KNN em Java

Thiago Alexandre Nakao França

Universidade Tecnológica Federal do Paraná – UTFPR
COCIC – Coordenação do Curso de Bacharelado em Ciência da Computação
Campo Mourão, Paraná, Brasil
nakaosensei@gmail.com

Resumo

O objetivo deste trabalho foi desenvolver o classificador KNN em Java, avaliar o seu desempenho para dois conjuntos de treino/teste distintos, serão avaliados variados valores de k.

1. Introdução

KNN é um classificador de simples implementação, embora seja antigo ainda é muito usado para fins de pesquisa/comparação com outros classificadores.

Neste trabalho o KNN foi implementado em Java, serão avaliados dois conjuntos de treino/teste, no primeiro par existem 10 classes no conjunto de treino, onde cada classe tem 132 caracteristicas, e no segundo par existem 25 classes no conjunto de treino, onde cada classe tem 2048 caracteristicas, os conjuntos de testes podem ter mais classes mas não podem ter mais caracteristicas.

Foi usada a distância euclidiana para calcular a distancia entre as instancias, além disso, a normalização foi por MinMax.

2. Experimentos do primeiro conjunto

No primeiro conjunto, existem 10 classes e 132 caracteristicas, serão exibidas as primeiras duas matrizes de confusão, as demais podem ser obtidas executando o programa, seguem os resultados:

k = 1, Matriz de confusão:
Accuracy:92.7

```
94 0 0 1 0 1 1 0 1 2
0 90 3 0 1 2 2 0 2 0
0 0 89 4 0 0 0 5 2 0
0 0 7 86 0 3 1 0 1 2
0 1 0 0 96 0 2 0 0 1
0 2 0 2 0 95 0 0 0 1
0 0 0 0 0 0 99 0 1 0
0 0 0 0 1 0 0 98 0 1
0 1 0 0 2 2 3 0 86 6
0 0 0 0 3 0 0 1 2 94
```

k = 3
Accuracy: 90.3
Matriz de confusão:

```
96 0 1 1 0 0 0 1 1 0
0 88 3 0 3 2 0 0 3 1
0 0 90 2 2 0 2 2 2 0
1 0 5 90 0 0 1 2 0 1
0 1 0 0 95 0 1 2 0 1
0 5 0 5 0 80 2 0 4 4
0 1 1 0 0 1 95 0 1 1
0 0 0 0 0 0 0 96 0 4
1 2 0 1 2 1 4 1 79 9
0 0 0 0 0 1 0 1 4 94
```

k = 5
Accuracy: 85.3
Matriz de confusão:

k = 7
Accuracy: 79.9

k = 9
Accuracy: 74.3

k = 11
Accuracy: 68.10000000000001

k = 13
Accuracy: 65.60000000000001

k = 15
Accuracy: 61.1

k = 17
Accuracy: 57.8

k = 19
Accuracy: 54.400000000000006

3. Experimentos do segundo conjunto

No segundo conjunto existem 25 classes no conjunto de treino, onde cada classe tem 2048 caracteristicas, no entanto no conjunto de testes existem 115 classes distintas.

No segundo experimento, a matriz não será exibida no relatório, por ser muito grande.

K = 1
Accuracy: 11.304347826086957

K = 3
Accuracy:8.212560386473431

K = 5
Accuracy:7.536231884057972

K = 7
Accuracy:5.990338164251208

K = 9
Accuracy:4.9275362318840585

K = 11
Accuracy:4.154589371980676

K = 13
Accuracy:3.6714975845410627

K = 15
Accuracy:2.8019323671497585

K = 17
Accuracy:2.0289855072463765

K = 19
Accuracy:1.5458937198067633

4. Conclusões

Após a realização dos experimentos, foi notável que o uso de valores elevados de K levou a acuracia para valores cada vez menores, isso se deve ao fato de classes "fora de questão" passarem a ser consideradas pelo voto majoritário.