Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Gabriel Martelosso Cardoso

Avaliação do Particionamento Recursivo Baseado em Modelo para Predição.

Curitiba
2024

Gabriel Martelosso Cardoso

Avaliação do Particionamento Recursivo Baseado em Modelo para Predição.

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Cesar Augusto Taconeli

Curitiba
2024

Avalição Particionamento Recursivo Baseado em Modelo para Predição

Evaluation of the Model-Based Recursive Partitioning for Prediction

Gabriel M. Cardoso[1], Cesar A. Taconeli[2]

[1]Aluno do programa de Especialização em Data Science & Big Data, gabriel.martelosso@gmail.com
[2]Professor do Departamento de Estatística - DEST/UFPR, taconeli@gmail.com

Este trabalho investiga a eficácia do método de particionamento recursivo baseado em modelo (MOB) e sua expansão através de florestas aleatórias (MOB-RF) enquanto modelos preditivos. Avaliamos esses métodos em duas bases de dados distintas: Incidência de Diabetes (classificação) e Boston Housing (regressão). Os resultados indicam que MOB e MOB-RF são altamente eficazes, apresentando desempenho competitivo em várias métricas-chave. A análise também sugere que a adição da estrutura de random forest contribui fortemente na robustez do modelo.
Palavras-chave: particionamento recursivo baseado em modelo, florestas aleatórias, ensemble learning, modelagem preditiva

This study investigates the effectiveness of the Model-Based Recursive Partitioning (MOB) method and its extension through Random Forests (MOB-RF) as predictive models. We evaluate these methods on two distinct datasets: Diabetes Incidence (classification) and Boston Housing (regression). The results indicate that MOB and MOB-RF are highly effective, showing competitive performance across various key metrics. The analysis also suggests that the incorporation of the random forest structure significantly enhances the model's robustness.
Palavras-chave: model-based recursive partitioning, random forests, ensemble learning, predictive modeling

1. Introdução

Nos últimos anos, a aplicação de técnicas de machine learning em diversos domínios tem crescido exponencialmente, exigindo a escolha e implementação de modelos preditivos robustos que possam lidar com a complexidade e a variabilidade dos dados. A eficácia de um modelo preditivo não depende apenas de sua capacidade de ajuste aos dados de treinamento, mas também de sua capacidade de generalizar para novos dados não vistos, minimizando erros e maximizando a precisão das previsões (Hastie, Tibshirani, & Friedman, 2009).

Árvores de decisão e regressão são amplamente utilizadas em machine learning devido à sua simplicidade e interpretabilidade (Breiman et al., 1984). Elas descrevem a resposta em relação a uma sequência de separações nas covariáveis usando métricas de separabilidade como ganho de informação, entropia e erro quadrático médio (MSE). No entanto, essas árvores sofrem de problemas de overfitting, tornando-se altamente específicas aos dados de treinamento e sensíveis a variações (Breiman et al., 1984). Para combater o overfitting, técnicas de poda e métodos de ensemble, como florestas aleatórias, são comumente utilizados (Breiman, 2001).

Florestas aleatórias são um conjunto de árvores de decisão/regressão ajustadas em amostras bootstraping. Cada árvore fornece uma predição individual, e a predição final é determinada pela classe mais frequente ou pela média das predições. Isso resulta em melhor generalização, redução de overfitting e maior robustez à variância (Breiman, 2001).

O particionamento recursivo baseado em modelo (MOB) constrói árvores de decisão incorporando um modelo paramétrico no processo de particionamento (Zeileis, Hothorn, & Hornik, 2008). O MOB utiliza variáveis de partição para segmentar os dados e, dentro de cada segmento, ajusta um modelo paramétrico (como regressão linear ou logística) com as variáveis preditoras. Este método permite detectar mudanças nas relações entre as variáveis preditoras e a resposta ao longo de diferentes segmentos dos dados. Testes de

instabilidade são aplicados para verificar se as relações mudam significativamente, guiando a criação de novas divisões na árvore. Isso permite a construção de uma estrutura que particiona os dados em subconjuntos homogêneos, ajustando modelos locais que representam a relação única entre a resposta e as variáveis explicativas (Zeileis & Hornik, 2007).

O MOB tem sido amplamente utilizado na literatura para análises de subgrupos, permitindo identificar e modelar relações heterogêneas dentro dos dados (Seibold, Zeileis, & Hothorn, 2016). No entanto, o MOB apresenta problemas semelhantes às árvores de decisão tradicionais, como overfitting e alta complexidade (Zeileis et al., 2008). Para mitigar esses problemas, a expansão para MOB-RF combina o MOB com florestas aleatórias, resultando em menor overfitting, maior robustez à variância e fronteiras de decisão mais complexas (Garge, Bobashev, & Eggleston, 2013). Apesar dessas vantagens, MOB-RF depende de um tamanho amostral grande e é computacionalmente custoso.

O objetivo deste estudo é avaliar e explorar o desempenho do MOB e do MOB-RF no contexto de machine learning (predição), comparando-os com modelos mais comuns da literatura, como árvores de decisão, regressão logística e florestas aleatórias. Serão considerados dois casos de estudo: classificação da incidência de diabetes em indígenas Pima (Asuncion & Newman, 2007) e previsão dos preços de habitação em Boston (Harrison & Rubinfeld, 1978).

2. Metodologia

2.1. Bases de dados

Utilizamos a base de dados '*Diabetes incidence in Pima Indians*' para a análise de classificação, que é amplamente conhecida e utilizada em estudos sobre diabetes, disponível no repositório de aprendizado de máquina da UCI (Asuncion e Newman 2007), os campos da base são utilizados para prever a presença de diabetes.

Para análise de regressão foi o utilizado a base de dados Boston Housing, dados bastante populares desde sua análise por Breiman e Friedman (1985), nela utilizamos informações residenciais para prever os valores das habitações.

2.2. Modelos Avaliados

Neste estudo, foram aplicados e comparados os seguintes modelos. Árvore de decisão, que utiliza uma estrutura em árvore para tomar decisões baseadas em regras aprendidas a partir dos dados de treinamento; Regressão logística, um modelo estatístico utilizado para prever a probabilidade de ocorrência de um evento binário, baseado em uma ou mais variáveis independentes; MOB, o particionamento recursivo baseado em modelo que incorpora um modelo paramétrico no processo de particionamento, ajustando modelos locais nos nós da árvore; MOB-RF, a extensão do MOB que utiliza florestas aleatórias para reduzir o *overfitting* e aumentar a robustez; *Random Forest*, um *ensemble* de árvores de decisão ajustadas em amostras *bootstraping* para melhorar a generalização e reduzir a variância; *Support Vector Machine*, um modelo que utiliza hiperplanos em um espaço de alta dimensionalidade para a classificação dos dados; e Regressão Linear Simples, um modelo estatístico que estabelece uma relação linear entre uma variável dependente e uma ou mais variáveis independentes. Cada um desses modelos foi avaliado com base nas métricas definidas para os casos de classificação e regressão. Desta maneira serão utilizados, para fins de comparação, modelos bem explorados e comuns na literatura.

2.3. Métricas de Avaliação

As métricas foram escolhidas devido às suas propriedades estatísticas e relevância prática, proporcionando uma avaliação abrangente e equilibrada da performance dos modelos. As métricas de classificação foram selecionadas para capturar tanto a precisão quanto a capacidade discriminativa dos modelos, enquanto as métricas de regressão foram escolhidas para avaliar a precisão das previsões e a proporção da variabilidade explicada pelos modelos.

2.3.1. Métricas de classificação

Para avaliar a performance dos modelos de classificação, utilizamos as métricas de acurácia, precisão, recall, F1-score e AUC-ROC, conforme discutido em "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"de Trevor Hastie, Robert Tibshirani e Jerome Friedman.

A acurácia é uma métrica básica que mede a proporção de previsões corretas em relação ao total de previsões. É particularmente útil quando temos um conjunto de dados balanceado. No entanto, em casos de desbalanceamento de classes, a acurácia pode ser enganosa, uma vez que um modelo pode simples-

mente prever a classe majoritária para obter uma alta acurácia.

Para enfrentar essa limitação, utilizamos a precisão e o *recall*. A precisão mede a proporção de verdadeiros positivos entre todas as previsões positivas, sendo crucial em situações onde os falsos positivos têm um custo elevado. Por exemplo, em um modelo de detecção de fraudes, previsões incorretas de fraude podem resultar em desconforto para clientes honestos. O recall, por outro lado, mede a proporção de verdadeiros positivos entre todos os positivos reais, sendo essencial quando os falsos negativos têm um custo elevado, como na detecção de doenças graves, onde a falha em identificar todos os casos pode ser catastrófica.

O F1-score, a média harmônica da precisão e do recall, é utilizado para encontrar um equilíbrio entre essas duas métricas, especialmente em conjuntos de dados desbalanceados. A AUC-ROC, que mede a área sob a curva ROC (Característica de Operação do Receptor), avalia a capacidade do modelo de discriminar entre as classes positivas e negativas. Uma AUC-ROC alta indica um modelo com boa capacidade discriminativa, enquanto uma AUC-ROC próxima de 0,5 sugere que o modelo não é melhor que uma escolha aleatória.

2.3.2. Métricas de regressão

Para avaliar a performance dos modelos de regressão, utilizamos as métricas de Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R-Squared), conforme apresentadas em "Applied Linear Statistical Models"de John Neter, Michael H. Kutner, Christopher J. Nachtsheim e William Wasserman.

O MAE é a média dos erros absolutos entre as previsões e os valores reais, fornecendo uma medida clara e interpretável do erro médio. É menos sensível a outliers em comparação com o RMSE, tornando-o útil em cenários onde grandes desvios não são comuns. O RMSE, a raiz quadrada da média dos erros quadráticos, é mais sensível a grandes desvios, penalizando erros maiores de forma mais severa. Essa sensibilidade pode ser desejável em situações onde grandes erros são particularmente prejudiciais.

O Coeficiente de Determinação (R-Squared) mede a proporção da variabilidade dos dados que é explicada pelo modelo. Um R-Squared próximo de 1 indica que o modelo explica bem a variabilidade dos dados, enquanto um valor próximo de 0 indica o contrário. Esta métrica é amplamente utilizada por sua interpretabilidade e capacidade de fornecer uma visão geral da performance do modelo.

2.4. Validação e Ajuste de Hiperparâmetros

A validação cruzada k-fold foi utilizada para garantir que os modelos fossem robustos e generalizassem bem para dados não vistos. Realizamos uma busca de hiperparâmetros (Grid Search) para otimizar os parâmetros dos modelos, incluindo profundidade máxima da árvore, número mínimo de amostras por folha (MOB-RF e Florestas aleatórias).

3. Resultados

A Tabela 1 apresenta os resultados das métricas de avaliação para os modelos de classificação utilizado na base de dados de diabetes, apresentando sua médias e desvios padrões dentre os 10 k-folds. As métricas de classificação incluem a acurácia, precisão, recall (sensibilidade), F1-score e a área sob a curva ROC (ROC-AUC). É também calculado um Score Geral, que faz resumo das demais medidas.

Tabela 1: Desempenho dos Modelos na Classificação - Diabetes (Média e Desvio padrão nos 10 folds)

	Métricas					
	ACC	Prec.	Recall	F1	ROC AUC	Score geral
MOB	0.778 (0.06)	0.722 (0.08)	0.59 (0.1)	0.645 (0.07)	0.849 (0.05)	0.717 (0.07)
MOB RF	0.779 (0.04)	0.727 (0.09)	0.586 (0.07)	0.646 (0.06)	0.828 (0.05)	0.713 (0.06)
Log. Reg.	0.775 (0.04)	0.722 (0.1)	0.57 (0.1)	0.632 (0.08)	0.841 (0.03)	0.708 (0.07)
Dec. Tree	0.756 (0.03)	0.683 (0.07)	0.563 (0.15)	0.604 (0.09)	0.761 (0.08)	0.673 (0.09)
Rand. For.	0.772 (0.04)	0.703 (0.08)	0.594 (0.1)	0.64 (0.08)	0.816 (0.03)	0.705 (0.06)
SVM	0.769 (0.04)	0.728 (0.11)	0.542 (0.1)	0.614 (0.08)	0.838 (0.03)	0.698 (0.07)

Na Tabela 2 são apresentadas as medidas para os dados *Boston Housing*, nos modelos de regressão, observa-se as médias e desvios padrões nos resultados dos 10 folds. O Score Geral resume as demais medidas para obter uma nota geral.

Tabela 2: Desempenho dos Modelos na Regressão - Boston housing (Média e Desvio padrão nos 10 folds)

	Métricas			
	RMSE	R2	MAE	Score geral
MOB	3.61 (0.66)	0.84 (0.07)	2.56 (0.27)	0.79 (0.33)
MOB RF	3.52 (0.65)	0.84 (0.06)	2.51 (0.24)	0.82 (0.32)
Lin. Reg.	4.21 (0.87)	0.78 (0.08)	3.03 (0.32)	0.58 (0.42)
Dec. Tree	5.83 (0.36)	0.60 (0.06)	4.20 (0.45)	-
Rand. Forest	3.01 (0.53)	0.90 (0.03)	2.09 (0.33)	1.00 (0.30)
SVM	4.30 (0.84)	0.78 (0.10)	2.91 (0.31)	0.59 (0.42)

4. Discussão

4.1. Classificação de Diabetes

Os resultados para classificação de diabetes, apresentados na Tabela 1, mostram que os modelos MOB e MOB-RF obtiveram um desempenho superior aos demais em todas as métricas. MOB apresentou um acurácia de 0.778 (± 0.06), enquanto o MOB-RF alcançou 0.779 (± 0.04). Em termos de precisão, MOB obteve 0.722 (± 0.08) e MOB-RF 0.727 (± 0.09).

O *recall* do MOB foi de 0.590 (± 0.1), e do MOB-RF de 0.586 (± 0.07), destacando um bom equilíbrio entre precisão e sensibilidade. O F1-Score, que combina a precisão e o recall, foi de 645 (± 0.07) e 0.646 (± 0.06) para o MOB e MOB-RF, respectivamente. A métrica ROC-AUC foi de 0.849 (± 0.05) para o MOB e 0.828 (± 0.05) para o MOB-RF, indicando a capacidade de distinção entre as classes positivas e negativas.

Em comparação com os demais modelos, o MOB apresentou uma melhor percentual de 2.9% enquanto o MOB-RF apresentou 3.8%, de acordo com o Score Geral e o F1-Score, respectivamente. Além disso, a adição da estrutura de *random forest* resultou em uma redução de 17.1% no desvio padrão no Score geral e de 27.9% no F1-Score, quando comparamos o MOB-RF aos demais modelos, indicando maior robustez. Além disso, o desvio padrão nas medidas do MOB-RF foi 17.4% menor em relação ao MOB, reforçando que a implementação da estutura de *random forest* leva a uma maior robustez.

4.2. Regressão - Boston Housing

No estudo de regressão, os resultados apresentados na Tabela 2 indicam que os modelos MOB e MOB-RF superam a maioria dos modelos tradicionais em termos de RMSE, MAE e R-Squared. O MOB-RF obteve o menor RMSE, com um valor de 3.52 (± 0.65), enquanto o MOB apresentou um RMSE de 3.61 (± 0.66).

Em termos de R-Squared, o MOB-RF obteve 0.84 (± 0.06), enquanto o MOB resultou em 0.84 (± 0.07), indicando uma alta proporção da variabilidade na resposta é explicada pelos modelos. A métrica MAE foi de 2.51 (± 0.24) para MOB-RF e 2.56 (± 0.27) para o MOB, demonstrando a eficácia dos modelos na previsão dos valores de habitação.

Quando comparado com os demais modelos com exceção do *random forest*, o MOB-RF apresentou melhores percentuais de 24.5% no RMSE, 14.6% no *R-Squared* e 23.2% no MAE. Além disso, o MOB-RF apresentou reduções nos desvios padrões em relação aos demais modelos, com uma redução de 10.6% no RMSE, 25.4% no R-Squared e 32.2% no MAE, indicando uma maior robustez no modelo. Por fim, ouve um redução de 5% do desvio padrão das medidas no MOB-RF quando comparado com o MOB, indicando que a estrutura de *random forest* contribui em tornar o modelo mais robusto.

5. Conclusão

Os modelos MOB e MOB-RF demonstraram um desempenho geral de predição superior aos modelos tradicionais no cenário de classificação, e foram apenas ultrapassados pelo modelo de Random forest no caso de regressão, indicando que este métodos possam ser úteis num cenário demachine learning.

A estrutura de random forest incorporada no MOB-RF contribui expressivamente para a redução da variância nos resultados dos k-folds, como pode ser observado em ambos os cenários.

A proximidade dos resultados entre o MOB e o MOB-RF não favorece nem um nem outro olhando apenas para a predição. Porém a menor variância no MOB-RF indica que este é mais robusto e que pode ser mais indicado num contexto preditivo.

Os bons resultados nas predições e a redução da variância nos resultados indica que o MOB-RF é um modelo robusto com grande potencial preditivo, podendo competir e até superar os modelos comumente adotados para trabalhos preditivos.

Agradecimentos

Gostaria de expressar minha gratidão a todos os docentes que contribuíram para minha formação acadêmica ao longo deste curso. Em especial, agradeço aos professores Cesar Augusto Taconeli, Wagner Hugo Bonat, Anderson Ara, Paulo Justiniano Ribeiro Jr e Paulo R. Lisboa de Almeida cujas orientações, ensinamentos e apoio foram fundamentais para o desenvolvimento deste trabalho de conclusão de curso. Suas aulas, insights e conselhos não apenas ampliaram meu conhecimento, mas também inspiraram meu crescimento pessoal e profissional.

Também gostaria de agradecer à minha namorada, Sinaí Sánchez, pela sua paciência, compreensão e apoio incondicional durante todo este processo. Sua presença e encorajamento foram essenciais para que eu pudesse superar os desafios e alcançar este objetivo. Sou grato por estar ao meu lado em todos os momentos, tornando esta jornada mais leve e significativa.

Referências

[1] Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. Journal of Computational and Graphical Statistics, 17(2), 492-514.

[2] Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partitioning in R. Journal of Machine Learning Research, 16, 3905-3909.

[3] Zeileis, A., & Hornik, K. (2007). Generalized M-Fluctuation Tests for Parameter Instability. Statistica Neerlandica, 61(4), 488–508.

[4] Seibold, H., Zeileis, A., & Hothorn, T. (2016). Model-based recursive partitioning for subgroup analyses. The International Journal of Biostatistics, 12(1), 45-63. https://doi.org/10.1515/ijb-2015-0032

[5] Garge, N.R., Bobashev, G. & Eggleston, B. Random forest methodology for model-based recursive partitioning: the mobForest package for R. BMC Bioinformatics 14, 125 (2013).

[6] Breiman L (2001). “Random Forests.” Machine Learning, 45(1), 5–32.

[7] Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). Classification and Regression Trees. Wadsworth, California.

[8] Breiman L, Friedman JH (1985). “Estimating Optimal Transformations for Multiple Regression and Correlation.” Journal of the American Statistical Association, 80(391), 580–598.

[9] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[10] Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied Linear Statistical Models. McGraw-Hill/Irwin.

[11] Asuncion A, Newman DJ (2007). “UCI Repository of Machine Learning Databases.” URL http://www.ics.uci.edu/ mlearn/MLRepository.html.

[12] Bivand, Roger, Revisiting the Boston Data Set (Harrison and Rubinfeld, 1978): A Case Study in the Challenges of System Articulation (December 2, 2015). NHH Dept. of Economics Discussion Paper No. 30/2015, Available at SSRN: https://ssrn.com/abstract=2719454 or http://dx.doi.org/10.2139/ssrn.2719454