Implementação de rede *Self Organizing Map* (SOM)

por Eduardo S. M. de Vasconcelos
NUSP 7656724 | eduardovasconcelos@usp.br

São Carlos, 21 de maio de 2016

• • •

Introdução

Este trabalho tem como objetivo servir como objeto de avaliação parcial na disciplina Redes Neurais (SCC0570), ministrada pelo Dr. Pablo Andretta Jaskowiak durante o primeiro semestre letivo de 2016 no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC/USP). Neste trabalho foi implementada uma rede do tipo *Self Organizing Map* (SOM).

Este documento é dividido em três seções, sendo esta *Introdução* a primeira delas, seguida da seção *Descrição do problema*, onde discute-se a base de dados utilizada, seu pré-processamento e as características gerais da implementação realizada. Na seção *Resultados* é discutida a performance da rede obtida. Especial destaque é dado ainda aos casos negativos (i.e. para os quais a rede não tem boa performance). Discute-se ainda as dificuldades encontradas na realização do trabalho.

Para referência sobre a execução do código, vide arquivo *"README.txt"*, que acompanha este documento.

Descrição do problema

Optou-se pela utilização da base de dados *Iris flower data set*[1], compilada por Ronald Fisher em 1936. Foi utilizada uma versão obtida do *site* da *Universidade de Illinois*[2]. O motivo da escolha dessa base de dados foi sua grande popularidade e consequente existência de muito material com o qual comparar os resultados obtidos pela rede implementada. A *Iris flower data set* é constituída de 150 exemplos e 3 espécies de flores do gênero Iris, a saber: *I. setosa*, *I. versicolor* e *I. virginica*. A base tem 50 exemplos de cada espécie e cada exemplo tem 4 parâmetros numéricos, representando o comprimento e a largura das sépalas e o comprimento e a largura das pétalas da flor, além de um rótulo identificando a espécie à qual o exemplo corresponde. Uma característica desta base de dados é que uma das espécies (*I. setosa*) é linearmente separável das outras duas. As espécies *I. versicolor* e *I. virginica*, no entanto, não são linearmente separáveis e parecem formar um único *cluster* ao olharmos para um gráfico de dispersão da base de dados, conforme pode ser visto na Figura 1[3]. Tal fato refletiu nos resultados da rede, conforme será mostrado adiante.

1 R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". *Annals of Eugenics* 7 (2): 179–188.doi:10.1111/j.1469-1809.1936.tb02137.x
2 http://mste.illinois.edu/malcz/DATA/BIOLOGY/Fisher's_Irises.html
3 Extraída de https://upload.wikimedia.org/wikipedia/commons/5/56/Iris_dataset_scatterplot.svg

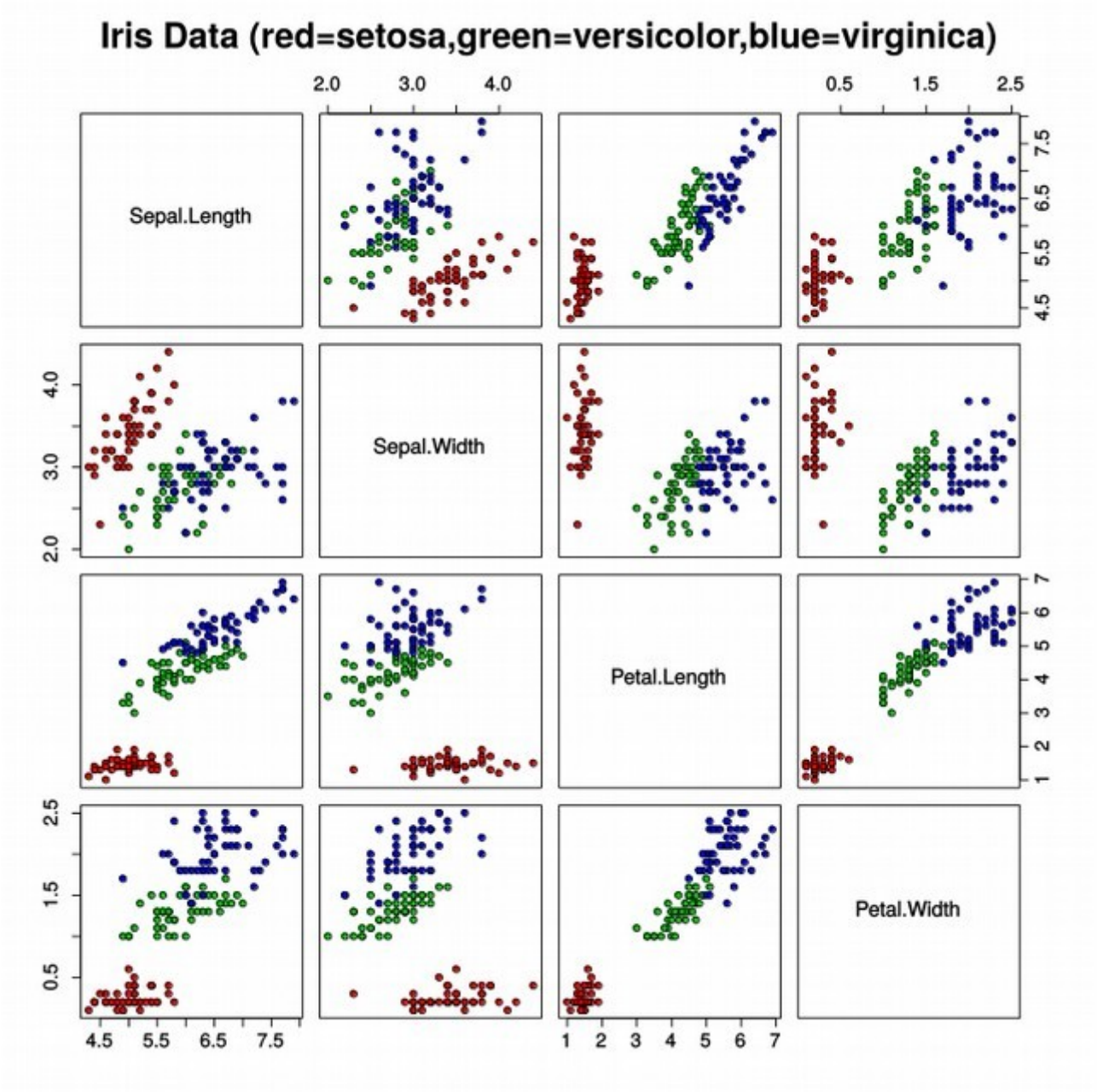

Figura 1. Gráfico de dispersão da base de dados *Iris data set*. Os parâmetros numéricos da base são *plotados* 2 a 2.

A base de dados não passou por nenhum tratamento prévio à execução da rede. Os dados simplesmente foram importados para uma matriz no ambiente de programação utilizado (Octave). Evidentemente, os rótulos foram suprimidos da matriz quando do treinamento da rede, uma vez que estes não devem ser utilizados no treinamento de uma rede SOM, que emprega aprendizado não supervisionado. Dessa forma, o conjunto de entrada X da rede é uma matriz 150x4, contendo os 150 exemplos da base e seus 4 parâmetros numéricos.

Quanto à arquitetura da rede, optou-se por realizar 3 testes, cada qual com uma arquitetura diferente. Optou-se pela estrutura de rede SOM vista em aula i.e. camada de entrada e retículo bidimensional na saída. O teste 1 foi realizado para um retículo bidimensional de lado 10 (portanto com 100 neurônios), o teste 2 foi realizado para um retículo bidimensional de lado 20 (portanto com 400 neurônios) e finalmente o teste 3 para foi realizado para um retículo bidimensional de lado 30 (portanto com 900 neurônios). A arquitetura da rede foi sempre representada por um vetor com dois

elementos, cada qual igual ao valor de lado escolhido para o retículo bidimensional da saída. Portanto, as arquiteturas dos 3 testes foram, respectivamente:

- Teste 1: (10, 10);
- Teste 2: (20, 20);
- Teste 3: (30, 30).

Todos os testes foram realizados com os seguintes parâmetros:

- Quantidade de épocas de treinamento (N): 100;
- Desvio-padrão inicial da gaussiana (σ_0): 0,5;
- Constante temporal de correção do desvio-padrão da gaussiana (t_1): 1,2;
- Taxa de aprendizado inicial (η_0): 0,5;
- Constante temporal de correção da taxa de aprendizado (t_2): 1,2.

Tanto o desvio-padrão da gaussiana quanto a taxa de aprendizado foram corrigidos exponencialmente em função da época de treinamento usando a fórmula:

$$p = p_0 \cdot \mathrm{e}^{-n/t}$$

Onde p representa o parâmetro a atualizar (σ ou η), n representa a época de treinamento e t é a respectiva constante de tempo (t_1 ou t_2).

Optou-se pelo uso de Octave para o desenvolvimento do trabalho. A linguagem foi escolhida pela vasta gama de operadores matriciais que implementa, o que permite focar na rede neural e escapar de problemas com programação de nível mais baixo.

Resultados

Conjunto de testes

Os resultados são listados de acordo com o teste realizado. Para cada teste é apresentado o mapa resultante, bem como um gráfico de validação, que mostra como estão distribuídas as classes no mapa obtido. Cada mapa é uma matriz de N linhas e N colunas (com N igual ao tamanho do lado do retículo bidimensional) na qual cada neurônio de saída é representado por uma posição. Sobre cada neurônio é desenhado um círculo vermelho de área proporcional ao número de exemplos de treinamento que o ativam. Foi definido um tamanho máximo para os círculos no mapa, de forma a torná-lo mais limpo. É provida ainda a matriz correspondente a cada mapa obtido, contendo os valores numéricos (i.e. número de exemplos que ativam o neurônio) correspondentes a cada posição do mapa.

TESTE 1 O teste 1 corresponde à rede com retículo bidimensional de lado 10. O mapa obtido é mostrado na Figura 2.

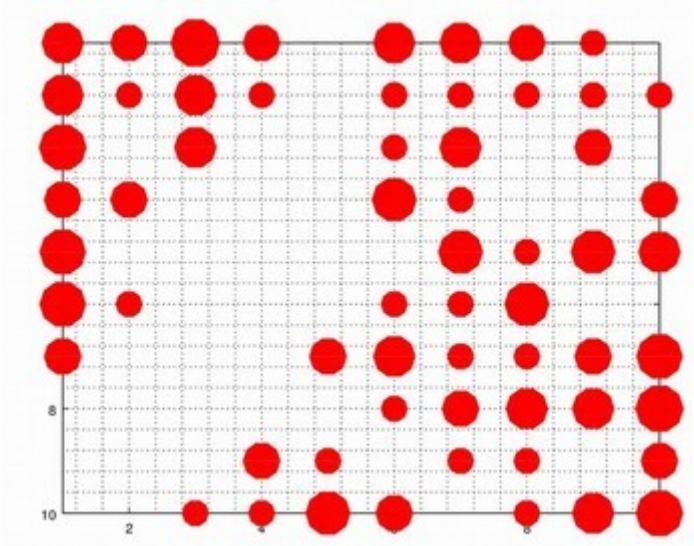

Figura 2. Mapa obtido para o Teste 1. Cada neurônio é representado por uma posição da matriz. Sobre cada neurônio é desenhado um círculo vermelho de área proporcional ao número de exemplos que o ativam.

Ao mapa apresentado na Figura 2 corresponde a matriz abaixo. Cada entrada da matriz representa o número de exemplos da base que ativam o neurônio correspondente à mesma posição no retículo bidimensional.

$$M_1 = \begin{bmatrix} 3 & 2 & 9 & 2 & 0 & 3 & 3 & 2 & 1 & 0 \\ 3 & 1 & 3 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 6 & 0 & 3 & 0 & 0 & 1 & 3 & 0 & 2 & 0 \\ 2 & 2 & 0 & 0 & 0 & 4 & 1 & 0 & 0 & 2 \\ 5 & 0 & 0 & 0 & 0 & 0 & 4 & 1 & 4 & 3 \\ 5 & 1 & 0 & 0 & 0 & 1 & 1 & 4 & 0 & 0 \\ 2 & 0 & 0 & 0 & 2 & 3 & 1 & 1 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 3 & 7 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 & 4 & 2 & 0 & 1 & 3 & 6 \end{bmatrix}$$

Claramente, a rede estabeleceu uma fronteira tênue entre dois grupos distintos. O gráfico de validação do teste, mostrado na Figura 3, permite analisar com mais detalhes o resultado obtido. Ele mostra o mapa do retículo bidimensional novamente (com cada neurônio representado por uma posição na matriz), mas sobre ele agora estão representadas as espécies (rótulos) de cada exemplo de treinamento que acabou por ativar o neurônio em questão. Os triângulos vermelhos representam a espécie *I. setosa* e as espécies *I. versicolor* e *I. virginica* são representadas pelos quadrados verdes e pelas circunferências azuis, respectivamente.

Observa-se que, de fato, o grupo isolado obtido refere-se à *I. setosa*, linearmente separável dos demais. O segundo grupo obtido no mapa é constituído por exemplares das outras duas espécies. Por mais que não haja uma fronteira clara entre eles, os exemplares de *I. versicolor* tendem a ficar na borda do mapa, enquanto os exemplares de *I. virginica* tendem a se concentrar mais próximos à diagonal secundária, logo antes da fronteira com a *I. setosa*. Há uma certa sobreposição entre a *I. versicolor* e a *I. virginica*, como pode ser observado em algumas posições da Figura 3.

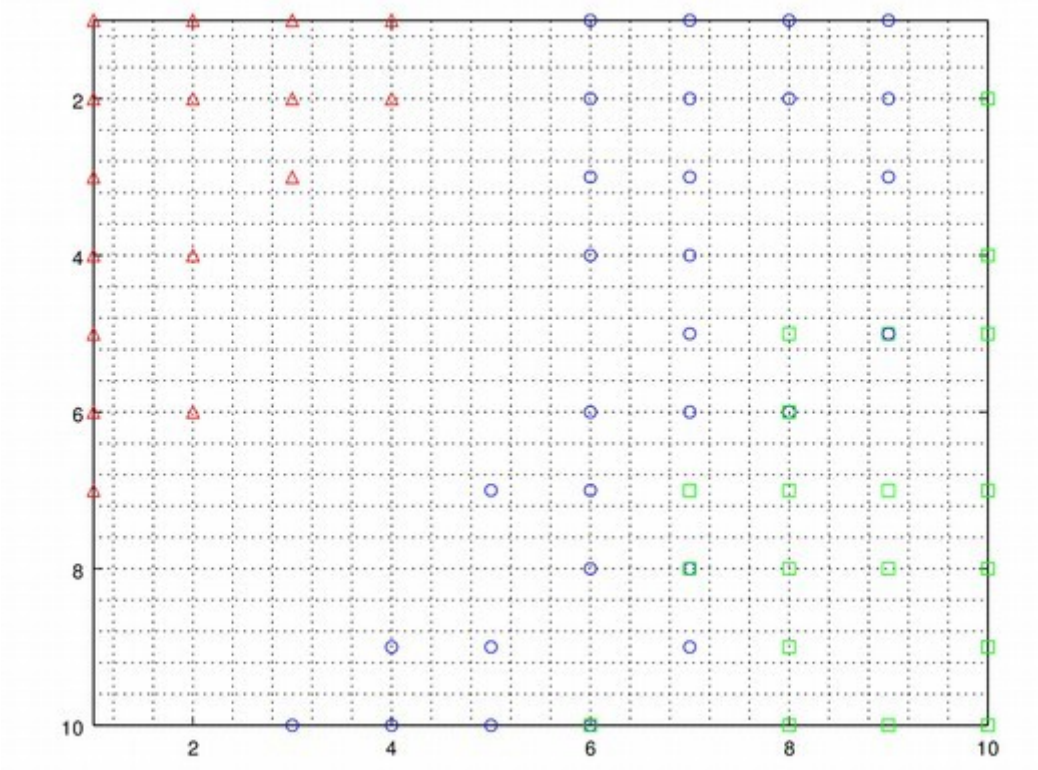

Figura 3. Gráfico de validação do Teste 1. Cada neurônio é representado por uma posição da matriz. Sobre cada neurônio é desenhado um símbolo correspondente a cada rótulo que o ativa. Os triângulos vermelhos representam a espécie *I. setosa*, os quadrados verdes representam a espécie *I. versicolor* e as circunferências azuis representam a espécie *I. virginica*.

TESTE 2 O teste 2 corresponde à rede com retículo bidimensional de lado 20. O mapa obtido é mostrado na Figura 4.

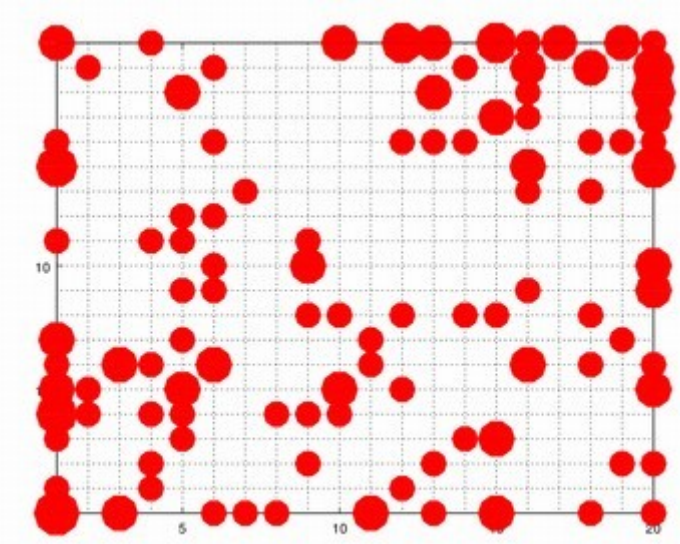

Figura 4. Mapa obtido para o Teste 2. Cada neurônio é representado por uma posição da matriz. Sobre cada neurônio é desenhado um círculo vermelho de área proporcional ao número de exemplos que o ativam.

Salvo pelas posições relativas entre os grupos que aparecem no mapa (cuja fronteira aparece mais alinhada com a diagonal principal do mapa do que com a diagonal secundária neste caso – o que não é de surpreender devido ao caráter não determinístico da inicialização do mapa), o resultado obtido no mapa do Teste 2 é muito similar ao obtido no Teste 1. A matriz correspondente ao mapa apresentado na Figura 4 é mostrada a seguir.

$$M_2 = \begin{matrix}
2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 3 & 2 & 0 & 3 & 1 & 2 & 0 & 2 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 2 & 0 & 3 \\
0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 & 2 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 4 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 2 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\
2 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\
3 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 2 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 2 & 0 & 1 & 0 & 2 & 0 & 0 & 1 & 0 & 1
\end{matrix}$$

Apareceram dois grupos separados por uma fronteira tênue. Para validar a solução, na Figura 5 é apresentada a distribuição das classes sobre o mapa em termos de seus rótulos. Os triângulos vermelhos representam a espécie *I. setosa* e as espécies *I. versicolor* e *I. virginica* são representadas pelos quadrados verdes e pelas circunferências azuis, respectivamente.

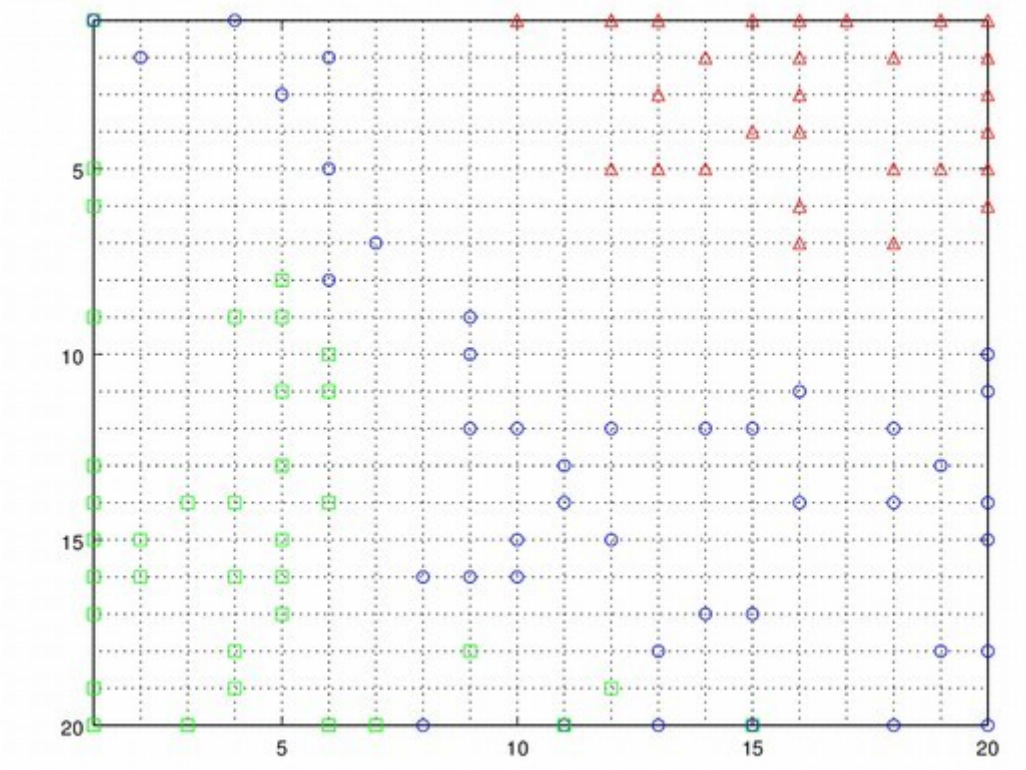

Figura 5. Gráfico de validação do Teste 2. Cada neurônio é representado por uma posição da matriz. Sobre cada neurônio é desenhado um símbolo correspondente a cada rótulo que o ativa. Os triângulos vermelhos representam a espécie *I. setosa*, os quadrados verdes representam a espécie *I. versicolor* e as circunferências azuis representam a espécie *I. virginica*.

Observa-se uma separação bastante clara entre os exemplares da espécie *I. virginica* e as outras duas espécies de íris, entras as quais, novamente, há uma certa sobreposição, ainda que ao observar o *plot* de validação exista separação entre elas no *cluster* que formam.

TESTE 3 O terceiro e último teste foi realizado com um mapa de lado 30. O mapa obtido é apresentado na Figura 6.

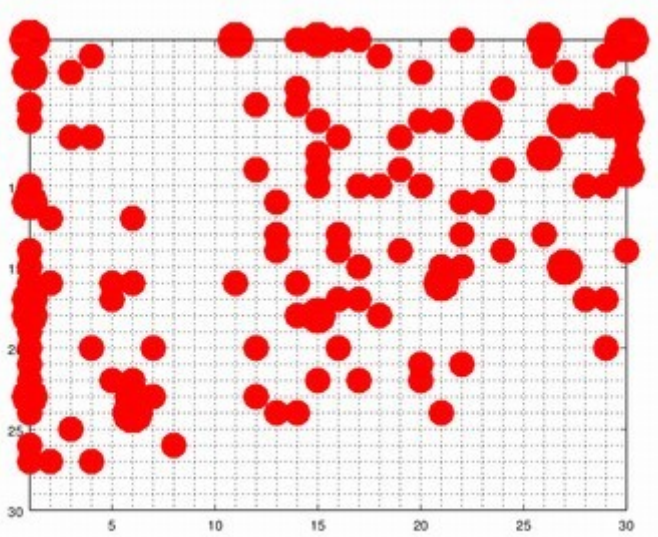

Figura 6. Mapa obtido para o Teste 3. Cada neurônio é representado por uma posição da matriz. Sobre cada neurônio é desenhado um círculo vermelho de área proporcional ao número de exemplos que o ativam.

O resultado obtido nos testes 1 e 2, aparentemente, se repete. Dois *clusters* apareceram mesmo num mapa maior, com 900 neurônios, e há uma clara fronteira entre eles. A matriz correspondente à distribuição da Figura 6 não é apresentada neste relatório[4]. O *plot* de validação do mapa obtido na Figura 6 é apresentado na Figura 7. Os triângulos vermelhos representam a espécie *I. setosa* e as espécies *I. versicolor* e *I. virginica* são representadas pelos quadrados verdes e pelas circunferências azuis, respectivamente.

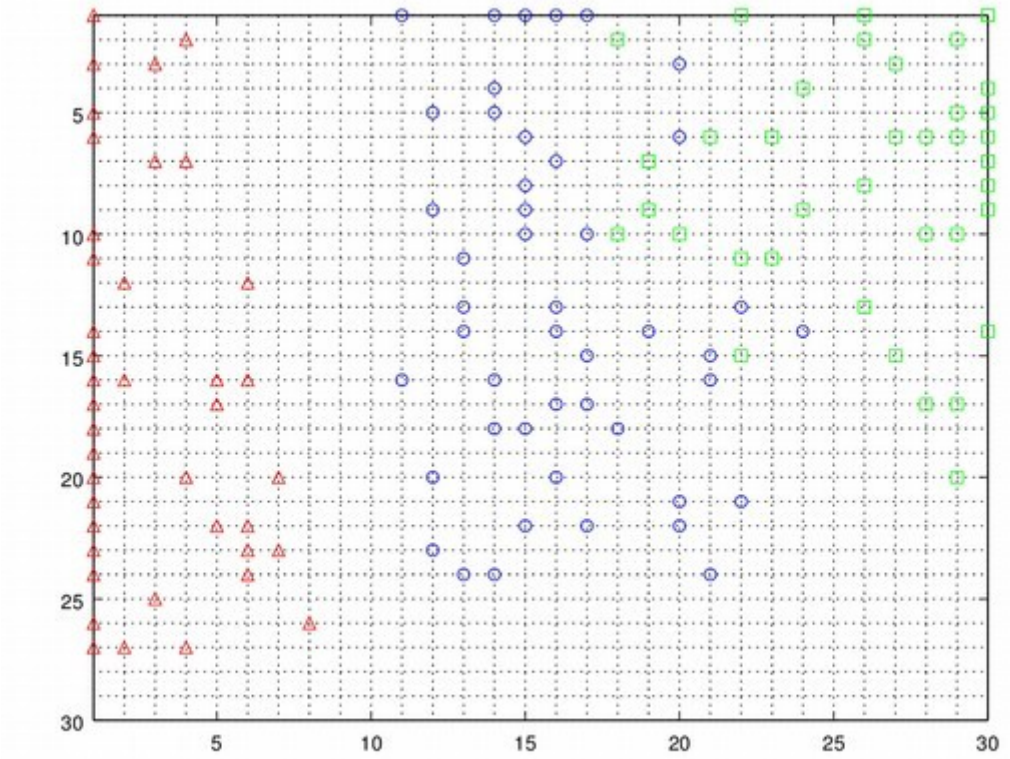

Figura 7. Gráfico de validação do Teste 3. Cada neurônio é representado por uma posição da matriz. Sobre cada neurônio é desenhado um símbolo correspondente a cada rótulo que o ativa. Os triângulos vermelhos representam a espécie *I. setosa*, os quadrados verdes representam a espécie *I. versicolor* e as circunferências azuis representam a espécie *I. virginica*.

Observa-se que os exemplares da espécie *I. setosa* são separados por uma fronteira muito clara do *cluster* formado pelos exemplares das outras duas espécies. Novamente, os representantes das espécies *I. versicolor* e *I. virginica* separam-se entre eles, mas a partir da inspeção unicamente do mapa obtido para o teste (exibido na Figura 6), não é possível, pelo menos para o autor deste documento, distinguir uma fronteira dentro do *cluster* que formam, isto é, uma fronteira de separação entre as espécies *I. versicolor* e *I. virginica*. Nesta configuração 30x30 do retículo bidimensional da rede, observa-se que não há sobreposição com respeito a classes que ativam nenhum neurônio, ou seja, cada neurônio especializou-se em ativar perante exemplos de uma única classe, o que representa uma vantagem com relação às configurações 10x10 e 20x20.

Casos negativos

A avaliação dos resultados obtidos mostra que a rede obtém performance razoável ao separar as classes em um mapa, no entanto, a fronteira entre as classes não é sempre clara. Independente da variação no tamanho do mapa nos casos testados, observa-se que dois *clusters* com fronteira muito

4 Devido ao seu tamanho (900 elementos), a matriz correspondente à Figura 6 tornaria o texto poluído e pouca informação poderia ser retirada de sua inspeção. No entanto, se o leitor desejar vê-la, ela pode ser encontrada no arquivo de serialização do Octave *N=100_30x30/simobj.txt*, distribuído junto a este documento (sob o nome de objeto *mat_y*).

tênue entre eles são formados: um que corresponde à espécie *I. setosa* apenas e outro que corresponde às espécies *I. versicolor* e *I. virginica*. Dentro deste, o gráfico de validação do resultado sempre mostra que as espécies estão separadas **dentro** do *cluster*, mas isso não é visível, ou ao menos não é claro através da inspeção do mapa retornado pela rede (i.e. da configuração final do retículo bidimensional da rede) apenas, o que é um aspecto negativo do resultado obtido.

A maior dificuldade de realização deste trabalho, assim como no primeiro projeto desenvolvido nesta disciplina, foi encontrar valores “bons” para os parâmetros de configuração para os quais a rede SOM realizasse a tarefa desejada, em particular os parâmetros σ_0, η_0, t_1 e t_2.