Previsão do Período de Recorrência de Câncer de Mama

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 the breast cancer diagnosis problem and uses a Multi-Layer Perceptron 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 will be explained and the results of the classification using them will be parsed. Due to a small dataset, prediction was not able to perform as expected in a medical problem, but improvements from baseline classification could be seen through the application of pre-processing methods and hyper-parameter tuning.**

Keywords—Multi-Layer Perceptron, MLP, Predicting, Breast Cancer.

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 procura expandir o problema do diagnóstico do CM, utilizando uma das diversas técnicas de ML, o Multi-Layer Perceptron (MLP), para a previsão da recorrência do CM segundo o período de recorrência definido em [2]. A pesquisa é feita com base em um banco de dados fornecido pelo INCA. Um desafio em particular do problema é a pouca quantidade de dados disponíveis para o aprendizado.

Ainda, será feita a comparação entre os experimentos realizados durante os estudos, demonstrando os efeitos da alteração de diferentes hiper parâmetros da MLP [3], e como técnicas de pré-processamento podem tanto auxiliar na melhora dos resultados quanto ter o efeito oposto dependendo da ocasião e aplicação.

Os experimentos da pesquisa foram feitos no Weka, um software *open source* de *machine learning*, em conjunto com algoritmos executados em Python e R.

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, seleção de atributos e normalização. 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, e trará sugestões sobre outras possíveis abordagens e/ou soluções para o problema em foco.

II. METODOLOGIA

Neste capítulo, todo o desenvolvimento do trabalho será exposto, incluindo maiores detalhes sobre o banco de dados utilizado, técnicas de pré-processamento testadas, técnicas de seleção de atributos, detalhes das redes MLP utilizadas e seus hiper parâmetros.

A. Banco de dados

Neste tópico, será exposto as características do banco de dados e como algumas delas podem influenciar positivamente ou negativamente uma classificação utilizando MLP. 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 redundâncias, pois não seriam de ajuda na utilização do MLP. 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) que indicava 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: normalização *one-hot encoding* (1-N), e normalização dos dados numéricos (min-max), seleção de variáveis com Relief-F e Lasso Multinomial [4], e superamostragem.

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 de classificação foram feitos tanto com a utilização da normalização 1-N quanto sem essa normalização. Apenas uma variável categórica foi considerada numérica em todos os casos (*Neoplasm Disease Stage American Joint Committee on Cancer Code)*, por conter mais de 12 categorias sequencialmente organizadas. Todas as categorias numéricas foram normalizadas entre 0 e 1 através da normalização min-max.

Testes foram feitos também com a utilização de dois métodos de seleção de variáveis, o Relief-F e o Lasso Multinomial. A utilização dos métodos visava a diminuição da quantidade de variáveis, extraindo-se aquelas que não contribuíam informativamente com as possíveis classes. O Relief-F é um método iterativo que diferencia as variáveis e as ordena com uma nota, de acordo com comparações feitas entre observações. O Lasso Multinomial é uma técnica de regressão, onde para cada classe, identifica os coeficientes dos atributos, sendo os atributos mais importantes aqueles com maiores coeficientes.

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 extraídos, uma vez que saíam do objetivo principal da pesquisa. Com isso, restaram 344 dados para analisar, número pequeno para uma tarefa de classificação com redes neurais.

Para ultrapassar esse obstáculo, outro método de pré-processamento utilizado foi a superamostragem. Como os dados se encontravam desbalanceados, característica que influencia negativamente o desempenho da classificação de um MLP, os dados de classes menores foram 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.

TABELA I. BASE DE DADOS PRÉ E PÓS-SUPERAMOSTRAGEM

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	

C. Multi-Layer Perceptron

A rede MLP foi utilizada para realizar a classificação dos dados já mencionados e processados. Diversos experimentos foram feitos com o objetivo de alcançar o melhor resultado, alterando tanto a configuração da rede MLP, e seus neurônios na camada escondida, as variáveis de entrada, utilizando normalização ou não, e ainda, testando diferentes combinações de hiper parâmetros.

Como se sabe, o MLP é um aproximador universal, e seu desempenho de aproximação dependerá da correta configuração dos parâmetros mencionados, em conjunto com um pré-processamento eficaz dos dados utilizados. Para a definição de alguns hiper parâmetros, algumas métricas e heurísticas já são conhecidas. A métrica de Hecht-Nielsen diz que, para N_h, sendo o número de neurônios na camada escondida e N_i o número de entradas da rede, temos que $N_h \leq 2 * N_i + 1$. A métrica de Baum-Haussler mostra que, para N_o sendo o número de neurônios na camada de saída, ε o erro desejado no teste e N o número de padrões na base, $N_h \leq \frac{N*\varepsilon}{N_i+N_o}$. Outras heurísticas conhecidas foram consideradas enquanto os experimentos estavam sendo executados, como $N_h = \frac{N_i+N_o}{2}$, $N_h = N_i$, $N_h = N_o$ e por fim $N_h = N_i + N_o$. Essas heurísticas, apesar de não possuírem provas, vêm sendo aceitas como um guia principalmente pelos resultados de testes empíricos.

As configurações utilizadas serão detalhadas nos experimentos mencionados no próximo capítulo, de resultados, acompanhados do desempenho da classificação para os melhores resultados obtidos de acordo com pré-processamentos utilizados (com/sem seleção de atributos, com/sem hiper amostragem, com/sem normalização binária) e a quantidade de neurônios na camada escondida.

Além disso, como os experimentos foram conduzidos no software Weka, as ativações da camada de saída utilizadas foram sigmoides compreendidas no intervalo [0,1].

III. Resultados

Neste capítulo, serão expostos os resultados dos experimentos com melhor desempenho, com base nas configurações utilizadas. É importante salientar que todos os experimentos foram realizados com uma validação cruzada de 10 *folds* e por conta da pequena quantidade de dados, um conjunto de teste não foi utilizado, uma vez que qualquer diminuição da quantidade de dados utilizados acarretaria em uma piora importante de desempenho do modelo.

Os experimentos foram feitos com parâmetros-base, nos quais a taxa de aprendizado é fixa em 0,3, o momento utilizado é de 0,2 , o treinamento ocorre durante 1000 épocas e a validação é realizada em 20% dos dados.

A. *Experimentos sem superamostragem*

Inicialmente, os experimentos foram feitos sem a utilização da técnica de superamostragem, utilizando então os 344 dados disponíveis após a limpeza e pré-processamentos do banco de dados.

1) Sem one-hot encoding:

O melhor resultado obtido na classificação sem a utilização da superamostragem foi também sem a utilização da normalização 1-N para as variáveis categóricas.

Como mostra a Tabela II, foi alcançada uma acurácia de aproximadamente 78,2%, porém, na figura 1, pode-se observar a matriz de confusão da classificação realizada. Nota-se que a classe "Late", representada pela letra c, não foi classificada uma única vez. Isso ocorre por conta do desbalanceamento da base, prejudicando a classificação dos grupos menos representativos. Nesse caso, de 344 dados ao todo, apenas 22 eram representados por esse grupo. Os hiper parâmetros utilizados para obter esse resultado foram:

- Número de neurônios de entrada: 18
- Número de neurônios na camada de saída: 4
- Número de neurônios na camada escondida: 11

TABELA II. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO SEM A UTILIZAÇÃO DE SUPERAMOSTRAGEM E NORMALIZAÇÃO 1-N

		Desempenho de classificação
Acurácia		78,1977%
MAE		0,165
RMSE		0,2788
Precisão	Classe a	0,705
	Classe b	0,494
	Classe c	-
	Classe d	0,908

```
  a   b   c   d   <-- classified as
 31  24   0   7 |   a = 0
 11  41   0   9 |   b = 1
  1  17   0   4 |   c = 2
  1   1   0 197 |   d = 3
```

Fig. 1. Matriz de confusão da classificação sem a utilização de superamostragem e normalização 1-N

Os métodos de seleção de variáveis não auxiliaram para um melhor desempenho. A utilização do Relief-F resultou em um desempenho igual ao resultado mencionado, com uma acurácia de aproximadamente 78,2%. A utilização do Lasso Multinomial, da mesma forma, não melhorou o desempenho da classificação e o desempenho foi pior que o mostrado nos resultados, atingindo uma acurácia de aproximadamente 77,9%.

2) Com one-hot encoding:

O melhor desempenho utilizando a normalização 1-N obteve um desempenho menor que o caso anterior, alcançando uma acurácia de aproximadamente 77,6%

A tabela III mostra mais detalhadamente as métricas de desempenho para esse experimento, qye teve como hiper parâmetros:

- Número de neurônios de entrada: 18
- Número de neurônios na camada de saída: 4
- Número de neurônios na camada escondida: 16

TABELA III. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO SEM A UTILIZAÇÃO DE SUPERAMOSTRAGEM E COM NORMALIZAÇÃO 1-N

		Desempenho de classificação
Acurácia		77,6163%
MAE		0,1673
RMSE		0,2835
Precisão	Classe a	0,647
	Classe b	0,487
	Classe c	-
	Classe d	0,912

Pode-se perceber que o desempenho alcançado foi ligeiramente pior. A maior quantidade de atributos, nesse caso, piorou a capacidade da rede classificar os dados. A seguir são expostos os resultados dos melhores experimentos realizados utilizando a superamostragem.

Novamente, analisando a matriz de confusão na fig.2 observa-se que a classe c não foi reconhecida.

```
  a   b   c   d   <-- classified as
 33  23   0   6 |   a = 0
 14  38   0   9 |   b = 1
  1  17   0   4 |   c = 2
  3   0   0 196 |   d = 3
```

Fig. 2. Matriz de confusão da classificação sem a utilização de superamostragem e com normalização 1-N

B. *Experimentos com superamostragem*

Após a primeira etapa de experimentos, a técnica da superamostragem foi utilizada, com o intuito de diminuir a escassez dos dados e tornar possível a detecção da classe c, que possuía poucos dados.

1) *Sem one-hot encoding*

O melhor resultado sem a utilização da normalização 1-N obteve acurácia de aproximadamente 81,6%. Analisando esse resultado, observa-se que houve uma melhora com relação aos experimentos que não utilizaram a superamostragem. Isso ocorreu devido ao aumento de dados para serem aprendidos pela MLP. A tabela IV apresenta as métricas de desempenho detalhadas para a rede utilizada, que possuía os hiper parâmetros:

- Número de neurônios de entrada: 18
- Número de neurônios na camada de saída: 4
- Número de neurônios na camada escondida: 10

TABELA IV. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO COM A UTILIZAÇÃO DE SUPERAMOSTRAGEM E SEM NORMALIZAÇÃO 1-N

		Desempenho de classificação
Acurácia		81,6216%
MAE		0,1286
RMSE		0,2699
Precisão	Classe a	0,882
	Classe b	0,657
	Classe c	0,826
	Classe d	0,880

Nesse caso, a classe c pôde ser identificada e seus dados classificados. Já é visível a melhora no desempenho gerada pela superamostragem. A fig.3 mostra a matriz de confusão para essa configuração

```
  a   b   c   d   <-- classified as
 97  16   5   6 |   a = 0
 11  90   8  13 |   b = 1
  0  27  76   7 |   c = 2
  2   4   3 190 |   d = 3
```

Fig. 3. Matriz de confusão da classificação com a utilização de superamostragem e sem normalização 1-N.

2) *Com one-hot encoding*

Também foi testado, como nos experimentos sem superamostragem, a base com normalização categórica 1-N. Nesse experimento, os resultados foram ainda melhores que o anterior, alcançando uma acurácia de aproximadamente 83,2%, como mostra a tabela V. Os hiper parâmetros utilizados foram:

- Número de neurônios de entrada: 28
- Número de neurônios na camada de saída: 4
- Número de neurônios na camada escondida: 28

TABELA V. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO COM A UTILIZAÇÃO DE SUPERAMOSTRAGEM E COM NORMALIZAÇÃO 1-N

		Desempenho de classificação
Acurácia		83,2432%
MAE		0,1121
RMSE		0,2572
Precisão	Classe a	0,828
	Classe b	0,724
	Classe c	0,852
	Classe d	0,892

Na fig.4 observa-se a matriz de confusão para a classificação:

```
  a   b   c   d   <-- classified as
 96  17   6   5 |   a = 0
 12  92   6  12 |   b = 1
  2  11  92   5 |   c = 2
  6   7   4 182 |   d = 3
```

Fig. 4. Matriz de confusão da classificação com a utilização de superamostragem e com normalização 1-N.

Por fim, utilizando ainda os mesmos hiper parâmetros, um último experimento foi feito com o intuito de melhorar o desempenho de classificação. A "paciência" do *early stopping* foi aumentada para 100 épocas, enquanto anteriormente, apenas 5 épocas eram utilizadas. Em outras palavras, a partir de uma época na qual o desempenho de classificação dos dados de validação não melhorasse, o algoritmo testaria mais 4 épocas. Após a alteração, esse tempo de espera seria de 100 épocas ao todo. Isso pôde ajudar o algoritmo a treinar por mais tempo e melhorar seus resultados, uma vez que o desempenho da classificação dos dados de validação tende a melhorar, porém, não é uma melhora constante. Em outras palavras, é comum que de uma época para outra, o desempenho seja pior, e volte a melhorar posteriormente. O *early stopping* é utilizado para prevenir a piora consistente do treinamento.

Nesse último experimento, vê-se que o desempenho melhorou ainda mais, devido ao ajuste do parâmetro do *early stopping*. Alcançou-se uma acurácia de aproximadamente 86,6%, como pode ser visto na tabela VI. Observando a fig.5, nota-se que os erros de classificação foram em menor quantidade.

TABELA VI. MÉTRICAS DE DESEMPENHO NA CLASSIFICAÇÃO COM A UTILIZAÇÃO DE SUPERAMOSTRAGEM E COM NORMALIZAÇÃO 1-N

		Desempenho de classificação
Acurácia		88,6486%
MAE		0,0697
RMSE		0,2178
Precisão	Classe a	0,876
	Classe b	0,831
	Classe c	0,906
	Classe d	0,915

```
  a   b   c   d   <-- classified as
 96  17   6   5 |   a = 0
 12  92   6  12 |   b = 1
  2  11  92   5 |   c = 2
  6   7   4 182 |   d = 3
```

Fig. 5. Matriz de confusão da classificação com a utilização de superamostragem e com normalização 1-N. Paciência do *early stopping* aumentada para 100 épocas.

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, selecionando os atributos mais relevantes e significativos para uma classificação utilizando o MLP.

Após essa limpeza, por conta do escopo da pesquisa, uma quantidade pequena de dados é 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 da técnica de superamostragem, com o objetivo de aumentar a base e diminuir o desbalanceamento existente.

Os resultados mostraram que a utilização dessa técnica foi substancial para uma melhor classificação dos pacientes. Técnicas de seleção de variáveis também foram testadas, porém os resultados não foram satisfatórios. Devido à pouca quantidade de dados, utilizar uma menor quantidade de atributos resultou também em uma classificação menos acurada.

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, 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.

Por fim, como sugestão de uma nova abordagem, observando os resultados de todos os experimentos fica claro que a classe “Survival”, representada pela letra d, consegue ser diferenciada mais facilmente. Por conta disso, seria possível utilizar uma classificação em dois estágios. O primeiro diferenciando casos com recorrência e sem recorrência. E o segundo, tomando os casos com recorrência e classificando-os em relação ao período de recorrência.

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] B. Al-Shargabi, F. Al-Shami, R. S. Alkhawaldeh, “Enhancing multi-layer perceptron for breast cancer prediction”, International Journal of Advanced Science and Technology, 2019.

[4] T. Hastie, R. Tibishirani, M. Wainwright, “Statistical learning with sparsity the lasso and generalizations”, 2016.

[5] A. T. Azar, S. A. El-Said, “Probabilistic neural network for breast cancer classification”, Neural Computing and Applications, vol. 23, 2013