – Brazil

caioponte@edu.unifor.br, {caminha,vasco}@unifor.br

Abstract. *Random Forest is a popular and effective algorithm to solve classification and regression problems. The Random Forest predictions are made considering that each tree has an equal contribution to the final result. This work proposes a new method of weighting regression trees in order to improve the model performance. Our strategy is motivated to use statistical dispersion measures, such as standard deviation or standard error of the mean, as indicators of the quality of the prediction in the leaves. The proposed weighting strategy has been compared with other weighting methods. In this comparison, it was observed that it reduced the Mean Absolute Error in about 30 % of the studied datasets.*

Resumo. *Floresta Aleatória é um algoritmo popular e efetivo na resolução de problemas de classificação e regressão. As predições de uma Floresta Aleatória são feitas considerando que cada árvore possui igual contribuição no resultado final. Este trabalho propõe um novo método de ponderação de árvores de regressão com o objetivo de melhorar o poder de predição do modelo. Nossa estratégia é motivada em utilizar medidas de dispersão estatística, como desvio padrão ou erro padrão da média, como indicadores de qualidade da predição na folha. A estratégia de ponderação proposta foi comparada com outros métodos de ponderação. Nessa comparação observou-se que a mesma reduziu o Erro Absoluto Médio em cerca de 30% dos conjuntos de dados estudados.*

1. Introdução

Florestas aleatórias *(Random Forest - RF)* é um método de aprendizagem de máquina largamente utilizado atualmente em problemas de regressão e classificação. Ele se baseia na geração de florestas, compostas, cada uma, por várias árvores de decisão que fazem, juntas, predições melhores do que uma única árvore [Dietterich et al. 2002]. A intuição por trás de métodos desse tipo é de que a diversidade representada por várias árvores montadas com seleção aleatória de atributos e de exemplos (para evitar que as árvores sejam correlacionadas entre si), reduz sobreajuste dos dados de treinamento *(overfit)* [Sagi and Rokach 2018]. Tipicamente, cada árvore que forma uma floresta tem igual colaboração na predição final. Quando se trata de classificação a predição final é decidida por voto majoritário. Para regressão, a decisão final é uma média entre as decisões individuais.

Nossa pesquisa investiga novas formas de capturar a participação individual das árvores de decisão em RF para problemas de regressão. Tal preocupação é abundante na literatura quando se trata de problemas de classificação, [Amaratunga et al. 2008, Kim et al. 2011, Tsymbal et al. 2006], mas rara em problemas de regressão. O processo de predição de valor de uma variável alvo para uma árvore de regressão é feito com base na média dos valores desta variável alvo de todos os exemplos de treinamento que estão representados por uma determinada folha da árvore. Portanto, a qualidade da predição de uma árvore dependerá, da qualidade de predição de suas folhas. Por sua vez, a qualidade de predição das folhas depende das distribuição dos valores da variável alvo. Caso esses valores se agrupem em torno de um valor médio, devem levar à predições mais corretas. No entanto, valores alvos muito heterogêneos podem aumentar o erro do modelo. Lança-se aqui uma hipótese que a utilização de medidas estatísticas de dispersão dos valores alvos pode ser útil para melhorar o poder de predição das árvores e consequentemente das florestas.

Propõem-se utilizar medidas de dispersão que sirvam como base para a ponderação das árvores de decisão de uma *RF*, de tal maneira que, quanto maior a heterogeneidade dos valores alvos menor o peso aquela predição tem na combinação para a resposta final da floresta. Dessa forma, no processo de treinamento associará um valor de dispersão, que pode ser, por exemplo, o desvio padrão da variável alvo, para cada folha. Quando passado uma nova amostra de teste, resgatará o valor da dispersão da folha de cada árvore dessa amostra, aplicando uma combinação linear, na qual os pesos são inversamente proporcionais ao valor do desvio padrão, por exemplo. A intuição sobre a dispersão citada na hipótese também pode abordar estratégias de como agregar a predição na própria folha, de tal maneira, a ajudar a alcançar o objetivo de aprimorar o poder de predição da floresta. Além disso, pode acontecer de que a distribuição de predições tenha uma assimetria muito forte (alto *skewness*). Nesse caso, considerar a mediana e/ou a moda poderia ser mais adequado como medida de centralidade do que a média.

Com o intuito de compreender o impacto dessas questões, realizou-se experimentos para testar as suposições feitas, em 31 conjuntos de dados. Nesses experimentos, comparou-se a estratégia proposta neste trabalho com outros métodos encontrados na literatura para ponderação das predições das árvores de decisões. Essas metodologias, foram adaptadas para abordar problemas de regressão, uma vez que foram criadas para classificação. Os resultados mostraram que ao considerar medidas de dispersão como ponderadores das predições finais, o Erro Absoluto Médio (Mean Absolute Error - MAE) dos regressores foi reduzido em cerca de 30% dos conjuntos de dados avaliados. Mesmo quando melhorias não foram observadas, o MAE não aumentou para nenhum conjuntos de dados, indicando que a estratégia proposta tem robustez para ser aplicada sem maiores prejuízos.

2. Fundamentação Teórica

O *RF* é um *ensemble* constituído por L árvores de decisão que farão previsões que serão combinadas para compor o resultado final do modelo. O *RF* possui algumas característica desejáveis como: é tão rápido quanto outros algoritmos de *Bagging* ou *Boosting*; possui desempenho competitivo com outros métodos; é relativamente robusto a ruídos ou *outliers*; e é facilmente paralelizável [Breiman 2001]. *RF* possui dois processos aleatórios para a construção da floresta. O primeiro é a criação de conjuntos de dados *bootstrap*,

que consiste de uma técnica de amostragem de dados. Considere o conjunto de dados de treinamento $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, onde x_i é um vetor de atributos da amostra i que possui dimensionalidade n, y_i é valor alvo que se deseja prever e m é número de amostras de treino de D. No método *bootstrap*, seleciona-se aleatoriamente amostras de D, de tal forma a construir um novo conjunto de dado B_j do mesmo tamanho de D. Essa amostragem é feita com reposição, ou seja, quando é selecionada uma nova amostra ela já pode ter sido selecionada antes. Isso fará com que essas instância com repetição tenha maior peso no processo de treinamento de uma determinada árvore de decisão.

É importante destacar que existirão amostras fora do conjunto B_j, denominadas amostras *Out-of-Bag (OOB)*. No *RF*, o procedimento *bootstrap* é realizado L vezes, um para cada árvore criada, resultando os conjuntos $B_1, B_2, ..., B_L$. O segundo procedimento aleatório é o Subconjunto Aleatório de Atributos [Ho 1998], que altera o processo de treinamento de uma árvore de decisão, de tal maneira que, toda vez que cria-se uma novo nó na árvore, seleciona-se aleatoriamente um subconjunto dos atributos para serem candidatos na escolha do atributo daquele nó. Normalmente escolhe-se o valor de $\sqrt{n}$ para ser o tamanho do subconjunto de atributos a serem selecionados aleatoriamente.

3. Trabalhos Relacionados

Existem diversos trabalhos na literatura que lidam com formas de melhorar o desempenho de *ensembles* através de ponderações. Por exemplo, ponderações de classes em problemas de classificações para conjuntos de dados com classes não balanceados ou mesmo ponderações de atributos modificando o método de Subconjunto Aleatório de Atributos, de modo a considerar atributos mais relevantes na predição como um *preferential attachment* [Amaratunga et al. 2008]. Em [Kim et al. 2011] é proposto um método de ponderação tanto de instâncias quanto de preditores, de tal forma que instâncias difíceis de acertar terão peso maior e que preditores que acertam tais instâncias difíceis também terão pesos maiores.

No entanto, este trabalho foca em estratégias que tratam exclusivamente de ponderação no processo de combinação das árvores de uma *RF*. O processo de combinação dos regressores para compor a predição final da floresta pode ser divido como estático ou dinâmico [Puuronen et al. 1999]. Na combinação estática, os pesos atribuídos para cada árvore da *Random Forest* são calculados uma única vez e, portanto, são sempre fixos.

A estratégia de combinação padrão do algoritmo *RF* considera que todas as árvores têm igual contribuição para o resultado final da floresta, deste modo é considerada uma estratégia estática. Um outro exemplo seria considerar alguma medida de desempenho, como acurácia (para classificação) ou erro absoluto médio (para regressão), como peso da árvore, de tal forma que, quanto maior o acurácia mais aquele modelo terá impacto na predição final das instâncias de teste [Li et al. 2010]. Nessa última estratégia, pode-se adaptar, por exemplo, a medida de erro sobre o conjunto de dados *OOB* de cada árvore, para um problema de regressão. O peso de cada árvore j pode ser definido como $w_j = 1/E_j(X^j_{OOB})$, na qual $E_j(X^j_{OOB})$ é o erro médio da árvore j no conjunto OOB dessa árvore. Esse método será nomeado neste trabalho de $WOOB$.

Já no modo dinâmico, o peso de uma predição do modelo leva em conta uma

instância de teste. A ponderação descrita em [Tsymbal et al. 2006], chamada *Dynamic Integration (DI)*, leva em conta a instância de teste para encontrar seus k vizinhos mais próximos. A partir desses k vizinhos, é possível verificar como uma árvore desempenha realizando a predição desses vizinhos que também pertencem ao conjunto *OOB* e, portanto, são amostras não vistas no treinamento para aquela árvore. *Dynamic Integration* foi definido inicialmente para problema de classificação, mas em [Rooney et al. 2004] são realizados experimentos para regressão. Para fins de comparações, nesse trabalho define-se os pesos de *DI* da seguinte forma: dado uma nova instância de teste x, o peso da predição da árvore j para tal instancia é igual a

$$w_j(x) = \frac{\sum_{i=1}^{k} I_{OOB_j}(x_i) \cdot \sigma(x, x_i) \cdot E_j(x_i)}{\sum_{i=1}^{k} I_{OOB_j}(x_i) \cdot \sigma(x, x_i)} \tag{1}$$

onde OOB_j é o conjunto *Out-of-Bag* da árvore j; k é o número de vizinhos no conjunto OOB_j; $I()$ é uma função indicadora, ou seja, $I_{OOB_j}(x_i)$ retornará 1 se $x_i \in OOB_j$ ou 0, caso contrário; cada x_i representa os atributos de cada um dos k vizinhos; $\sigma(x, x_i)$ é um coeficiente baseado em distância e $E_j(x_i)$ é o erro da instância x_i do conjunto OOB_j da árvore j.

Em [Kleiman] é proposto uma outra forma de ponderação dinâmica em *Random Forest* que estende o conceito de DI, nomeada como *Instance-based Out-of-Bag Weighting (IOOBW)*. Nessa estratégia, não se realiza o *k-NN* nem se calcula o coeficiente $\sigma(x, x_i)$. Em vez disso, é considerado uma medida de desempenho das amostras do conjunto *OOB*, dentro da folha da árvore que a instância caiu. Como essa estratégia é construída baseado em problemas de classificação, será adaptado aqui para problemas regressão, de forma similar ao *DI*. Em *IOOBW* o peso da predição de uma determinada árvore do *ensemble* é igual a

$$w_j(x) = \frac{\sum_{i=1}^{l} I_{OOB_j}(x_i) \cdot I_{Leaf_i}(x, x_i) \cdot E_j(x_i)}{\sum_{i=1}^{l} I_{OOB_j}(x_i) \cdot I_{Leaf_i}(x, x_i)} \tag{2}$$

onde $Leaf_i(x, x_i)$ é o conjunto das amostras OOB_j que pertencem à mesma folha que x na árvore j e l é número de instância desse conjunto.

4. Estratégia de Ponderação Proposta

Neste artigo é proposta uma nova abordagem de ponderação dinâmica paras as predições dos regressores de uma *RF*. Essa estratégia utiliza as informações presentes na folha, mas diferentemente das estratégias citadas nos Trabalhos Relacionados, esta ponderação não leva em conta nenhuma medida de acurácia. Em vez disso serão utilizadas medidas estatísticas para mensurar a heterogeneidade dos valores alvos das amostras de treinamento dentro de um nó folha para ponderar as predições das árvores. Pode-se utilizar, por exemplo, o desvio padrão (*STD*) ou erro padrão da média (*SEM*) como medidas de dispersões estatísticas.

A intuição dessa estratégia é utilizar a dispersão dos dados para saber o nível de heterogeneidade dos valores alvos dentro de uma folha. Essa estratégia é aplicada para problemas de regressão, na qual as predições de uma árvore de decisão é, normalmente, realizada agregando os valores alvos via uma média aritmética. Pode-se perceber que, se

a distribuição das variáveis alvos das amostras utilizadas no treinamento for Gaussiana (principalmente com desvio padrão baixo), aquela folha irá predizer uma valor confiável, uma vez que essas amostras devam ser muito parecidas devido a sua baixa dispersão. Isso se aproxima do conceito de pureza, presente nos algoritmos de árvores de classificação. A ponderação da predição da árvore, através do desvio padrão, por exemplo, dada uma amostra x, é igual a

$$w_j(x) = \sqrt{\frac{\sum_{i=1}^{l} I_{IOB_j}(x_i) \cdot I_{Leaf_i}(x, x_i) \cdot (y_i - \bar{y})^2}{\sum_{i=1}^{l} I_{IOB_j}(x_i) \cdot I_{Leaf_i}(x, x_i)}} \tag{3}$$

onde IOB_j é o conjunto *In-of-Bag*, ou seja, os dados de treinamentos da árvore j, y_i é o valor alvo da amostra i que a instância x caiu e $\bar{y}$ é a média de todas os valores alvos y_i. Sumarizando, o peso da predição é tomado com o desvio padrão dos valores alvos dentro da folha que a amostra de teste caiu, $std(Y_{Leaf})$, como ilustrado na Figura 1, na qual mostra um exemplo da predição de uma amostra de teste x_0. Cada uma das L árvores irá indicar o valor da predição $\hat{y}_1, \hat{y}_2, ..., \hat{y}_L$ e será calculado o valor do peso (w_i) de cada predição baseado no desvio padrão dos valores alvos das B amostras de treinamento na qual a amostra x_0 caiu. Por fim, realizará uma combinação linear entre entre as predições e os pesos de todas as árvores para obter o resultado final do modelo para dada a amostra x_0.

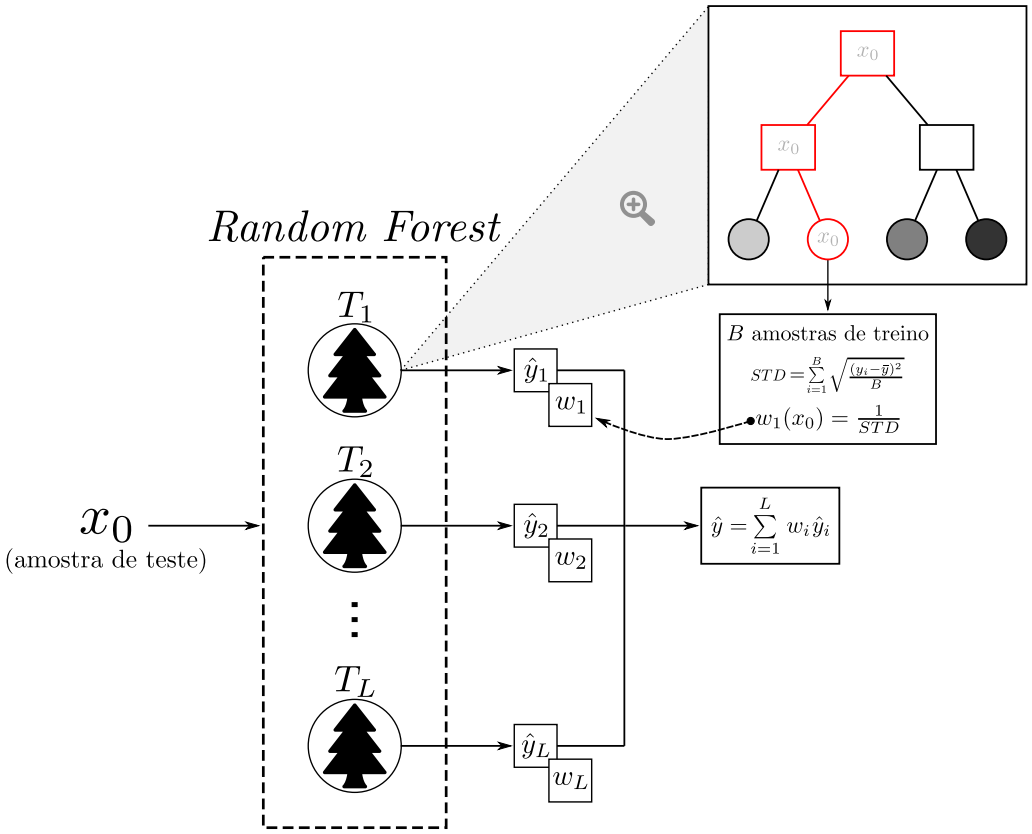

Figura 1. Exemplo da ponderação no processo de predição.

Adicionalmente, serão experimentadas outras formas de agregar o resultado dado pela folha de uma árvore de decisão. O padrão do algoritmo *CART*, por exemplo, para

Tabela 1. Descrição dos conjuntos de dados. A variável $\mathrm{m_1}$ representa o número de instâncias do conjunto de dados, $\mathrm{n_{num}}$ representa o número de atributos numéricos e $\mathrm{n_{cat}}$ o número de atributos categóricos.

Conj. de Dados	m	n_{num}	n_{cat}	*Conj. de Dados*	m	n_{num}	n_{cat}
autoHorse	159	17	8	autoMpg	392	4	3
autoPrice	159	15	0	bodyfat	252	14	0
bostonHousing	506	12	1	breastTumor	286	1	8
cholesterol	299	6	7	cloud	108	4	2
cpu	209	6	1	echoMonths	106	6	3
fishcatch	158	5	2	fruitfly	125	2	2
lowbwt	189	2	7	meta	264	19	2
pbc	276	10	8	pharynx	195	1	10
pwLinear	200	10	0	quake	2178	3	0
sensory	576	0	11	servo	167	0	4
strike	625	5	1	veteran	137	3	4
UJIIndoorLoc Lat	1115	521	0	UJIIndoorLoc Lon	1115	521	0
house price (kaggle)	2919	36	43	abalone	4177	7	1
power	9568	4	0	pumadyn8	8192	8	0
pumadyn32	8192	32	0	ailerons	13750	40	0
crime	1993	100	0				

indução de árvores de regressão é a média dos valores da folha [Breiman et al. 1984]. Além da média aritmética, outras medidas, como moda e mediana, podem ser avaliadas. Se faz importante esse tipo de análise, pois pode-se pensar em diversas distribuições que os valores alvos podem tomar dentro de um nó folha. Em distribuições muito próximas de Gaussianas, a média é uma boa escolha. No entanto, quando se tem uma distribuição Gaussiana com uma assimetria muito forte, ou seja, com o *skewness* alto, a média pode não ser a melhor escolha, sendo necessário considerar medidas como mediana ou a moda, que podem tornar o modelo menos vulnerável as predições equivocadas.

5. Metodologia

Nesta seção será detalhada a metodologia para realizar os experimentos a fim de comparar as estratégias de ponderação citadas anteriormente. Nos experimentos foram utilizados a implementação do *Random Forest* para regressão baseada na biblioteca de aprendizado de máquina *Scikit-learn* [1]. Ao todo, foram utilizados 31 conjuntos de dados de regressão, obtidos de alguns repositórios, como: *UCI machine learning repository* [2], *Weka Wiki repository* [3], *Delve* [4] e *Kaggle* [5]. Esses conjuntos de dados são reais e simulados, possuindo um grande espectro no que tange ao número de instâncias e número de atributos. A Tabela 1 sumariza os detalhes de cada um desses dados.

Em relação aos parâmetros da *RF*, foram utilizadas, como modelo base do *ensemble*, árvores de decisões treinadas a partir do algoritmo *CART* (padrão da biblioteca *Scikit-learn*). Em todos os experimentos, a *RF* tinha 200 árvores de decisão, possuindo

[1]https://scikit-learn.org/

[2]https://archive.ics.uci.edu/

[3]https://waikato.github.io/weka-wiki/datasets/

[4]http://www.cs.toronto.edu/ delve/data/datasets.html

[5]https://www.kaggle.com/

uma profundidade máxima de 20. Foi escolhido o valor de $\sqrt{n}$ (n é o número de atributos do conjunto de dados), como sendo o valor de atributos candidatos na criação de um novo nó da árvore. Um ponto importante para a metodologia proposta nesse trabalho, é que deva existir uma quantidade razoável de amostras dentro de uma folha. Se for tomado o desvio padrão como a medida de dispersão, por exemplo, de poucos valores dentro da folha, pouca relevância essa informação tem. Por isso, foi escolhido como o número mínimo de amostras dentro de uma folha o valor de 20 amostras.

Relativo ao treinamento da *RF*, utilizou-se *k-fold cross validations*, no qual o número de *folds* implicará no número de execuções diferentes para cada experimento. Escolheu-se 20 como sendo o número de *folds* a ser utilizado em vez do padrão de 10 *folds*, pois assim é possível obter mais execuções. As variáveis categóricas foram modeladas como *One Hot Encoder*. Relativo ao processo de predição, no qual as estratégias de ponderação estão inclusas, o método proposto utilizou três medidas de dispersão: o desvio padrão (*STD*), erro padrão da média (*SEM*) e desvio padrão da moda (*STD* Moda). Esse último, altera a fórmula padrão do *STD*, substituindo a diferença entre cada elemento e a média para a diferença entre cada elemento e a moda. Isso porque testou-se também uma variação da forma de agregação dos valores alvos das amostras dentro de uma folha, considerando, em vez da média (comumente usada em *RF*), o valor da moda. Dessa forma teria mais sentido utilizar a agregação, via moda, a nível de folha com ponderação *STD* da Moda.

Para *DI*, utilizou-se o valor 10 vizinhos dentro do conjunto *OOB* de cada árvore e o coeficiente de distância (σ) baseado na distância Euclidiana. Nenhum parâmetro adicional é necessário para WOOB e IOOBW. Como medida de erro, foi utilizado o erro absoluto médio (*MAE*).

6. Resultados

A Tabela 2, ilustra os resultados das 20 execuções de cada um dos *folds*, indicando a média do *MAE* $\pm$ erro padrão da média (*SEM*) do *MAE*. Dentre os 20 valores dos erros para cada conjunto de dados foram retirados os valores extremos, considerando apenas os valores dentro do intervalo LS (limite superior) e LI (limite inferior), conhecidos como *inliers*. O limite superior é definido como $LS = Q_3 + 1.5 * AIQ$, e o limite inferior como $LI = Q_1 - 1.5 * AIQ$, no qual Q_1 e Q_3 são o primeiro e o terceiro quartil, respectivamente, e $AIQ = Q_3 - Q_1$ representa a amplitude interquartil. Esse é o método padrão utilizado em Diagramas de Caixas na detecção de *outliers*. Os erros destacados em negrito na Tabela 2 indicam quais são as estratégias vencedoras, considerando a média do *MAE* dos valores dos erros *inliers* $\pm$ *SEM*. Quando nenhum elemento na linha estiver em negrito é porque todas as estratégias obtiveram resultados empatados.

Tabela 2. Comparações dos resultados das *RF* para diversas ponderações. Os valores da Tabela representam o erro absoluto médio (*MAE*). Os valores que estão destacados foram as estratégias com menores erros. Caso nenhuma valor, na linha, estiver destacado é porque todos os modelos empataram.

Conj. de Dados	*RF*	RF Média Pond. STD	RF Moda Pond. STD Moda	RF Média Pond. SEM	RF DI	RF IOOBW	RF WOOB
autoMpg	2.56 ± 0.19	2.50 ± 0.20	2.48 ± 0.27	2.50 ± 0.20	2.63 ± 0.19	2.51 ± 0.19	2.56 ± 0.19
bostonHousing	2.98 ± 0.27	**2.79 ± 0.32**	**2.38 ± 0.24**	**2.81 ± 0.33**	3.13 ± 0.28	**2.59 ± 0.26**	2.96 ± 0.27
cholesterol	36.12 ± 1.06	36.63 ± 1.20	37.15 ± 1.37	36.61 ± 1.21	36.15 ± 1.02	37.03 ± 1.21	36.11 ± 1.06
breastTumor	7.85 ± 0.37	7.85 ± 0.37	7.92 ± 0.35	7.85 ± 0.37	7.86 ± 0.37	7.85 ± 0.37	7.84 ± 0.37
pbc	776.35 ± 69.82	766.73 ± 73.23	719.21 ± 72.40	766.88 ± 73.67	778.96 ± 69.54	774.72 ± 71.42	775.95 ± 69.80
quake	0.15 ± 0.00	0.15 ± 0.00	0.15 ± 0.00	0.15 ± 0.00	0.15 ± 0.00	0.15 ± 0.00	0.15 ± 0.00
sensory	0.65 ± 0.03	0.65 ± 0.03	0.66 ± 0.03	0.65 ± 0.03	0.65 ± 0.03	0.65 ± 0.03	0.65 ± 0.03
meta	67.93 ± 11.84	**49.47 ± 9.67**	**42.49 ± 10.41**	**49.14 ± 9.63**	209.17 ± 49.59	**55.27 ± 11.51**	67.48 ± 11.82
strike	230.54 ± 22.81	**178.53 ± 23.38**	**207.06 ± 33.04**	**178.30 ± 23.29**	231.85 ± 19.57	**181.10 ± 22.61**	230.56 ± 22.83
autoHorse	**7.87 ± 0.71**	**7.49 ± 0.51**	8.54 ± 0.27	**7.55 ± 0.50**	10.08 ± 0.97	**7.26 ± 0.48**	**7.68 ± 0.69**
autoPrice	**1463.84 ± 165.38**	**1485.91 ± 200.47**	**1494.02 ± 228.23**	**1491.30 ± 202.45**	1879.78 ± 251.41	**1488.07 ± 197.00**	**1464.14 ± 162.61**
bodyfat	2.65 ± 0.10	**2.22 ± 0.10**	**2.00 ± 0.14**	**2.23 ± 0.09**	2.89 ± 0.13	**2.18 ± 0.07**	2.44 ± 0.09
cloud	**0.31 ± 0.02**	**0.28 ± 0.03**	0.39 ± 0.04	**0.28 ± 0.03**	0.34 ± 0.02	**0.27 ± 0.03**	**0.30 ± 0.02**
cpu	25.50 ± 2.19	**11.29 ± 1.23**	18.40 ± 2.33	**11.72 ± 1.23**	48.49 ± 3.55	**11.79 ± 1.20**	23.90 ± 2.68
echoMonths	9.63 ± 0.67	9.23 ± 0.75	9.16 ± 0.99	9.19 ± 0.76	10.51 ± 0.73	9.68 ± 0.82	9.55 ± 0.69
fishcatch	**81.57 ± 14.51**	**62.37 ± 12.47**	**74.48 ± 14.41**	**62.56 ± 12.39**	137.30 ± 22.14	**60.10 ± 12.12**	**74.67 ± 13.57**
fruitfly	11.69 ± 0.88	11.53 ± 0.90	11.14 ± 0.90	11.53 ± 0.90	11.68 ± 0.88	11.64 ± 0.88	11.69 ± 0.87
lowbwt	339.95 ± 52.86	340.86 ± 49.83	365.19 ± 49.81	342.13 ± 49.34	374.33 ± 53.42	340.16 ± 48.34	338.38 ± 50.42
abalone	1.51 ± 0.10	1.46 ± 0.13	1.59 ± 0.19	1.46 ± 0.13	1.52 ± 0.10	1.48 ± 0.12	1.51 ± 0.10
ailerons	1.49 ± 0.04	**1.44 ± 0.04**	**1.35 ± 0.06**	**1.44 ± 0.04**	1.50 ± 0.04	**1.40 ± 0.05**	1.48 ± 0.04
UJIIndoorLoc Lat	13.44 ± 0.30	11.01 ± 0.27	**9.39 ± 0.28**	10.98 ± 0.27	15.43 ± 0.33	10.97 ± 0.27	13.37 ± 0.29
UJIIndoorLoc Lon	16.48 ± 0.41	11.92 ± 0.32	**9.28 ± 0.25**	11.50 ± 0.26	20.15 ± 0.31	12.00 ± 0.32	16.29 ± 0.41
house price (kaggle)	28638.13 ± 333.13	28498.63 ± 349.72	28923.86 ± 294.96	28471.71 ± 356.79	28746.50 ± 338.11	28522.80 ± 349.71	28636.98 ± 333.28
power	2.60 ± 0.03	**2.51 ± 0.02**	**2.48 ± 0.02**	**2.51 ± 0.02**	2.70 ± 0.02	**2.51 ± 0.02**	2.59 ± 0.03
pumadyn8	1.45 ± 0.01	1.12 ± 0.00	**1.03 ± 0.01**	1.12 ± 0.00	1.58 ± 0.01	1.11 ± 0.00	1.42 ± 0.01
pumadyn32	0.02 ± 0.00	**0.01 ± 0.00**	**0.01 ± 0.00**	**0.01 ± 0.00**	0.02 ± 0.00	**0.01 ± 0.00**	**0.01 ± 0.00**
crime	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00
pharynx	274.41 ± 17.55	262.92 ± 18.42	259.57 ± 21.37	262.38 ± 18.54	283.25 ± 17.48	263.10 ± 18.27	269.04 ± 17.66
pwLinear	**1.94 ± 0.08**	**1.86 ± 0.08**	**1.91 ± 0.08**	**1.92 ± 0.09**	2.15 ± 0.09	**1.90 ± 0.09**	**1.89 ± 0.08**
servo	**0.73 ± 0.05**	**0.62 ± 0.06**	**0.65 ± 0.07**	**0.62 ± 0.06**	0.88 ± 0.04	**0.62 ± 0.06**	**0.68 ± 0.05**
veteran	82.88 ± 7.91	72.19 ± 9.17	74.31 ± 12.76	72.07 ± 9.14	84.75 ± 7.60	74.80 ± 8.84	82.95 ± 7.95

Pode-se verificar pela Tabela 2, em uma visão geral, que a estratégia proposta não piora o resultado para nenhum conjunto de dados, o que também acontece, na maior parte do tempo, para as outras estratégias de ponderação. De modo geral, as estratégias são equivalentes entre si, nas quais existem conjuntos de dados que todos os métodos possuem resultados empatados (autoMpg, cholesterol, breastTumor, pbc, quake, sensory, echoMonths, fruitfly, lowbwt, abalone, house price kaggle, crime, pharynx, veteran) quando considerado os intervalos de $\pm SEM$. Isso representa 14 dos 31 conjuntos de dados. Os resultados de *IOOBW* se assemelham com os resultados da ponderação proposta nesse trabalho. Em 14 conjuntos de dados (bostonHousing, meta, strike, autoHorse, autoPrice, bodyfat, cloud, cpu, fishcatch, ailerons, power, pumadyn32, pwLinear, servo), *IOOBW* empata com o a ponderação proposta aqui. Existem 11 conjuntos de dados que a estratégia de ponderação proposta aqui, diminui o erro em relação ao *RF* padrão (bostonHousing, meta, strike, bodyfat, cpu, ailerons, UJIIndoorLoc Lat, UJIIndoorLoc Lon, power, pumadyn8, pumadyn32). Desses 11, existem três conjuntos de dados (UJIIndoorLoc Lat, UJIIndoorLoc Lon, pumadyn8) que a estratégia *RF Moda Pond. STD Moda* se destacou e foi a melhor abordagem dentre todas os tipos de ponderações. Os três conjuntos de dados têm apenas variáveis numéricas e uma quantidade relativamente grande de instâncias. Vale ressaltar que, a adaptação aqui proposta para o método *DI* não obteve resultados satisfatórios, uma vez que ele piorou em seis conjuntos de dados se comparado com *RF* padrão (autoHorse, autoPrice, cloud, fishcatch, pwLinear, servo).

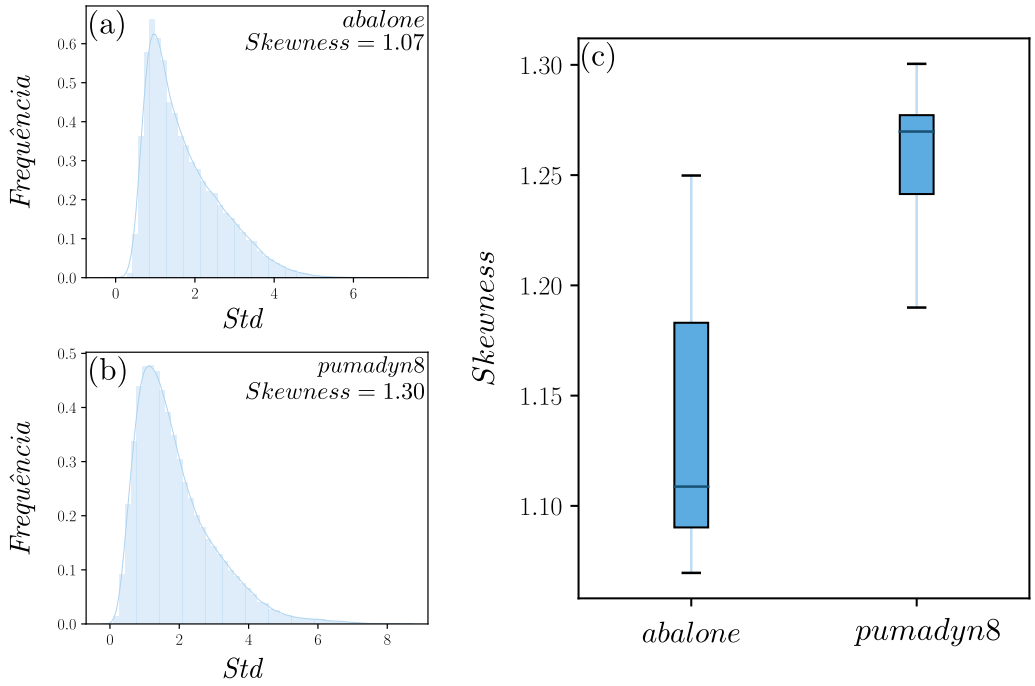

Figura 2. Caracterização dos resultados através do *skewness*. Em (a) e (b) é ilustrado o histograma do desvio padrão dos valores alvos das folhas das árvores para o conjunto de dados *abalone* e *pumadyn8*, respectivamente, de uma das 20 execuções. Em (c), sumariza, a partir de um Diagrama de Caixas, os valores dos *skewness* dos valores *STD* de 20 execuções.

A Figura 2 ilustra uma caracterização em cima de dois conjuntos de dados: *abalone* (no qual não houve melhora nos resultados) e *pumadyn8* (no qual houve a uma melhora significativa nos resultados). Para cada uma das 20 execuções, a distribuição dos

valores do desvio padrão nas folhas foi plotada e, em seguida, calculado o *skewness* (assimetria) desses valores. As Figuras 2 (a) e (b) ilustram a distribuição do desvio padrão de todas as folhas para cada árvore de uma determinada execução da *RF*. Percebe-se que, em ambos os conjuntos de dados, existem assimetrias para direita. Isso implica que há algumas folhas que são muito mais heterogêneas que a maioria. A ponderação atuará principalmente nessas folhas, afim de diminuir o impacto delas na predição final. A Figura 2(c) resume todos os 20 valores de *skewness* para esses dois conjuntos de dados (uma para cada execução), ilustrando que em *pumadyn8* a ponderação das predições tem impacto maior, uma vez que, as folhas com dispersão alta terão sua contribuição diminuída. Em *abalone*, como a dispersão é menor, os pesos terão menor impacto se comparado com *pumadyn8*.

A Figura 3 sumariza a caracterização, realizada na Figura 2, para todos os conjuntos de dados, analisando o *skewness* da distribuição dos STD das folhas. Quando analisam-se os conjuntos de dados que, na Tabela 2, a estratégia proposta empata com a RF padrão, observa-se que 13 dos 20 conjuntos de dados possuem *skewness* médio menor que 1. A área azul, na Figura 3, ilustra a região com valor de *skewness* menor que 1. Isso indica que esses 65% dos conjuntos de dados que não tem assimetria ou tem assimetria baixa, não impacta tanto no processo da ponderação proposta. Quando analisam-se os conjuntos de dados que a estratégia proposta melhora o resultado em relação a RF padrão, percebe-se que 9 dos 11 conjuntos de dados possuem *skewness* médio maior que 1 (região verde). Ou seja, cerca de 80% dos conjuntos de dados que tem assimetria alta, ocorre a melhora nos resultados. De forma geral, aqui é proposto o raciocínio que conjuntos de dados com assimetria alta tendem a melhorar os resultados. As caixas coloridas de vermelho representam os conjuntos de dados que vai contra este raciocínio.

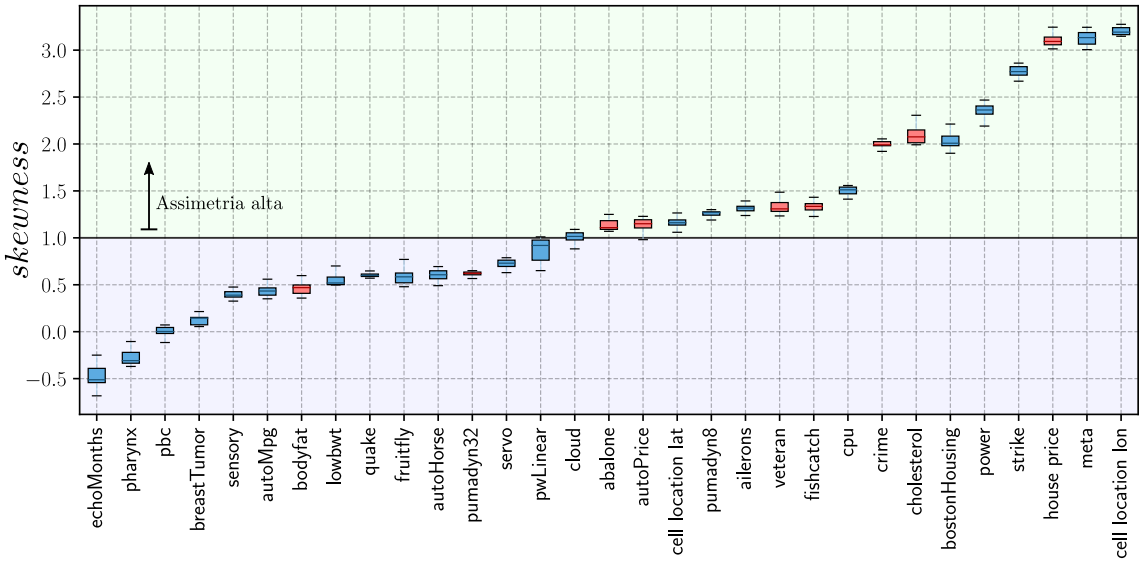

Figura 3. Diagramas de Caixa dos valores de *skewness* dos STD para todos os conjuntos de dados. Cada caixa representa a distribuição dos valores de *skewness* para cada uma das 20 execuções. As caixas estão ordenadas pela média do valor de *skewness*. A linha marcada em *skewness=1* representa um limiar para os valores muito assimétricos, ou seja, a área azul representa a região de pouca ou nenhuma assimetria, enquanto a área verde a região de alta assimetria.

7. Conclusão

Neste trabalho, levantou-se uma hipótese a qual a dispersão dos valores alvos dentro uma folha de uma árvore de regressão pode ser útil para capturar questões sobre a qualidade da predição daquela determinada folha. Buscou-se contribuir como uma nova forma de ponderação das predições das árvores de uma *Random Forest*, considerando o peso da predição inversamente proporcional a dispersão dentro de uma folha, com o objetivo de melhorar o poder de predição do floresta. Houve uma comparação do método proposto, com outras formas de ponderação encontradas na literatura, sendo preciso fazer adaptações nesses modelos, uma vez que a maioria dos métodos encontrados se aplicam para problemas de classificação. Foram testados 31 conjuntos de dados e pode-se afirmar que o método proposto não piorou o resultado em nenhum deles. Destaca-se ainda que a estratégia de ponderação proposta diminuiu o erro significativamente em 30% dos conjuntos de dados testados, quando comparado com *RF* padrão.

Referências

Amaratunga, D., Cabrera, J., and Lee, Y.-S. (2008). Enriched random forests. *Bioinformatics*, 24(18):2010–2014.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees.

Dietterich, T. G. et al. (2002). Ensemble learning. *The handbook of brain theory and neural networks*, 2:110–125.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844.

Kim, H., Kim, H., Moon, H., and Ahn, H. (2011). A weight-adjusted voting algorithm for ensembles of classifiers. *Journal of the Korean Statistical Society*, 40:437–449.

Kleiman, R. Instance-based out-of-bag weighting in random forests.

Li, H. B., Wang, W., Ding, H. W., and Dong, J. (2010). Trees weighting random forest method for classifying high-dimensional noisy data. In *2010 IEEE 7th International Conference on E-Business Engineering*, pages 160–163. IEEE.

Puuronen, S., Terziyan, V., and Tsymbal, A. (1999). A dynamic integration algorithm for an ensemble of classifiers. In *International symposium on methodologies for intelligent systems*, pages 592–600. Springer.

Rooney, N., Patterson, D., Anand, S., and Tsymbal, A. (2004). Dynamic integration of regression models. In *International Workshop on Multiple Classifier Systems*, pages 164–173. Springer.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Tsymbal, A., Pechenizkiy, M., and Cunningham, P. (2006). Dynamic integration with random forests. In *European conference on machine learning*, pages 801–808. Springer.