UNIVERSIDADE FEDERAL DO ESTADO DO RIO GRANDE DO NORTE

Disciplina / Área de conhecimento:

ECT2702 - TÓPICOS AVANÇADOS EM INFORMÁTICA I - T01 (2019.2)

SOM -Self-Organizing Map

Luís Fernando Tavares

Natal, 2019

**Sumário**

1 Introdução……………………...………….........................…………..............................………......03

2 Metodologia.……………...…………………...................................................................…......04

3 Códigos…………………….......................................................................................……......06

4 Experimentos………………...........................................................….....................……......11

02

**1 Introdução**

O problema abordado neste relatório é: uma base de dados que contém registro de alunos da disciplina de LOP (Lógica de Programação), contendo informações sobre as submissões de atividades feitas por semana e sua situação na matéria, o objetivo é saber se a relação entre as entregas das atividades e a situação do aluno (aprovado ou reprovado), de acordo com os dados obtidos do semestre.

A base de dados possui diversos dados coletados dos alunos durante o semestre, tais como:

* semana 1~21: Possuem a quantidade de atividades que foram submetidas naquela semana.
* situacao: Situação dos alunos no semestre (aprovado, aprovado por nota, reprovado, reprovado por media e falta, reprovado por nota).
* anoTurma: Ano que foi cursado a disciplina.
* turma: subturmas da disciplina

03

**2 Metodologia**

O modelo de Machine Learning utilizado para resolver este problema é o Mapa Auto-Organizado (MAO ou SOM — Self-Organizing Map), é um tipo de rede neural artificial (RNA) treinada usando aprendizado não supervisionado para produzir uma representação discreta de baixa dimensão (geralmente bidimensional) do espaço de entrada das amostras de treinamento, chamada de mapa, em outras palavras, ele cria um mapeamento de uma distribuição de alta-dimensão em uma grade regular de baixa-dimensão.

Os mapas auto-organizados diferem de outras redes neurais artificiais, pois aplicam aprendizado competitivo em oposição ao aprendizado de correção de erros (como retropropagação com descida gradiente) e no sentido de que eles usam uma função de vizinhança para preservar as propriedades topológicas do espaço de entrada.

As etapas de mapeamento do SOM começam com a inicialização dos vetores de peso. A partir daí, um vetor de amostra é selecionado aleatoriamente e o mapa de vetores de peso é pesquisado para descobrir qual peso melhor representa essa amostra. Cada vetor de peso tem pesos vizinhos que estão próximos a ele.

O peso escolhido é recompensado por se tornar mais parecido com o vetor de amostra selecionado aleatoriamente. Os vizinhos com esse peso também são recompensados ​​por se tornarem mais parecidos com o vetor de amostra escolhido.

É capaz de compactar informações preservando os relacionamentos topológicos e as métricas mais importantes dos dados originais.

O algoritimo funciona da seguinte forma:

Primeramente os pesos de cada nó são inicializados e um vetor é escolhido aleatoriamente no conjunto de dados de treinamento.

Cada nó é examinado para calcular quais pesos são mais parecidos com o vetor de entrada. O nó vencedor é conhecido como Melhor Unidade de Correspondência (MUC ou BMU — Best Matching Unit).

Então a vizinhança da BMU é calculada. A quantidade de vizinhos diminui com o tempo.

O peso vencedor é recompensado por se tornar mais parecido com o vetor de amostra. Os vizinhos também se tornam mais parecidos com o vetor de amostra. Quanto mais próximo um nó da BMU, mais seus pesos são alterados e mais distante o vizinho fica da BMU, menos ele aprende.

Por fim a etapa 1 é repitida até que um critério de parada não seja atingido.

As aplicações práticas da rede SOM são varias,alguns exemplos são: reconhecimento de voz ,organização automática de documentos numa biblioteca, visualização de registros financeiros etc.

04

Os atributos selecionados foram as semanas 1~21, pois indicam os números de submissões de atividades por semana feitas pelos alunos, para serem comparadas com a situação dos alunos, aprovados ou reprovados.

Esses atributos foram selecionados com o fim de saber se a relação entre o numero de submissões por semana das atividades solicitadas da matéria com a situação dos alunos (aprovação ou reprovação).

05

**3 Códigos**

**1º Importação:**



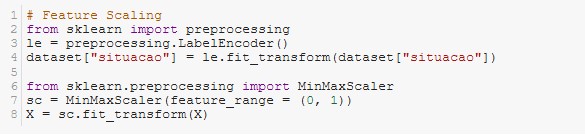
Esse código faz a importação dos dados da base de dados, no caso em .csv, denominada de lop\_submissao\_semana. O dataset.head mostra uma certa quantidade de dados contidos em cada linha da base.

**2º Seleção dos atributos:**



Esse código permite a escolha dos atributos (X) que são usados para comparação do parâmetro (Y), sendo este as coordenadas de certas partes pré-selecionadas para a base de dados, no caso será usada para determinar o número de submissões por semana data a situação do aluno (indicado pelos números de 0 a 4).

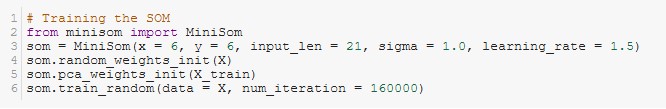
**3º** Escala de recursos**:**



Ele transformará seus dados de forma que sua distribuição tenha um valor médio 0 e desvio padrão de 1. Dada a distribuição dos dados, cada valor no conjunto de dados terá o valor médio da amostra subtraído e depois será dividido pelo desvio padrão de todo o conjunto de dados e possui um codigo que transforma a situação dos alunos da base de dados em numeros (como mostrado na etapa de cima).

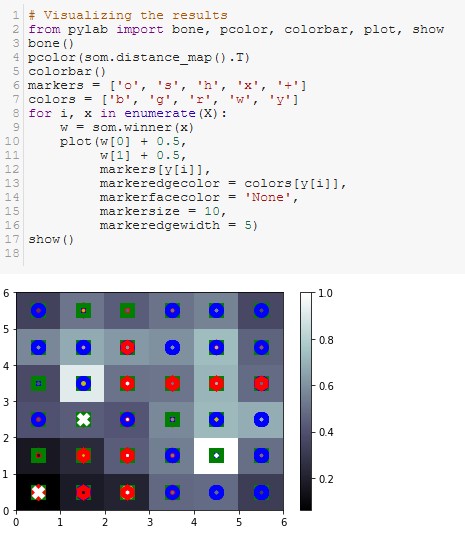
**06**

**4º Treinamento I:**

****

Esse código inicia a separação dos atributos para serem treinados e testados para a criação da matriz (mapa), determinando o tamanho da matriz, número da entrada de dados, a taxa de aprendizado e o número de interações que o treinamento irá ter.

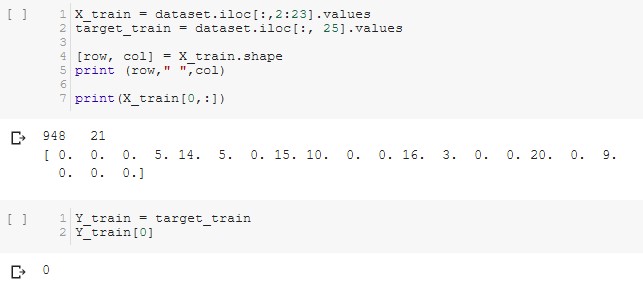
5**º Matriz I (Mapa):**

****

Esse código mostra a montagem e a matriz (mapa) da rede SOM, mostrando onde está posicionado os clusters e (respectivamente) as situações dos alunos de 0 a 4 como mostrado anteriormente usando cores (respectivamente azul, verde, vermelho, branco e amarelo).

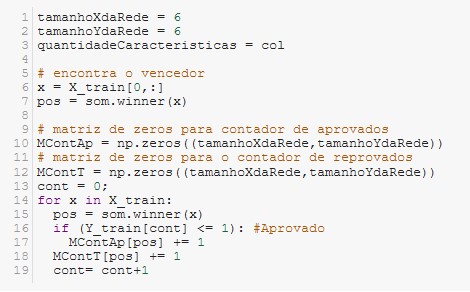
07

**6º Treinamento II:**

****

Treinamento da rede SOM para criação de gráficos e matrizes para representar a relação entre as submissões e a situação do aluno.

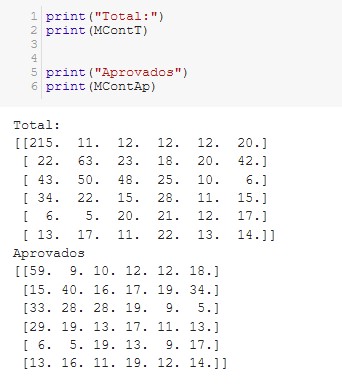
7**º Matriz II (parte 1):**

****

Mostra a montagem das duas matrizes, a matriz Total, que mostra o número total de submissões feitas pelos alunos em cada cluster no mapa da rede SOM, e a matriz Aprovados, que mostra o número de submissões feitas pelos alunos aprovados em cada cluster no mapa da rede SOM.

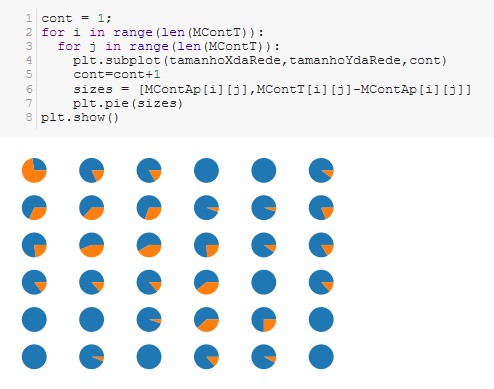
08

8**º Matriz II (parte 2):**

****

**Impressão das matrizes Total e Aprovados.**

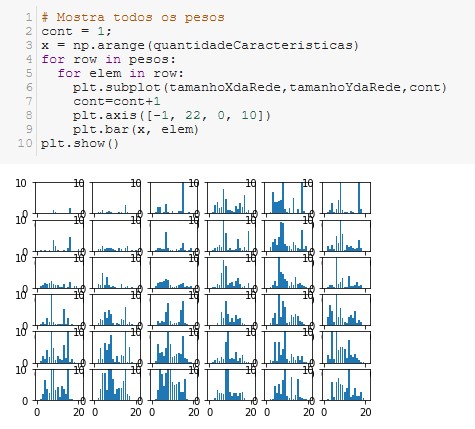
9**º Gráficos I (Pizza):**

****

**09**

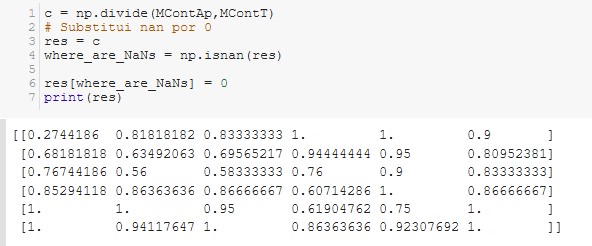
Os gráficos criados usando como base as matrizes Total e Aprovados, posicionado de acordo com os clusters, a qual os aprovados são representados pela cor azul e a cor laranja são os reprovados.

10**º Gráficos II (Histograma):**

****

Os gráficos criados usando como base as matrizes Total e Aprovados, posicionado de acordo com os clusters, mostrando o número de submissões por clusters.

11**º Resultados:**



A matriz mostra a porcentagem de aprovados em cada clusters (Aprovados/Total).

10

**4 Experimentos**

Os tipos de testes executados foram para, de acordo com os dados fornecidos, ter um gral de certeza de que o número de submissões feitas pelos alunos tinha relação com a situação dos alunos (aprovados e reprovados), utilizado a criação de um mapa de clusters para criar glomerado da situação do aluno, em seguida a criação de duas matrizes que utilizavam os clusters para determinar o número de submissões (matriz Total) e o número de submissões de alunos aprovados (matriz Aprovado), utilizados na criação de dois gráficos, um para determinar a porcentagem de aprovados por clusters (Pizza) e outro para o número de submissões por clusters (Histograma), e uma matriz que mostra a porcentagem de aprovação de cada cluster.

Parâmetro avaliado foi a situação dos alunos, se aprovado ou reprovado, o qual foi comparado com o numeros de submissões feitas pelos alunos durante as 21 semanas do semestre.

Os resultados obtidos foram que na maiorias dos casos os clusters que tinham cerca de 60% a 100% de aprovados tinham alto indice de submissões ao longo do semestre, principalmente no meio e no fim dele, enquanto nos outros clusters de reprovados acima de 40% tinham baixo indice de submissões ao longo do semestre, principalmente no meio dele.

Em conclusão, o numero alto de submissões, na maioria do casos simbolizam a maior probabilidade de que o aluno seja aprovado enquanto o numero baixo de submissões simbolizam a maior probabilidade de que o aluno seja reprovado.

11