Algoritmos de Machine Learning para Predição da Sobrevida do Câncer de Mama

Machine Learning Algorithms for Prediction of Breast Cancer Survival

Algoritmos de aprendizaje automático para la predicción de la supervivencia del cáncer de mama

Pablo Deoclecia dos Santos[1], Erika Yahata[2,3], Talita Santos Piheiro[1], Fellipe Soares de Oliveira[1], Priscyla Waleska Simões[1,2,3]

1 Programa de Pós-Graduação em Engenharia Biomédica, Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas-CECS, Universidade Federal do ABC-UFABC, São Bernardo do Campo (SP), Brasil.
2 Programa de Pós-Graduação em Engenharia da Informação, Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas-CECS, Universidade Federal do ABC-UFABC, Santo André (SP), Brasil.
3 Curso de Engenharia Biomédica, Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas-CECS, Universidade Federal do ABC-UFABC, São Bernardo do Campo (SP), Brasil.

Autor correspondente: Profa Dra. Priscyla Waleska Simões
E-mail: pritsimoes@gmail.com

Resumo

Objetivo: O presente artigo apresenta uma análise comparativa de algoritmos de Aprendizado de Máquina aplicados à predição da Sobrevida do Câncer de Mama. **Métodos:** Estudo descritivo que considerou dados de 1.570 pacientes com câncer de mama estágio I-III. A técnica *Synthetic Minority Oversampling Technique* foi aplicada devido ao desbalanceamento do conjunto de dados. Foram considerados no estudo os algoritmos *Naive Bayes*, *Random Forest*, *Multilayer Perceptron* e AdaBoost, e como estratégia de aprendizagem a validação cruzada. **Resultados:** O modelo desenvolvido a partir do algoritmo Random Forest apresentou maior acurácia (96,2%; IC95%: 95,5%-96,9%) e especificidade (97,4%; IC95%: 96,6%-98,2%); e o modelo desenvolvido a partir do AdaBoost, maior sensibilidade (95,3%; IC95%: 94,3%-96,4%). **Conclusão:** Assim, dentre os modelos apresentados em nosso estudo, o desenvolvido a partir do algoritmo Random Forest apresentou, no geral, as melhores medidas de avaliação na predição da sobrevida do Câncer de Mama.

Descritores: Análise de Sobrevida; Aprendizado de Máquina; Câncer mama.

Abstract

Objective: This paper aims to show a comparative analysis of Machine Learning algorithms applied to Breast Cancer Survival prediction. **Methods:** Descriptive study that considered data from 1,570 patients with stage I-III breast cancer. The Synthetic Minority Oversampling Technique was applied due to an imbalance in the dataset. The Naive Bayes, Random Forest, Multilayer Perceptron and AdaBoost algorithms were considered in the study, and cross-validation as a learning strategy. **Results:** The model developed from the Random Forest algorithm showed greater accuracy (96.2%; 95%CI: 95.5%-96.9%) and specificity (97.4%; 95%CI: 96.6%-98.2%); and the model developed from AdaBoost, greater sensitivity (95.3%; 95%CI: 94.3%-96.4%). **Conclusion:** Thus, among the models presented in our study, the one developed from the Random Forest algorithm presented, in general, the best evaluation measures in the prediction of breast cancer survival.

Keywords: Survival Analysis; Machine Learning; Breast cancer.

Resumen

Objetivo: Este estudio tiene como objetivo mostrar un análisis comparativo de los algoritmos de Aprendizaje automático aplicados a la predicción de la supervivencia al cáncer de mama. **Métodos:** Estudio descriptivo que consideró datos de 1.570 pacientes con cáncer de mama en estadio I-III. Se aplicó la técnica de sobremuestreo de minorías sintéticas debido a un desequilibrio en el conjunto de datos. Se consideraron en el estudio los algoritmos Naive Bayes, Random Forest, Multilayer Perceptron y AdaBoost, y la validación cruzada como estrategia de aprendizaje. **Resultados:** El modelo desarrollado a partir del algoritmo Random Forest mostró mayor precisión (96,2%; IC95%: 95,5%-96,9%) y especificidad (97,4%; IC95%: 96,6%-98,2%); y el modelo desarrollado a partir de AdaBoost, mayor sensibilidad (95,3%; IC95%: 94,3%-96,4%). **Conclusión:** Así, entre los modelos presentados en nuestro estudio, el desarrollado a partir del algoritmo Random Forest presentó, en general, las mejores medidas de evaluación en la predicción de la supervivencia del cáncer de mama.

Palabras clave: Análisis de Supervivencia; Aprendizaje automático; Cáncer de mama.

Introdução

O câncer de mama é o mais frequentemente diagnosticado e a principal causa de morte relacionada ao câncer em mulheres em todo o mundo [1]. No Brasil, excluindo os tumores de pele não melanoma, o câncer de mama é o de maior incidência em mulheres de todas as regiões, com taxas mais altas nas regiões Sul e Sudeste [2]. Para o ano de 2022 foram estimados 66.280 casos novos, o que representa uma taxa de incidência de 43,74 casos novos por 100 mil mulheres [2].

O câncer de mama apresenta alguns subtipos com características biológicas distintas que levam a diferenças de resposta ao tratamento e desfecho clínico [3]. A detecção em estágios avançados da doença resulta no tratamento com pior prognóstico; assim, o rastreamento tem contribuído na detecção precoce e melhora da sobrevida [4].

Com o recente avanço da Tecnologia da Informação e Comunicação na Saúde, o crescimento exponencial dos dados de saúde, o Aprendizado de Máquina (*Machine Learning*), a partir de seus algoritmos, por exemplo, possibilita identificar padrões ocultos nos dados para a predição clínica [5]. Os modelos de Aprendizado de Máquina podem auxiliar a minimizar falsos positivo e falsos negativos, pois busca explorar padrões e relacionamentos entre um grande número de casos e prever o resultado de uma doença usando casos históricos armazenados em conjuntos de dados [6].

Por outro lado, a sobrevida do câncer reflete a agressividade da doença, a eficácia do tratamento e fatores de risco. Enquanto as taxas de sobrevida hospitalar são normalmente usadas para avaliar os cuidados prestados em um determinado hospital, a sobrevida baseada na população reflete a eficácia da estratégia geral de controle do câncer de mama [7].

Alguns estudos têm apresentado bons resultados na predição do câncer de mama à partir de algoritmos de Aprendizado de Máquina [8], no entanto, observa-se a lacuna de modelos de Aprendizado de Máquina aplicados à sobrevida do Câncer de Mama que possam ser incorporados à prática clínica [9]. Mediante o exposto, o presente artigo apresenta um comparativo de algoritmos de Aprendizado de Máquina aplicados à predição da Sobrevida do Câncer de Mama.

Métodos

Estudo descritivo que considerou um *dataset* público oriundo de um estudo realizado no hospital Sun Yat-sen Memorial, em Guangzhou na China, e conduzido entre janeiro de 2000 a dezembro de 2010 com dados de 1.570 pacientes com câncer de mama estágio I-III tratadas em ambiente hospitalar.

A Tabela 1 ilustra os atributos considerados no estudo. No pré-processamento foi realizada a categorização e dicotomização dos atributos numéricos pela média. A técnica de reamostragem e imputação de dados sintéticos *Synthetic Minority Oversampling Technique* (SMOTE) foi aplicada, devido ao desbalanceamento do *dataset* [10].

Tabela 1 – Atributos de amostra.

Atributo	Descrição
Desfecho	Óbito, Sobrevida
Sobrevida Global	Menor que 85 anos, Maior ou igual a 85 anos
Sobrevida Livre	Menor que 80 anos, Maior ou igual a 80 anos
Idade	Menor que 50 anos, Maior ou igual a 50 anos
Anti Inflamatório	Não, Sim
Grau	Grau 3, Grau 2, Grau 1
Quimioterapia	Não, Sim
Linfonodos	Positivo, Negativo
Tamanho do Tumor	T3, T2, T1
Cirurgia Axilar	Dissecção axilar, Biópsia sentinela
Cirurgia Mamária	Mastectomia, Conservadora
Receptor de Estrogênio	Negativo, Positivo
Receptor de Progesterona	Negativo, Positivo
HER2	Positivo, Negativo
Subtipo Molecular	Triplo negativo, HER2 Positivo, Luminal
Glóbulos Brancos	Menor que 6.7 células x 10^9 por litro, Maior ou igual a 6.7 células x 10^9 por litro
Hemoglobinas	Menor que 126 gramas por litro, Maior ou igual a 126 gramas por litro
Plaquetas	Maior ou igual a 242 células x 10^9 por litro, Menor que 242 células x 10^9 por litro

Linfócitos	Maior ou igual a 2 células x 10^9 por litro, Menor que 2 células x 10^9 por litro
Neutrófilos	Menor que 4.1 células x 10^9 por litro, Maior ou igual a 4.1 células x 10^9 por litro
Monócitos	Maior ou igual a 0.4 células x 10^9 por litro, Menor que 0.4 células x 10^9 por litro
Relação Neutrófilo-Linfócito	Maior ou igual a 2.3 células, Menor que 2.3 células
Relação Plaqueta-Linfócito	Maior ou igual a 135 células, Menor que 135 células
Relação Linfócito-Monócito	Menor que 5.4 células, Maior ou igual a 5.4

Foram considerados no estudo os algoritmos *Naive Bayes*, *Random Forest* (RF), *Multilayer Perceptron* (MLP) e AdaBoost, e a validação cruzada (10-fold) como método de amostragem, e realizados alguns ajustes nos modelos.

Utilizamos a correlação de Pearson para avaliar a contribuição de cada atributo em relação ao atributo alvo (desfecho). Assim, observamos baixa correlação ($r_p < 0,1$) do atributo-alvo (desfecho - óbito, sobrevida) em relação à três atributos (Glóbulos Brancos, Relação Plaqueta-Linfócito e Neutrófilos) que foram excluídos dos modelos.

Para a avaliação dos modelos, adotamos a acurácia, a sensibilidade e a especificidade. A ferramenta WEKA v. 3.8.6 [11, 12] foi utilizada no estudo.

Resultados

A Tabela 2 apresenta as características da amostra antes e após a reamostragem com a aplicação do SMOTE no *dataset* original devido ao desbalanceamento da classe que desejamos analisar no nosso estudo, o desfecho Óbito e Sobrevida.

Tabela 2 – Características da amostra.

Atributos	Dataset	
	Original n = 1570 n (%)	**Reamostragem n = 2924 n (%)**
Desfecho		
Óbito	108 (6,90)	1462 (50,00)
Sobrevida	1462 (93,10)	1462 (50,00)
Sobrevida Global		
Menor que 85 anos	870 (55,50)	2100 (71,90)
Maior ou igual a 85 anos	698 (44,50)	822 (28,10)
Sobrevida Livre		
Menor que 80 anos	857 (54,70)	2162 (74,00)
Maior ou igual a 80 anos	711 (45,30)	760 (26,00)
Idade		
Menor que 50 anos	890 (56,70)	1515 (51,80)
Maior ou igual a 50 anos	679 (43,30)	1408 (48,20)
Anti Inflamatório		
Não	445 (28,50)	983 (33,70)
Sim	1114 (71,50)	1930 (66,30)
Grau		
Grau 3	531 (38,10)	1217 (44,30)
Grau 2	760 (54,60)	1428 (52,00)
Grau 1	102 (7,30)	102 (3,70)
Quimioterapia		
Não	229 (14,60)	255 (8,70)
Sim	1341 (85,40)	2669 (91,30)
Linfonodos		
Positivo	627 (42.00)	1601 (56,30)
Negativo	865 (58,00)	1245 (43,70)
Tamanho do Tumor		
T3	105 (7,20)	131 (4,60)
T2	510 (34,70)	1388 (49,20)
T1	853 (58,10)	1303 (46,20)

Cirurgia Axilar		
Dissecção axilar	740 (47,70)	1677 (57,70)
Biópsia sentinela	811 (52,30)	1228 (42,30)
Cirurgia Mamária		
Mastectomia	879 (56,00)	1946 (66,60)
Conservadora	691 (44,00)	978 (33,40)
Receptor de Estrogênio		
Negativo	437 (27,80)	1025 (35,10)
Positivo	1133 (72,20)	1899 (64,90)
Receptor de Progesterona		
Negativo	471 (30,00)	1045 (35,80)
Positivo	1098 (70,00)	1878 (64,20)
HER2		
Positivo	346 (22,10)	804 (27,50)
Negativo	1220 (77,90)	2116 (72,50)
Subtipo Molecular		
Triplo negativo	225 (14,30)	517 (17,70)
HER2 Positivo	344 (21,90)	865 (29,60)
Luminal	1001 (63,80)	1542 (52,70)
Hemoglobinas		
Menor que 126 gramas por litro	696 (44,30)	1120 (38,30)
Maior ou igual a 126 gramas por litro	874 (55,70)	1804 (61,70)
Plaquetas		
Maior ou igual a 242 células x 10^9 por litro	733 (46,70)	1162 (39,70)
Menor que 242 células x 10^9 por litro	837 (53,30)	1762 (60,30)
Linfócitos		
Maior ou igual a 2 células x 10^9 por litro	725 (46,20)	1201 (41,10)
Menor que 2 células x 10^9 por litro	845 (53,80)	1723 (58,90)
Monócitos		
Maior ou igual a 0.4 células x 10^9 por litro	897 (57,10)	2025 (69,30)
Menor que 0.4 células x 10^9 por litro	673 (42,90)	899 (30,70)
Relação Neutrófilo-Linfócito		
Maior ou igual a 2.3 células	561 (35,70)	1252 (42,80)
Menor que 2.3 células	1009 (64,30)	1672 (57,20)
Relação Linfócito-Monócito		

Menor que 5.4 células	1017 (64,80)	2208 (75,50)
Maior ou igual a 5.4	553 (35,20)	716 (24,50)

A Tabela 3 apresenta a avaliação dos modelos. O algoritmo RF apresentou maior acurácia (96,2%; IC95%: 95,5%-96,9%), isto significa que o modelo apresentou 96,2% de classificações corretas.

O melhor resultado de sensibilidade foi observado no modelo AdaBoost (95,3%; IC95%: 94,3%-96,4%), ou seja, o poder do modelo AdaBoost em predizer que foram a óbito as pacientes que realmente tinham evoluído a óbito de acordo com o desfecho clínico apresentado no *dataset*.

O maior valor de especificidade proveio do algoritmo *Random Forest* (97,4%; IC95%: 96,6%-98,2%), ou seja, o poder do modelo Random Forest em predizer que evoluíram para sobrevida as pacientes que realmente sobreviveram de acordo com o desfecho clínico apresentado no *dataset*.

Tabela 3 – Avaliação dos modelos.

Modelo	**Medida de avaliação (IC95%)**		
	Acurácia	**Sensibilidade**	**Especificidade**
Naive Bayes	83,8% (82,5%-85,1%)	88,2% (86,5%-89,8%)	79,4% (77,3%-81,5%)
Random Forest	96,2% (95,5%-96,9%)	95,1% (94,0%-96,2%)	97,4% (96,6%-98,2%)
Multilayer Perceptron	93,5% (92,6%-94,4%)	93,8% (92,6%-95,1%)	93,2% (91,9%-94,5%)
AdaBoost	96,1% (95,4%-96,8%)	95,3% (94,3%-96,4%)	96,9% (96,0%-97,7%)

Os resultados de previsão do nosso estudo mostram que o modelo desenvolvido a partir do algoritmo *Random Forest* apresentou maior acurácia (96,2%; IC95%: 95,5%-96,9%) e especificidade (97,4%; IC95%: 96,6%-98,2%); e o modelo desenvolvido a partir do AdaBoost, maior sensibilidade (95,3%; IC95%: 94,3%-96,4%). Assim, dentre os modelos apresentados em nosso estudo, o desenvolvido a partir do algoritmo *Random Forest* apresentou, em geral, as melhores medidas de avaliação na predição da sobrevida do câncer de mama.

Discussão

O presente artigo apresenta uma análise comparativa de algoritmos de *Machine Learning* aplicados à predição da Sobrevida do Câncer de Mama. Nesse contexto, buscando discutir nossos achados, uma meta-análise da Indonésia que incluiu estudos publicados de 2000 a 2018 sobre a acurácia diagnóstica de diferentes algoritmos de Aprendizado de Máquina para predição do câncer de mama por meio dos algoritmos *Support Vector Machine*, Redes Neurais Artificiais, Árvores de Decisão, *Naive Bayes*, e *K-Nearest Neighbor*; concluiu que o modelo desenvolvido a partir do algoritmo *Support Vector Machine* apresentou as melhores medidas de avaliação (acurácia de 99,51%, sensibilidade de 95,45% e especificidade de 77,86%), em comparação com os demais algoritmos [13].

Nesse contexto, um estudo realizado no Irã no período de 1999 a 2007 avaliou a relação de fatores de risco na predição da sobrevida do câncer de mama por meio dos algoritmos *Naive Bayes, Random Forest*, 1-*Nearest Neighbor*, AdaBoost, *Support Vector Machine*, *Radial Basis Function Neural*, e MLP; e concluiu que o modelo desenvolvido a partir do algoritmo Random Forest foi o de maior acurácia (96,0%), sensibilidade (96,0%) e especificidade (98,0%) [6].

Num outro estudo desenvolvido na Malásia no período de 1993 e 2017 que avaliou alguns fatores clínicos (sexo, menopausa, histórico familiar, entre outros), na predição da sobrevida do câncer de mama por meio dos algoritmos *Support Vector Machine*, *Random Forest*, Árvores de Decisão, e MLP; concluiu que o modelo desenvolvido pelo algoritmo *Random Forest* foi o que apresentou maior acurácia (83,3%), sensibilidade (93,7%) e especificidade (76,8%)[14].

Nosso estudo apresentou no modelo mais bem avaliado (*Random Forest*), acurácia de 96,2% (IC95%: 95,5%-96,9%). Na comparação do mesmo modelo aos estudos supracitados, a acurácia apresentada em nosso estudo foi maior do que no estudo do Irã [6] (acurácia de 96,0%) e da Malásia [14] (83,3%).

O MLP é uma rede neural artificial (RNA) que utiliza retro propagação e apresenta pelo menos três camadas de neurônios para classificação de dados: uma camada de

entrada, uma camada oculta e uma camada de saída [15]. O AdaBoost é um algoritmo de *boosting*, o qual funciona executando repetidamente algoritmos de aprendizado fraco em várias distribuições sobre os dados de treinamento [16]. O algoritmo *Naive Bayes* faz o uso do Teorema de Bayes [17] baseando-se na suposição de aleatoriedade [18].

O modelo desenvolvido a partir do algoritmo *Random Forest* apresentou, no geral, os melhores resultados em nossa pesquisa. Dentre as características que podem explicar tal desempenho, pode-se destacar que é um algoritmo que consiste em uma coleção de árvores de regressão de base aleatória que, na sua saída, resultam em uma única variável aleatória. Essas árvores aleatórias são combinadas para formar a estimativa de regressão agregada [19]. O *Random Forest* muda como a classificação ou as árvores de regressão (*Regression Tree)* são construídas; em árvores clássicas cada nó é dividido usando a melhor divisão entre todas as variáveis; no *Random Forest*, cada nó é dividido usando o melhor entre um subconjunto de preditores escolhidos aleatoriamente naquele nó [20]. Apresenta dois parâmetros (o número de variáveis no subconjunto aleatório em cada nó e o número de árvores na floresta), e geralmente não é muito sensível aos seus valores [20].

Conclusão

Considerando a grandeza do *dataset* utilizado no nosso estudo, o balanceamento e os ajustes realizados, os resultados são promissores em relação a utilização de algoritmos de ML para auxiliar na predição da sobrevida do câncer de mama. Dentre os modelos apresentados em nosso estudo, o desenvolvido a partir do algoritmo *Random Forest* apresentou, no geral, as melhores medidas de avaliação. Trabalhos futuros podem avaliar outros modelos de Aprendizado de Máquina a partir de algoritmos como o *Random Survival Forest*, que consideram a censura inerente aos dados.

Agradecimentos

Esse estudo foi financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - *Finance Code* 001, Conselho Nacional de

Desenvolvimento Científico e Tecnológico (CNPq) e Universidade Federal do ABC (UFABC).

O estudo foi realizado no âmbito do Centro de Pesquisa Aplicada em Inteligência Artificial BI0S – Brazilian Institute of Data Science, apoiado pelo processo nº 2020/09838-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

1. Hassan MA, Ates-Alagoz Z. Cyclin-Dependent Kinase 4/6 Inhibitors Against Breast Cancer. Mini Rev Med Chem. 2022.
2. INCA. Estimativa 2020. In: Saúde Md, editor. Incidência de Câncer no Brasil. Brasil: Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA); 2019.
3. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World J Clin Oncol. 2014;5(3):412-24.
4. Milosevic M, Jankovic D, Milenkovic A, Stojanov D. Early diagnosis and detection of breast cancer. Technol Health Care. 2018;26(4):729-59.
5. Trister AD, Buist DSM, Lee CI. Will Machine Learning Tip the Balance in Breast Cancer Screening? JAMA Oncol. 2017;3(11):1463-4.
6. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. Technol Health Care. 2016;24(1):31-42.
7. Nandakumar A, Anantha N, Venugopal TC, Sankaranarayanan R, Thimmasetty K, Dhar M. Survival in breast cancer: a population-based study in Bangalore, India. Int J Cancer. 1995;60(5):593-6.
8. Puja G, Shruti G. Breast Cancer Prediction using varying Parameters of Machine Learning Models. Procedia Computer Science. 2020;171:593-601.
9. Pinheiro TS, Yahata E, Santos PDd, Oliveira FSd, Takahata AK, Suyama R, et al. Machine Learning e Análise Multivariada aplicados à Sobrevida do Câncer Mama. Journal of Health Informatics. 2022;14(0).
10. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002;16:321-57.
11. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009;11(1):10–8.
12. Frank E, Hall M, Witten I. Appendix B - The WEKA workbench. In: Ian HW, Eibe F, Mark AH, Christopher JP, editors. Data Mining (Fourth Edition). Fourth Edition ed: Morgan Kaufmann; 2017. p. 553-71.

13. Nindrea RD, Aryandono T, Lazuardi L, Dwiprahasto I. Diagnostic Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis. Asian Pac J Cancer Prev. 2018;19(7):1747-52.

14. Kalafi EY, Nor NAM, Taib NA, Ganggayah MD, Town C, Dhillon SK. Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data. Folia Biol (Praha). 2019;65(5-6):212-20.

15. Le Thien MA, Redjdal A, Bouaud J, Seroussi B. Deep Learning, a Not so Magical Problem Solver: A Case Study with Predicting the Complexity of Breast Cancer Cases. Stud Health Technol Inform. 2021;287:144-8.

16. Freund Y, Schapire RE, editors. A desicion-theoretic generalization of on-line learning and an application to boosting. Computational Learning Theory; 1995 1995//; Berlin, Heidelberg: Springer Berlin Heidelberg.

17. Henry R, Meltzer MI. Etymologia: Bayesian Probability. Emerg Infect Dis. 2017;23(1):28.

18. Krishnan S. 6 - Machine learning for biomedical signal analysis. In: Krishnan S, editor. Biomedical Signal Analysis for Connected Healthcare: Academic Press; 2021. p. 223-64.

19. Biau G. Analysis of a Random Forests Model. Journal of Machine Learning Research. 2010;13.

20. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.