Predicting Breast Cancer Recurrence

Diego Patrik S. da Silva[1], Giuseppe F. Neto[1]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)
52.171-900 – Recife – PE – Brasil

dpatrikone@gmail.com, fiorentinogiuseppebcc@gmail.com

Abstract. *Breast cancer is the most common and the leading cause of cancer death among women worldwide. In Brazil, the National Cancer Institute (INCA) estimates that in 2018/2019 about 59.700 new cases will be diagnosed. Early stage detection and treatment are important to significantly reduce the chance of death. In such case, the worse fear of a patient is the recurrence of the cancer. This paper aims to predict whether a patient will face a recurrence using a multilayer perceptron artificial neural network based on the University Medical Centre breast cancer dataset.*

Resumo. *Câncer de mama é o tipo de câncer mais comum e o que mais mata mulheres em todo o mundo. No Brasil, o Instituto Nacional de Câncer (INCA) estima que para 2018/2019 sejam diagnosticados 59.700 novos casos. A detecção precoce e tratamento são muito importantes para reduzir significamente o risco de morte. Nesses casos, o maior medo para o paciente é a recorrência do câncer. Esse trabalho visa predizer se o paciente vai enfrentar uma recorrência usando uma rede neural artificial (RNA) do tipo Multilayer Perceptron (MLP) baseado no conjunto de dados de câncer de mama da University Medical Centre.*

1. Introdução

Segundo dados da Agência Internacional para a Pesquisa do Câncer, o câncer de mama é o tipo de câncer mais comum e que mais mata mulheres em todo o mundo. A previsão do Inca (Instituto Nacional de Câncer) é de que em 2018/2019 ocorram 59.700 casos de câncer de mama entre mulheres no Brasil. A detecção precoce aumenta significativamente a chance de sobrevivência do paciente que sofre dessa doença. Mas o maior problema é predizer a recorrência do câncer. Recorrência é quando o câncer volta a aparecer após o tratamento, no mesmo ou em outro lugar, podendo isso acontecer até 20 anos depois.

A análise de registros médicos já existentes permite que algoritmos de *machine learning* façam previsões sobre a saúde do paciente com um certo grau de certeza. O objetivo deste trabalho é usar classificação para predizer se um paciente vai enfrentar a recorrência do câncer.

A base de dados utilizada contém 10 atributos, incluindo o atributo classe, e leva em consideração aspectos como idade do paciente e tamanho do tumor, etc. Testamos esses dados da University Medical Centre em uma rede neural do tipo Multilayer Perceptron. A linguagem de programação Java foi utilizada para fins de implementação.

2. Base de dados

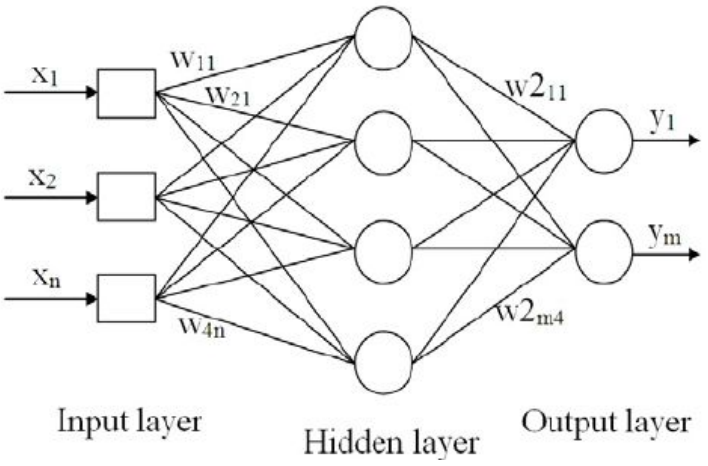
A base de dados usada neste estudo é fornecida pela University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia, por meio do repositório de *machine learning* da UCI. O conjunto de dados contém 10 atributos e um total de 286 instâncias. Consideramos todos os atributos nos testes:

1. Age: idade do paciente quando o diagnóstico foi realizado;
2. Menopause: status de menopausa do paciente;
3. Tumor size: tamanho do tumor (em mm);
4. Inv-nodes: número de glândulas linfáticas que transportam câncer metastático;
5. Node caps: se o tumor substitui os gânglios linfáticos e permite invadir tecidos próximos ou não;
6. Degree of malignancy: grau do tumor;
7. Breast: em qual seio o tumor foi diagnosticado;
8. Breast quadrant: o seio pode ser dividido em quatro quadrantes;
9. Irradiation: se o paciente foi submetido a terapia com radiação ou não;
10. Class: sem recorrência ou recorrência.

Atributos	Valores
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause	lt40, ge40, premeno
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
node-caps	yes, no
deg-malig	1, 2, 3
breast	left, right
breast-quad	left-up, left-low, right-up, right-low, central
irradiation	yes,no
class	no-recurrence-events, recurrence-events

3. Algoritmo

Neste trabalho, usamos uma rede neural artificial do tipo multilayer perceptron (MLP) para a classificação.



Uma rede neural artificial é uma ferramenta poderosa capaz de representar relações complexas de entrada-saída. Uma MLP é composta de 3 camadas: camada de entrada, camadas ocultas e camada de saída. O algoritmo de treinamento mais utilizado em modelos MLP é o Backprogation, que se baseia na aprendizagem por correção de erros. O algoritmo de Backprogation é um tipo de aprendizado supervisionado, quando o valor de saída é gerado o erro é calculado e seus valores são retro-propagados para entrada, os pesos são ajustados e os valores são novamente calculados.

Implementamos a MLP em Java, ajustando a quantidade de camadas, taxa de aprendizagem, o número de neurônios e definindo a sigmóide como função de ativação, tanto para a camada oculta, como para a camada de saída.

4. Resultados

A base de dados da University Medical Centre foi dividida em 60% para fins de treino e 40% para fins de teste. Fizemos experimentos alterando entre 0.0 e 1.0 a taxa de aprendizagem e medimos a performance da MLP baseado nos resultados de acurácia, precisão e relevância.

Onde

$$\text{acurácia} = \frac{Verdadeiros\ Positivos + Verdadeiros\ Negativos}{Número\ Total\ de\ Exemplos}$$

$$\text{precisão} = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Positivos}$$

$$\text{relevância} = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos}$$

Taxa de Aprendizado	Acurácia	Precisão	Relevância
0.0	28.9%	0.289	1.0
0.1	66.6%	0.407	0.333
0.2	69.2%	0.473	0.545
0.3	63.1%	0.395	0.515
0.4	**73.6%**	0.545	0.545
0.5	68.4%	0.463	0.575
0.6	67.5%	0.437	0.424
0.7	67.5%	0.433	0.393
0.8	72.8%	0.529	0.545
0.9	71.9%	0.514	0.545
1.0	68.4%	0.451	0.424

Testamos diferentes funções de ativação, obtivemos sempre melhores resultados com a função sigmóide.

Transfer function	acurácia	precisão	relevância	taxa de aprendizagem
sigmoidal	73.6%	0.545	0.545	0.4
hyperbolic	61.1%	0.410	0.636	0.1
heavyside	71.0%	0.333	0.030	0.4

5. Conclusão

Esse trabalho discute sobre a predição da recorrência de câncer de mama, usando uma rede neural e dados de uma base de dados como experimento. Como conclusão, alcançamos o objetivo de usar uma rede neural do tipo MLP para predizer a recorrência de um câncer, alcançando até 73% de acurácia. Um próximo passo seria definir quais dos atributos são mais relevantes.

6. Referências

Breast Cancer Data Set.
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Estatísticas para Câncer de Mama
http://www.oncoguia.org.br/conteudo/estatisticas-para-cancer-de-mama/6562/34/

Ahmad, Aamir (2013) “Pathways to Breast Cancer Recurrence”.
https://www.hindawi.com/journals/isrn/2013/290568/