Previsão do Período de Recorrência de Câncer de Mama Utilizando Técnicas de Data Mining

Gabriel Baruque
Engenharia Elétrica
Pontifícia Universidade Católica
Rio de Janeiro, Brasil
gabriel@baruque.com.br

***Abstract*—Breast cancer diagnosis has been a benchmark for machine learning techniques for decades now. A more current problem is the prediction of breast cancer recurrence, due to its importance in medical area, since many deaths involving this disease happen because of recurrence. This paper expands This paper expands the breast cancer diagnosis problem and uses a set of machine learning classification methods in order to predict whether the patient will develop recurrence and at which possible time window, according to a classification made by a reference paper. Alongside, some pre-processing methods were used and explained. The results were interesting but some considerations had to be made. Due to a small dataset, a cross-validation was used in all experiments, with no data separated only for test.**

Keywords—Classification, Decision Tree, Random Forest, Relief-F, Feature Selection, Oversampling, SMOTE-NC.

I. INTRODUÇÃO

O diagnóstico do câncer de mama (CM) tem servido como *benchmark* para métodos de inteligência artificial (IA) e *machine learning* (ML) por décadas. Tem sido considerado um problema de solução não-trivial, principalmente pelos dados ruidosos e em geral escassos [1] para um problema de ML.

Aproximadamente todas as mortes relacionadas a CM são causadas pela recorrência e/ou a metástase do CM, ao invés do tumor primário. Além disso, a metástase geralmente não acontece no mesmo período que o tumor primário e o tempo de recorrência pode variar bastante [2]. Realizar a previsão do tempo de recorrência do CM, principalmente para uma recorrência tardia, constitui um desafio atual na área.

Este trabalho visa expandir o problema do diagnóstico do CM, utilizando alguns métodos de Data Mining já consolidados atualmente: Árvore de Decisão, Vizinhos Mais Próximos (KNN) e *Random Forest*. Além desses métodos, diversas técnicas de pré-processamento também são utilizadas para que seja possível realizar uma melhor classificação do banco de dados.

Os experimentos da pesquisa foram realizados em Python, utilizando a biblioteca Scikit Learn, com o auxilio do software Weka, para realizar alguns pré-processamentos.

A pesquisa abordará o banco de dados utilizado, percorrendo suas características, como quantidade de atributos e observações, e tipos de variáveis presentes. Discorrerá sobre os métodos de pré-processamento utilizados, tanto para limpeza e extração de dados, como para seleção de atributo. Detalhará os experimentos feitos, explorando e justificando os resultados. Por fim, concluirá com observações sobre os resultados obtidos, analisando possíveis problemas encontrados no decorrer dos experimentos em conjunto com possíveis soluções.

II. METODOLOGIA

Neste capítulo, todo o desenvolvimento do trabalho será exposto, incluindo maiores detalhes sobre o banco de dados utilizado, técnicas de pré-processamento e seleção de atributos, além de detalhes dos métodos de classificação utilizados.

A. Banco de dados

Neste tópico, serão expostas as características do banco de dados e como algumas delas podem influenciar positivamente ou negativamente uma classificação utilizando diferentes métodos. Os principais atributos serão mencionados, porém não serão detalhados, pois o detalhamento médico/biológico extrapola o escopo do trabalho.

Inicialmente o banco de dados fornecido pelo INCA continha 1084 observações e 96 atributos. Dentre os atributos, 39 eram de características genômicas, o que não seria abordado pela pesquisa, e foram então extraídos do banco de dados. Diversos outros atributos, principalmente administrativos, foram retirados, além de atributos redundantes, pois não seriam de ajuda nas classificações. Após essa extração inicial, mantiveram-se 42 atributos que possivelmente seriam utilizados. Por fim, uma análise mais detalhada de cada um mostrou que muitos desses atributos apresentavam uma quantidade muito grande de valores faltantes ou a variância existente era muito pequena ($\leq 5\%$). Nesses casos, novamente foi feita a extração das variáveis, pois não ajudariam no processo de classificação, restando por fim 18 variáveis além da classe.

As variáveis utilizadas foram:

- *Cancer Type Detailed*
- *Subtype*
- *Diagnosis Age*
- *Sex*
- *Ethnicity Category*
- *Race Category*
- *Neoplasm Disease Stage American Joint Committee on Cancer Code*
- *Aneuploidy Score*
- *Buffa Hypoxia Score*
- *Ragnum Hypoxia Score*
- *Winter Hypoxia Score*

- *Fraction Genome Altered*
- *In PanCan Pathway Analysis*
- *Mutation Count*
- *New Neoplasm Event Post Initial Therapy Indicator*
- *Prior Diagnosis*
- *Radiation Therapy*
- *Tissue Prospective Collection Indicator*

A classe foi formada pelo conjunto de dois atributos: *Progression-free Interval Time (PFI.time)* e *Progression-free Interval Status (PFI.status)*. O primeiro, indicando o tempo no qual o paciente desenvolveu a recorrência da doença, enquanto o segundo, sendo um atributo binário, indicava se a recorrência havia acontecido ou não. A união desses atributos tornou possível a criação de uma variável auxiliar (que substituiu as duas mencionadas), indicando o período de recorrência do CM do paciente:

- Early: recorrência ≤ 2 anos
- Mid: 2 ≤ recorrência ≤ 5 anos
- Late: recorrência > 5 anos
- Survival: sem recorrência após 5 anos

B. Pré-processamentos

Como mencionado anteriormente, a base de dados foi tratada inicialmente com alguns métodos de pré-processamento, entre eles filtros de variância e análise de dados faltantes. Outros métodos utilizados ao decorrer dos experimentos foram a seleção de variáveis com Relief-F e superamostragem por replicação e utilizando a técnica *Synthetic Minority Oversampling Technique - Nominal Continuous* (SMOTE-NC).

Variáveis com mais de 40% de valores faltantes foram excluídas da base de dados original, uma vez que a realização do preenchimento desses valores tenderia a enviesar os dados para um determinado padrão que não pode ser comprovado. Outras variáveis com valores faltantes foram preenchidas segundo a moda (categóricas), média (numéricas) ou ainda com relações existentes entre estas e outras variáveis (e.g. *Ethnicity* e *Race category*).

Testes foram feitos também com a utilização do método de seleção de variáveis Relief-F. A utilização do método visa a diminuição da quantidade de variáveis, extraindo-se aquelas que não contribuem informativamente com as possíveis classes. O Relief-F é um método iterativo que diferencia as variávei, identificando-as como fracamente ou fortemente relevantes, de acordo com comparações feitas entre as observações.

Mais de 700 dados da base se enquadravam fora dos limites das classes de interesse à pesquisa, ou seja, possuíam características de não-recorrência e tempo menor que 5 anos. Por esse motivo, esses dados foram retirados da base, uma vez que saíam do objetivo principal da pesquisa. Com isso, restaram 344 dados para analisar, número relativamente pequeno para ter alguma representatividade no problema em questão.

Para ultrapassar esse obstáculo e ainda solucionar o problema do desbalanceamento das classes, dois métodos de superamostragem foram utilizados: a superamostragem por replicação e por SMOTE-NC.

O desbalanceamento dos dados influencia negativamente grande parte dos métodos de classificação, então, no método da replicação, os dados das classes menores são replicados, a fim de tornar a base mais balanceada, e tornar os resultados mais robustos. Na tabela I é mostrada a base antes e depois da superamostragem por relicação.

TABELA I. BASE DE DADOS PRÉ E PÓS-SUPERAMOSTRAGEM POR REPLICAÇÃO

Classes	**Quantidade de dados**		
	Pré-superamostragem	***Pós-superamostragem***	***replicações***
Early	62	124	1
Mid	61	122	1
Late	22	110	4
Survival	199	199	0
Total	344	555	

Outra alternativa foi a utilização do SMOTE-NC. Essa técnica sintetiza novos dados da classe minoritária (com menos dados) através de uma relação dos dados dessa classe com seus vizinhos mais próximos. Na tabela 2, vemos um exemplo da técnica, que foi generalizada para lidar com atrinutos contínuos e categóricos:

TABELA II. PSEUDO-CÓDIGO DA GERAÇÃO DE NOVAS AMOSTRAS UTILIZANDO SMOTE-NC

F1 = 1 2 3 A B C (seja esta uma amostra para a qual calcularemos os vizinhos mais próximos)
F2 = 4 6 5 A D E
F3 = 3 5 6 A B K
A distância euclidiana entre F2 e F1 seria:
Eucl = raiz[(4-1)2 + (6-2)2 + (5-3)2 + Med2 + Med2]
Med é a mediana dos desvios padrões dos atributos contínuos das classes minoritárias
O termo da mediana é incluído duas vezes para os parâmetros de número 5: B→D e 6: C→E, que diferem para os dois vetores de parâmetros: F1 e F2

Com essa técnica, todas as classes foram balanceadas e por fim passaram a ter a mesma quantidade de dados, como mostra a tabela 3.

TABELA III. BASE DE DADOS PRÉ E PÓS-SUPERAMOSTRAGEM POR SMOTE-NC

Classes	**Quantidade de dados**	
	Pré-superamostragem	***Pós-superamostragem***
Early	62	199
Mid	61	199
Late	22	199
Survival	199	199
Total	344	796

C. *Classificações*

A classificação do banco de dados foi feita com a utilização de 3 métodos já bem conhecidos e consolidados: Árvore de Decisão, K vizinhos mais próximos (KNN) e Random Forest.

1) *Árvore de Decisão*

Esse método é hoje em dia um dos mais utilizados por possuir fácil implementação e apresentar resultados interpretáveis, adicionando um interesse a mais no método. É capaz de discriminar as classes realizando a divisão do espaço definido pelos atributos em sub-espaços, e estes por sua vez, sendo associados a uma classe.

É estruturado conforme uma árvore, possuindo nós, ramos e folhas. Os nós representam testes de um atributo, onde este é escolhido de forma a separar os padrões em classes distintas da melhor forma possível. Os ramos representam o resultado do teste feito nos nós, e as folhas são o "final da árvore", representando as classes.

A árvore aplicada na pesquisa utiliza como critério de avaliação da divisão dos ramos, o índice GINI. Além disso, faz as divisões dos ramos considerando a melhor divisão. Não foi utilizada a poda da árvore para alcançar melhores resultados.

2) *Vizinhos mais próximos (KNN)*

O KNN é um método de aprendizado baseado em instância, que utiliza a distância entre os dados para gerar o conhecimento sobre a estrutura deles. A Fig.1 Mostra graficamente como a quantidade de vizinhos a serem considerados podem alterar a classificação com esse método.

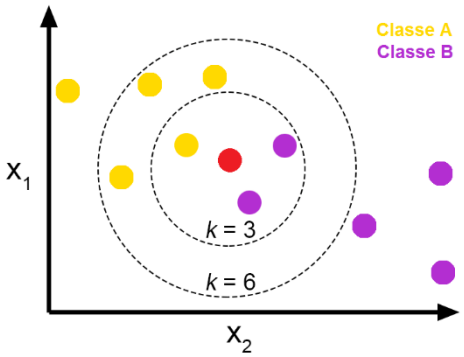

Fig. 1. Exemplo gráfico do KNN. Com 3 vizinhos, a bola rosa seria classificada como Classe B. Com 6 vizinhos, seria classificada como Classe A

3) *Random Forest*

Consiste em um método *ensemble* onde um grupo de classificadores combinam resultados individuais, formando um "consenso" ou opinião agregada. No caso do Random Forest, é um conjunto de árvores de decisão geradas aleatoriamente, e seus resultados são combinados para que se tenha uma saída única.

No trabalho, uma floresta com 100 árvores foi utilizada, fazendo uso do critério Gini para medir a qualidade da divisão realizada, assim como utilizado na árvore de decisão. Por fim, foi utilizada a seleção de classificadores como um dos parâmetros da *random forest*, indicando que cada componente da árvore era treinado em partes específicas do espaço de recursos. Portanto, cada componente da floresta se torna "especialista" nos padrões de uma tarefa simplificada.

III. RESULTADOS

Neste capítulo, os diversos resultados serão expostos, identificando características importantes de cada um deles. É importante salientar que os experimentos foram realizados levando-se em conta uma validação cruzada de 10 *folds*, e por conta da pequena quantidade de dados, um conjunto de teste não foi utilizado. As métricas consideradas no resultado dos experimentos foram: acurácia média (MA), macro precisão média (MMP), macro *recall* médio (MMR) e macro F1-*score* médio (MMF1). As métricas "micro", também muito vistas em análises de problemas com mais de duas classes, não foram utilizadas pois todas as classes são consideradas nos cálculos, e portanto todas essas métricas teriam o resultado equivalente à acurácia média. Por isso, optou-se por utilizar a acurácia média mais as métricas "macro", fornecendo um resultado mais amplo de cada experimento.

A. *Dados originais*

Inicialmente os experimentos foram feitos com os dados originais, após a limpeza e métodos de pré-processamento essenciais, como análise e preenchimento de valores faltantes e filtros de variância, restando 344 dados disponíveis para treinamento. Estes passos de pré-processamento foram realizados em todos os experimentos, antecedendo a utilização da seleção de variáveis ou superamostragem.

Os resultados do experimento inicial, expostos na tabela 4, foram aquém do esperado. Baixos valores nas métricas mostram que o banco de dados possui características relativamente difíceis de serem classificadas. O método que obteve melhores resultados foi o *Random Forest*, mas ainda assim, aquém do esperado para ser considerada uma boa classificação em um problema, principalmente na área da saúde.

TABELA IV. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO DO BANCO DE DADOS ORIGINAL

Método	**Métrica**			
	MA	**MMP**	**MMR**	**MMF1**
Árvore de Decisão	66,87	44,50	43,85	42,60
KNN	47,64	28,49	26,09	25,50
Random Forest	**75,61**	**50,86**	**51,23**	**50,08**

B. *Dados originais + Relief-F*

A utilização da seleção de atributos foi responsável por pequenas melhores nos métodos do KNN e *Random Forest*, mas ainda assim, os resultados não sofreram uma alteração muito consistente, aumentando em poucos pontos percentuais.

Os resultados são mostrados na tabela 5, onde novamente o melhor desempenho alcançado foi com o método *Random Forest*. Este é um padrão que será repetido em todos os testes feitos.

TABELA V. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO DO BANCO DE DADOS ORIGINAL COM SELEÇÃO DE ATRIBUTOS POR RELIEF-F

Método	**Métrica**			
	MA	**MMP**	**MMR**	**MMF1**
Árvore de Decisão	63,97	40,81	43,94	41,48
KNN	51,49	32,71	31,59	31,39
Random Forest	**77,61**	**55,02**	**54,27**	**53,52**

C. Dados com superamostragem - Replicação

A utilização da superamostragem alterou de forma drástica o resultado dos métodos de classificação. Em especial da Árvore de Decisão e R*andom Forest*. Uma melhora muito grande pôde ser observada com essas técnicas.

Os resultados são expostos na tabela 6, e mais uma vezo *Random Forest* obteve as melhores métricas.

TABELA VI. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO DO BANCO DE DADOS COM SUPERAMOSTRAGEM POR REPLICAÇÃO

Método	**Métrica**			
	MA	**MMP**	**MMR**	**MMF1**
Árvore de Decisão	96,94	96,88	97,86	97,19
KNN	49,90	48,90	53,07	49,13
Random Forest	**99,46**	**99,45**	**99,62**	**99,52**

D. Dados com superamostragem – SMOTE

Esta técnica de superamostragem também foi capaz de melhorar os resultados da classificação se comparado aos primeiros experimentos sem superamostragem. Foi ainda capaz de melhorar o desempenho do classificador KNN de forma considerável, o que os outros métodos não realizaram com a mesma eficácia.

Os resultados obtidos nos experimentos serão discutidos no capítulo de conclusão. Para este experimento, podem ser observados na tabela 7.

TABELA VII. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO DO BANCO DE DADOS COM SUPERAMOSTRAGEM SMOTE

Método	**Métrica**			
	MA	**MMP**	**MMR**	**MMF1**
Árvore de Decisão	77,28	77,87	77,28	77,10
KNN	68,98	68,10	68,97	64,56
Random Forest	**91,00**	**90,63**	**90,10**	**89,89**

E. Dados com SMOTE + Relief-F

Por fim, um último experimento foi realizado, unindo a técnica do SMOTE com a seleção de atributos Relief-F. Era esperado que desse modo, os benefícios de um espaço de variáveis com dimensões menores, aliado a um número maior de dados, pudesse gerar resultados promissores, principalmente para um banco de dados tão pequeno. De fato, isso ocorreu, porém apenas para a Árvore de Decisão. Os outros métodos continuaram com resultados similares ao experimento anterior.

Ainda assim, o método que obteve melhores métricas foi o *Random Forest*. Os resultados deste experimento podem ser observados na tabela 8.

TABELA VIII. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO DO BANCO DE DADOS COM SMOTE + RELIEF-F

Método	**Métrica**			
	MA	**MMP**	**MMR**	**MMF1**
Árvore de Decisão	81,18	81,28	81,17	80,94
KNN	68,34	68,95	68,32	65,65
Random Forest	**89,34**	**89,90**	**89,35**	**89,10**

F. Considerações

Embora as métricas alcançadas em alguns experimentos sejam muito boas e talvez até surpreendentes, é muito importante salientar que os métodos não foram testados em um grupo separado de dados. Além disso, as técnicas de superamostragem, em especial a técnica de replicação, podem gerar um viés nos padrões existentes na base de dados, fazendo com que os métodos sejam capazes de classifica-los com maior facilidade, sobretudo métodos como Árvores de Decisão e *Random Forest*, que se utilizam de heurísticas que levam em conta os atributos que melhor dividem as classes (homogeneidade, entropia). Isso de forma alguma faz o modelo ser mais preciso quando for utilizado em um problema real, uma vez que os dados de teste teriam um comportamento provavelmente diferente daquele gerado pela replicação dos dados, ou mesmo pela sintetização de novos.

É de extrema importância que mais dados sejam coletados para que seja possível realizar uma divisão em treino e teste, fazendo uso dos métodos disponíveis e testando-os em observações nunca antes "vistas" pelo método.

IV. CONCLUSÃO E TRABALHOS FUTUROS

O objetivo da pesquisa era a classificação de pacientes em um banco de dados fornecido pelo INCA, com respeito à recorrência do câncer de mama. Inicialmente, foram realizados diversos métodos de pré-processamento para organizar e limpar a base de dados, retirando atributos que não seriam utilizados na pesquisa e que prejudicariam a classificação.

Após essa limpeza, por conta do escopo da pesquisa, uma quantidade pequena de dados foi utilizada, e classes desbalanceadas foram formadas. Nos primeiros experimentos, nota-se a dificuldade da classificação em uma base com essas características. Uma solução proposta foi a utilização das técnicas de superamostragem por replicação e SMOTE, com o objetivo de aumentar a base e diminuir o desbalanceamento existente.

Os resultados mostraram que a utilização dessas técnicas foi substancial para uma melhor classificação dos pacientes. A técnica de seleção de variáveis Relief-F foi também utilizada, porém não gerou melhoras significativas nos resultados obtidos. Devido à pouca quantidade de dados, utilizar uma menor quantidade de atributos resultou também em uma classificação menos acurada.

Dentro dos algoritmos de classificação utilizados, ficou claro o melhor desempenho do *Random Forest*, um método ensemble que se utiliza de várias árvores de decisão e une os resultados de cada uma. Em todos os experimentos seu desempenho superou o dos outros classificadores.

Para trabalhos futuros é importante ressaltar a necessidade da coleta de mais dados, para que seja possível realizar uma classificação mais robusta, de preferência sem a necessidade de utilizar a superamostragem, uma vez que essa técnica tem a desvantagem de replicar padrões já existentes, ou sintetizar novos dados, possivelmente enviesando o resultado. Além disso, com mais dados seria possível realizar a divisão da base em treino e teste, para avaliar o real desempenho do modelo, o que é essencial para qualquer solução com modelos de previsão ou classificação.

REFERENCES

[1] X. Yao, Y. Liu, “Neural networks for breast cancer diagnosis”, Proceedings of the 1999 Congress on Evolutionary Computation. 1999

[2] T. Takeshita, L. Yan, M. Asaoka, O. Rashid, K. Takabe, “Late recurrence of breast cancer is associated with pro-cancerous ummune microenvironment in the primary tumor”, Scientific Reports. Nature Research. 2019

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall e W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique”. Journal of Artificial Intelligence Research 16. 2002.