Comparação entre os métodos kNN, k-Means e Mapa de Kohonen usados na classificação de peixes

Igor Beilner[1]

[1]Universidade Federal da Fronteira Sul (UFFS)
Curso de Ciência da Computação – Chapecó – SC – Brasil

igor.beilner@gmail.com

Resumo. *A classificação de dados de forma automática é uma preocupação de diversos segmentos, seja na inspeção de qualidade de peças na indústria, ou na clusterização de uma base de dados comercial que busca agrupar seus clientes de acordo com seus interesses. Este trabalho tem o objetivo de comparar o desempenho de três técnicas de aprendizado de máquina aplicados à classificação de duas espécies de peixes.*

1. Introdução

O aprendizado de máquina tenta fazer com que um sistema inteligente aprenda a realizar determinada tarefa a partir da experiência [3], este aprendizado pode ocorrer de duas maneiras: supervisionado e não-supervisionado. Em um sistema supervisionado, o algoritmo recebe como parâmetro um conjunto de treinamento que servirá como base para a classificação da entrada. Já em um sistema não-supervisionado, o aprendizado acontece sem ter nenhuma intervenção externa, o que pode ser útil quando os dados não são conhecidas previamente.

Neste trabalho foram comparados o desempenho de dois métodos estatísticos e um método de redes neurais artificiais (RNA) na classificação de dois tipos de peixes: Robalo e Salmão. Em qualquer sistema de classificação, é necessário que se tenham características que distinguem objetos de classes diferentes, mas que permitam identificar as variações de objetos de um mesmo grupo. Nesta aplicação, as características de cada espécie, fornecidas por uma base de dados, foram a luminosidade refletida pelos peixes e suas respectivas larguras, a distribuição dos peixes é apresentada na figura 1. Os pontos na cor azul observados na imagem representam a espécie salmão e os pontos vermelhos, os robalos, em um total de 131 peixes, 74 são salmões e 57 robalos.

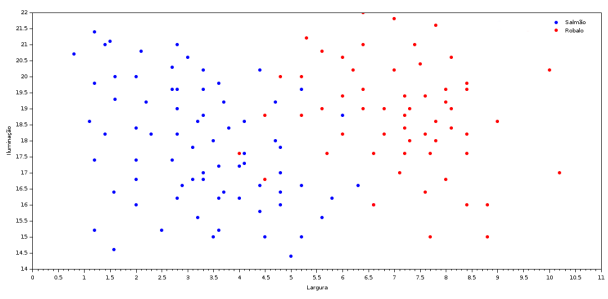

Figura 1. Distribuição dos peixes

2. Materiais e métodos

Foram usados três métodos para a classificação dos peixes, um supervisionado, o k-Nearest Neighbor, e dois não supervisionados, K-Means e Kohonen. As subseções a seguir apresentam cada um dos métodos utilizados.

2.1. Método k-Nearest Neighbor

Como mencionado anteriormente, o k-Nearest Neighbor (kNN) é um método de aprendizado supervisionado, pois, na etapa de treinamento é necessário fornecer as características que representem cada conjunto de interesse.

O kNN efetua a classificação de acordo com a semelhança do dado de entrada com os conjuntos de treinamento [5], esta semelhança pode ser descrita como a menor de distância do ponto de entrada à cada conjunto de treinamento.

Fazer uso de um grande conjunto de treinamento, não necessariamente, implica em uma melhor classificação dos dados, pois, pode ocorrer de os pontos deste conjunto não representarem a classe de maneira satisfatória [4]. Portando, a escolha deste conjunto deve ser feita de forma criteriosa, buscando encontrar o menor conjunto e que melhor represente a classe, isso diminui o custo computacional e melhor classifica os dados.

Nesta aplicação foi usado um conjunto de treinamento de doze pontos, e, para verificar a proximidade do ponto a ser classificado com o de treinamento, foi usado o cálculo da distância euclidiana ponderada, em que a distância entre um ponto p e outro ponto q é dada pela equação 1. A ponderação foi feita de modo que a luminosidade tenha 30% e a largura 70% de relevância.

$$D(p,q) = \sqrt{w_1(p_1 - q_1)^2 + w_2(p_2 - q_2)^2 + ... + w_n(p_n - q_n)^2} \tag{1}$$

A equação 1 mostra o cálculo da distância euclidiana entre dois pontos p e q, onde p_i e q_i são atributos correspondentes de cada ponto e w_i é a ponderação de cada atributo.

2.2. Método k-Means

Diferente do kNN, o k-Means é um método de classificação não-supervisionado, isso faz com que ele não dependa de um conjunto de treinamento definido a priori pelo usuário, esta característica faz com que o k-Means seja mais robusto, pois o deixa independente às variações dos dados, já que em um método supervisionado, em algum momento o conjunto de treinamento definido a priori pode deixar de representar sua respectiva classe.

O k-Means fica dependente apenas de um parâmetro k de entrada, que representa a quantidade de *clusters* responsáveis por agrupar as classes [6]. Ao ser inicializado, o algoritmo k-Means escolhe, de maneira arbitrária, k sementes para serem as centróides iniciais, a partir disso, os dados são agrupados aos *clusters* que possuem maior semelhança, nesta implementação, a semelhança é definida pela menor distância euclidiana ponderada que o dado a ser classificado possui em relação às centróides. Depois de ter classificado todos os dados, é calculada uma nova centróide, como mostra a equação 2.

$$C(\chi, n) = \frac{\sum_{i=1}^{n} (P_i \in \chi)}{n} \tag{2}$$

A equação 2 mostra o cálculo da nova centróide, onde χ é o conjunto de pontos da classe, *n* é o número de pontos da classe e P_i é um ponto do conjunto. O cálculo da nova centróide é executado iterativamente até que um critério de parada ϵ seja satisfeito, nesta aplicação o critério de parada utilizado foi o monitoramento da transição dos dados entre os *clusters*, que, quando os dados não mudam mais de um *cluster* para outro, ou seja, se estabilizam, é dito que os dados estão classificados.

2.3. Mapa de Kohonen

Os mapas de Kohonen são um método de redes neurais artificiais (RNA) [2] não-supervisionado inspirado na neurobiologia, em que neurônios competem entre si, sendo o neurônio vencedor o responsável por efetuar o reconhecimento da entrada [1].

A figura 2 ilustra a estrutura de um mapa de Kohonen, o mapa é composto por uma camada de entrada e outra de saída. A camada de entrada é composta por um vetor de características que representam o dado quantitativamente, cada atributo deste vetor está conectado a todos os neurônios da rede, estas ligações são chamadas de sinapses, cada sinapse da rede possui um peso responsável por indicar a maior similaridade entre o dado de entrada e o neurônio.

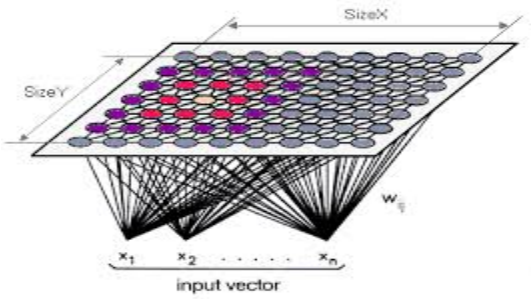

Figura 2. Mapa de Kohonen

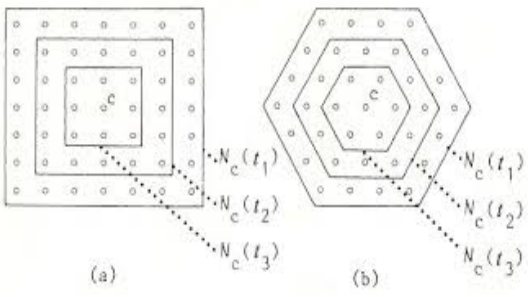

Figura 3. Vizinhança do neurônio vencedor

O algoritmo de uma rede de Kohonen é composto por duas etapas: aprendizado e aplicação. Na etapa de aprendizado existem duas fases: competitiva e colaborativa. A fase competitiva é a fase onde os atributos de entrada verificam sua similaridade com os neurônios da rede, nesta aplicação ao grau de similaridade entre neurônios e atributos é dado pela distância euclidiana ponderada, em que a menor distância indica maior similaridade. Juntamente com a fase competitiva, ocorre a colaborativa, que consiste em incrementar o peso das sinapses do neurônio excitado na fase competitiva e de seus vizinhos de acordo com uma taxa de aprendizado, taxa esta, que decresce com o passar do tempo, assim como o número de vizinhos também decresce. Nesta aplicação a taxa de aprendizado é dada pela equação 3 e o número de vizinhos é dado pela equação 4, onde n é o número de iterações da etapa de aprendizado, T é o máximo de iterações, η_0 é a taxa de aprendizado inicial e d_0 é o número inicial de vizinhos.

Nesta aplicação, a taxa inicial de aprendizado foi definida como 0.3 e o número inicial de vizinhos definido como a metade do número de linhas da matriz. A matriz que representa a rede, possui dimensões quadradas de 4x4, 8x8 e 16x16.

$$\eta(n) = \eta_0 \left(1 - \frac{n}{T}\right) \tag{3}$$

$$d(n) = \left\lceil d_0 \left(1 - \frac{n}{T}\right) \right\rceil \tag{4}$$

Quanto à vizinhança do neurônio vencedor, existem várias topologias, na figura 3 são apresentadas duas delas: a quadrada e a hexagonal. O tamanho da vizinhança tem o propósito de incrementar os pesos dos vizinhos do neurônio excitado na fase colaborativa. Nesta aplicação foi usada a topologia quadrada.

Após a etapa de aprendizado, é executada a etapa da aplicação, esta por sua vez, executa apenas a fase competitiva, que tem como objetivo realizar o reconhecimento a partir dos padrões definidos na etapa de aprendizado, através da excitação de neurônios similares, esta etapa efetua a classificação dos dados de entrada.

3. Resultados e discussão

Foram comparados os resultados obtidos pelos três classificadores. O desempenho de cada método é apresentado na tabela 1, são apresentados o número de salmões e robalos reconhecidos e também o erro associado a cada método, a linha *Erro salmões* corresponde aos salmões reconhecidos como robalos e, a linha *Erro Robalos* se referre aos robalos reconhecidos como salmões pelos respectivos métodos.

Os pontos escolhidos para o conjunto de treinamento de método kNN foram os que se encontram visualmente no centro de cada classe, de acordo com a figura 1. Para execução do k-Means, as sementes escolhidas foram os dois primeiros pontos da distribuição, que neste caso, pertencem a classe salmão, isso faz com que o algoritmo faça mais iterações para convergir.

Nos Mapas de Kohonen, os pesos das sinapses foram inicializados com uma função *rand()* com valores entre 0 e 1. A etapa de aprendizado fez uso de trinta pontos de cada espécie de peixe.

	kNN	k-Means	Kohonen 4x4	Kohonen 8x8	Kohonen 16x16
Salmões	72	72	94	80	76
Robalos	59	59	37	51	55
Erro Salmões	5	5	0	0	0
Erro Robalos	4	4	20	6	2

Tabela 1. Número de peixes reconhecidos em cada espécie e erro associado de acordo com os valores reais

Observando a tabela 1, verifica-se que o desempenho dos métodos kNN e k-Means aprensentam resultados idênticos, isso é reflexo da escolha do conjunto de treinamento do kNN, que neste caso, os pontos escolhidos estão a uma pequena distância da centróide de cada conjunto, o que implica na classificação identica, pois, o k-Means faz uso das centróides para fazer a clusterizacão.

Nas figuras 4 a 8 são apresentadas as classificações realizadas por cada um dos métodos propostos.

Ao analisar as figuras 4 e 5, que mostram as classificação feita pelos métodos kNN e k-Means, observa-se que o grande problema é construir uma fronteira de decisão que consiga identificar peixes que possuem características confusas em relação a suas classes.

Nas figuras 6, 7 e 8 são apresentados os resultados da classificação realizada pelos mapas de Kohonen em suas respectivas dimensões.

Observando as imagens é possível visualizar os efeitos causados de acordo com as dimensões do mapa, nota-se por exemplo, que quanto maior a dimensão, melhor é a classificação e torna-se possível classificar um peixe que possui características similares as de outra espécie.

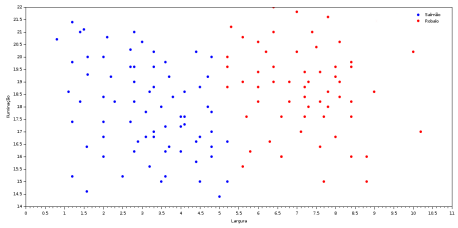

Figura 4. Classificacao com método kNN

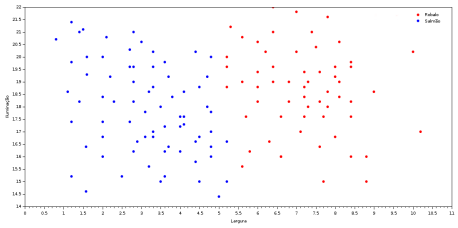

Figura 5. Classificacao com método k-Means

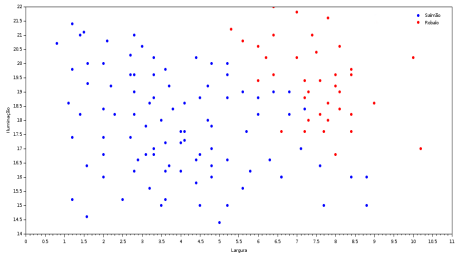

Figura 6. Classificacao com Kohonen 4x4

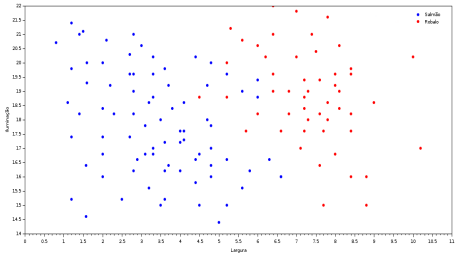

Figura 7. Classificacao com Kohonen 8x8

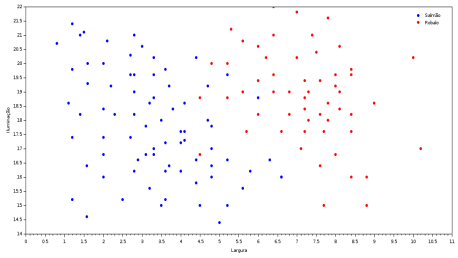

Figura 8. Classificacao com Kohonen 16x16

4. Conclusão

Analisando o comportamento de cada um dos métodos aprensentado na seção anterior, pode-se concluir que a escolha por um método para classificação de dados deve ser feita observando a necessidade da aplicação, pois, se a aplicação permite uma margem de erro na classificação, o método kNN ou k-Means pode ser indicado pelo fato de possuírem

implementação mais trivial. Já em sistemas mais críticos, onde não há tolerância de falhas, os mapas de Kohonen se comportam de maneira mais aceitável.

Referências

[1] Wonder Alexandre Luz Alves, SA de ARAÚJO, and André Felipe H Librantz. Reconhecimento de padrões de texturas em imagens digitais usando uma rede neural artificial híbrida. *Revistas Científicas de América Latina y el Caribe, España y Portugal*, 4:325–332, 2006.

[2] Marcio Brumatti. Redes neurais artificiais. *Vitória, Espírito Santo, ca*, 2005.

[3] Ben Coppin. *Inteligência Artificial*, volume 1. LTC, 2010.

[4] Carlos Andres Ferrero. *Algoritmo kNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia*. PhD thesis, Universidade de São Paulo, 2009.

[5] Leonardo Cavalheiro Langie and VLS LIMA. Classificação hierárquica de documentos textuais digitais usando o algoritmo knn. In *1o Workshop em Tecnologia da Informação e da Linguagem Humana*, volume 1, pages 1–10, 2003.

[6] Ricardo Linden. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, (4):18–36, 2009.