5
KNN (K – Nearest Neighbors)

5.1
Introdução

KNN é um classificador onde o aprendizado é baseado na analogia. O conjunto de treinamento é formado por vetores n-dimensionais e cada elemento deste conjunto representa um ponto no espaço n-dimensional.

5.2
Metodologia

Para determinar a classe de um elemento que não pertença ao conjunto de treinamento, o classificador KNN procura K elementos do conjunto de treinamento que estejam mais próximos deste elemento desconhecido, ou seja, que tenham a menor distância.

Estes K elementos são chamados de K-vizinhos mais próximos. Verifica-se quais são as classes desses K vizinhos e a classe mais freqüente será atribuída à classe do elemento desconhecido.

Abaixo tem-se as métricas mais comuns no cálculo de distância entre dois pontos, sendo que a mais utilizada, é a distância Euclidiana.

Seja $X= (x_1, x_2, ... , x_n)$ e $Y= (y_1, y_2, ... , y_n)$ dois pontos do $\Re^n$.

- A distância Euclidiana entre X e Y é dada por

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} .$$

- A distância Manhattan entre X e Y é dada por

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n| .$$

- A distância Minkowski entre X e Y é dada por

$$d(x,y) = \left(|x_1 - y_1|^q + |x_2 - y_2|^q + ... + |x_n - y_n|^q \right)^{1/q} , \text{ onde } q \in N .$$

Esta distância é a generalização das duas distâncias anteriores. Quando q = 1, esta distância representa a distância de Manhattan e quando q = 2, a distância Euclidiana.

Se cada variável possuir um peso relativo a sua importância, a distância Euclidiana ponderada pode ser representada como

$$d(x,y) = \sqrt{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + ... + w_n(x_n - y_n)^2}\ .$$

Pesos também podem ser aplicados às distâncias Manhattan e Minkowski.

KNN é um classificador que possui apenas um parâmetro livre (o número de K-vizinhos) que é controlado pelo usuário com o objetivo de obter uma melhor classificação.

Este processo de classificação pode ser computacionalmente exaustivo se considerado um conjunto com muitos dados. Para determinadas aplicações, no entanto, o processo é bem aceitável.

Na figura abaixo, tem-se um exemplo de classificação KNN com dois atributos, três classes e dois pontos desconhecidos 1 e 2. Deseja-se classificar estes dois pontos através dos 7 vizinhos mais próximos.

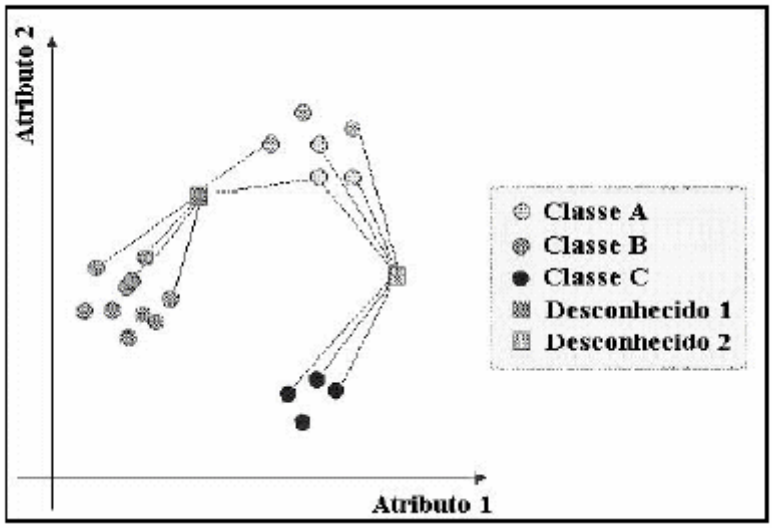

Figura 18 - Classificação pelo método KNN
Fonte: Gnecco et al., 2005

Analisando a classe predominante dos 7 vizinhos mais próximos, o ponto desconhecido 1 será classificado como um ponto pertencente a classe B e o ponto desconhecido 2 como um ponto pertencente a classe A.

Como já foi dito acima, este processo de classificação pode ser computacionalmente exaustivo; por este motivo, há uma variação mais rápida deste algoritmo.

Esse outro processo seleciona pontos que estão dentro de uma hiper-esfera de raio R (decidido pelo usuário) e a classe predominante dentro desta hiper-esfera será a classe do ponto desconhecido. A desvantagem deste processo é que pode existir hiper-esfera sem qualquer ponto.

A figura abaixo mostra o exemplo da figura 18 com o algoritmo modificado.

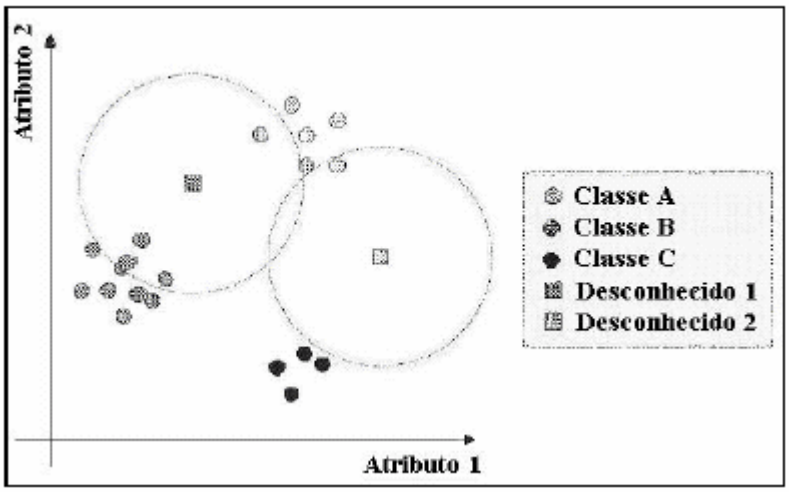

Figura 19 - Classificação pelo método KNN utilizando hiper-esferas de raio R
Fonte: Gnecco et al., 2005

Os vizinhos dos pontos desconhecidos 1 e 2 são os pontos pertencentes ao círculo centrado no ponto desconhecido 1 e 2, respectivamente.

O ponto desconhecido 1 será classificado como um ponto pertencente a classe B pois existem 5 pontos dentro ou parcialmente dentro do círculo centrado no ponto desconhecido 1. O ponto desconhecido 2 será classificado como um ponto pertencente a classe A pois existe somente um ponto da classe A dentro do círculo centrado no ponto desconhecido 2.