Classificador KNN e DMC

Savio Lopes Rabelo

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Programa de Pós-Graduação em Ciência da Computação
Campus Fortaleza – CE – Brasil

saviorabelo.ti@gmail.com

Resumo. *Este relatório descreve a implementação do KNN e DMC aplicada à classificação de padrões. A metodologia utilizada para a implementação é constituída por duas fases: treinamento e teste, com cada conjunto sendo composto por 80% e 20% das bases de dados, respectivamente, utilizando também uma busca em grade com validação cruzada k-fold. Foram usadas quatro bases de dados disponíveis online no repositório UCI Machine Learning. Os resultados são bastante satisfatórios, chegando em taxas de acerto em 100% em algumas bases.*

1. Introdução

KNN (do inglês, *K-Nearest Neighbors*) 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. 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 frequente será atribuída à classe do elemento desconhecido.

No classificador DMC (Distância Média dos Centróides), para cada classe é assumido um centro de massa (também conhecido como centróide). Um objeto x pertence a uma determinada classe y, quando a distância entre x e o centróide da classe y, for menor que todas as distâncias entre y e as outras classes do espaço de características. O primeiro passo do processo de classificação por distância mínima e o cálculo dos vetores médios (centróides) que representam cada classe por padrões. Para o cálculo da distância, diversas métricas podem ser utilizadas, como por exemplo, a distância Euclidiana.

Abaixo tem-se as métricas mais comuns no cálculo de distância entre dois pontos, sendo que a mais utilizada e a que foi utilizada nesse trabalho foi a distância Euclidiana. Seja $X = (x_1, x_2, \ldots, x_n)$ e $Y = (y_1, y_2, \ldots, y_n)$ dois pontos do $\mathbb{R}^n$.

A distância Euclidiana entre X e Y é dada por:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}. \quad (1)$$

A distância Manhattan entre X e Y é dada por:

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|. \quad (2)$$

A distância Minkowski entre X e Y é dada por:

$$d(x, y) = (|x_1 - y_1|^q + |x_2 - y_2|^q + \cdots + |x_n - y_n|^q)^{1/q}, \tag{3}$$

onde $q \in \mathbb{R}$.

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.

2. Metodologia

No primeiro momento foi realizada a separação do conjunto de dados em dois subconjuntos: treinamento e teste. Os valores utilizados para os conjuntos equivalem a 80% do conjunto original para a fase de treinamento e 20% do conjunto original para a fase de teste. Logo depois, os dados foram normalizados para eliminação de redundâncias indesejadas e também foram embaralhados. Por ser um problema com mais de 2 classes, foi exigida uma codificação diferente, Um-versus-Todos (do inglês *One-v-All*, OvA).

Além disso, também foi utilizada a busca em grade com validação cruzada *k-fold*. A busca em grade é uma busca com o objetivo de encontrar os melhores parâmetros. Já o método de validação cruzada *k-fold* consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros e calcula-se a acurácia do modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste. A Figura abaixo mostra o esquema realizado pelo *k-fold*. Ao final das k iterações calcula-se a acurácia sobre os erros encontrados, obtendo assim uma medida mais confiável sobre a capacidade do modelo de representar o processo gerador dos dados.

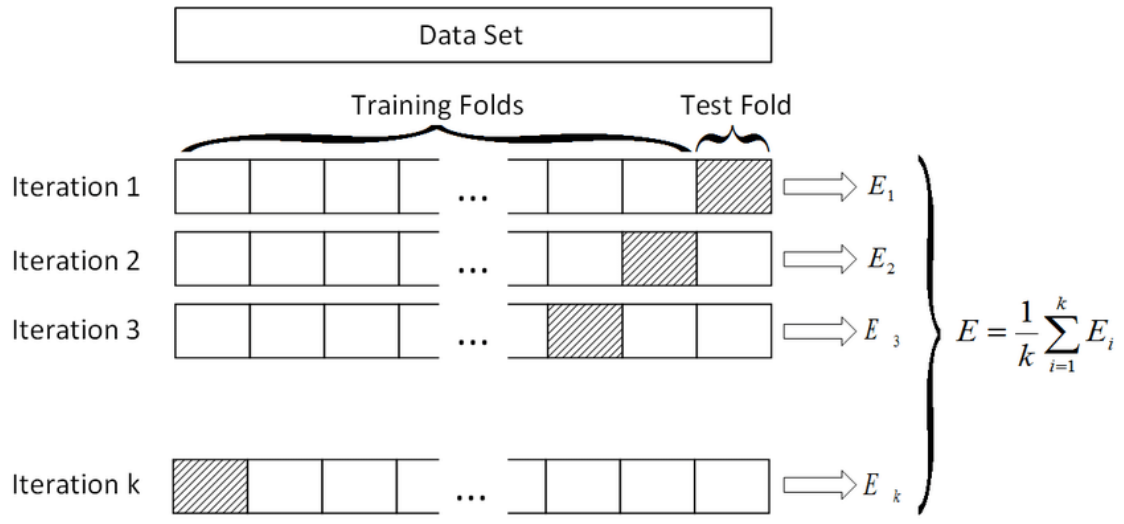


Figura 1. Método k-fold

Para a avaliação dos resultados alcançados na classificação, foram utilizados as seguintes métricas: a precisão ou valor preditivo positivo, taxa de sensibilidade ou taxa positiva verdadeira, especificidade ou taxa real negativa e acurácia. As Equações são apresentadas a seguir:

$$Precisao = \frac{VP}{VP + FP}, \tag{4}$$

$$Sensibilidade = \frac{VP}{VP + VN}, \tag{5}$$

$$Especificidade = \frac{VN}{N} = \frac{VN}{FP + VN}, \tag{6}$$

$$Acuracia = \frac{VP + VN}{P + N}, \tag{7}$$

onde P e N é o número de padrões de cada classe. VP é o verdadeiro positivo. VN é o verdadeiro negativo. FP é o falso positivo e FN é o falso negativo.

3. Conjuntos de Dados

. Para análise comparativas neste estudo, foram usados quatro conjuntos de dados: *Iris Flower Data Set*, *Vertebral Column Data Set*, *Dermatology Data Set* e *Breast Cancer Wisconsin Data Set*; todos disponíveis online no repositório *UCI Machine Learning* [Lichman 2013].

O banco de dados da Íris[1] é o conjunto mais conhecido que se encontra na literatura de reconhecimento de padrões. O conjunto de dados contém 3 classes de 50 instâncias cada, onde cada classe se refere a um tipo de planta de íris. Uma classe é linearmente separável das outras 2 classes.

Informações dos atributos:

1. Tamanho da sépala em cm
2. Largura da sépala em cm
3. Tamanho da pétala em cm
4. Largura da pétala em cm
5. Classe:
   (a) Iris Setosa
   (b) Iris Versicolour
   (c) Iris Virginica

Já o conjunto de dados da Coluna Vertebral[2] contém seis valores para características biomecânicas usadas para classificar pacientes ortopedistas em 3 classes (normal, hérnia de disco ou espondilolistese) ou 2 classes (normal ou anormal). Foi utilizado nessa prática o conjunto com três classes.

Informações dos atributos:

1. Incidência pélvica
2. Inclinação pélvica
3. Ângulo de lordose lombar

[1]Disponível em `https://archive.ics.uci.edu/ml/datasets/iris`
[2]Disponível em `http://archive.ics.uci.edu/ml/datasets/vertebral+column`

4. Inclinação sacra
5. Raio pélvico
6. Grau de espondilolistese
7. Classe:
    (a) Hérnia de Disco (DH)
    (b) Espondilolistese (SL)
    (c) Normal (NO)
    (d) Anormal (AB)

O banco de dados de Dermatologia[3] é constituído de 34 atributos. Esse banco é parte de um estudo que aponta o tipo de Eryhemato-Squamous Disease, uma doença de pele.

Informações dos atributos (valores de 0 a 3, exceto quando indicado):

1. Eritema
2. Escala
3. Fronteiras Definidas
4. Coceira
5. Fenômeno Koebner
6. Pápulas Poligonais
7. Pápulas Foliculares
8. Envolvimento da Mucosa Oral
9. Envolvimento no Joelho e no Cotovelo
10. Envolvimento do Couro Cabeludo
11. Histórico Familiar (0 ou 1)
12. Atributos Histopatológicos

    ⋮

33. Atributos Histopatológicos
34. Idade (Classe de 1 a 6)

E o banco de dados de Câncer de Mama[4] é constituído de 10 atributos. Informações dos atributos:

1. Número do código de amostra (número de identificação)
2. Clump Espessura (1 - 10)
3. Uniformidade do tamanho da célula (1 - 10)
4. Uniformidade da forma da Célula (1 - 10)
5. Adesão Marginal (1 a 10)
6. Tamanho Único de Células Epiteliais (1 - 10)
7. Núcleos Nus (1 - 10)
8. Cromatina Branda (1 a 10)
9. Nucleoli Normal (1 - 10)
10. Mitoses (1 - 10)
11. Classe (2 para benigno, 4 para maligno)

[3]Disponível em `http://archive.ics.uci.edu/ml/datasets/dermatology`

[4]Disponível em `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\%28Original\%29`

4. Simulações Computacionais

Para realizar os experimentos, foi utilizado um computador com a seguinte configuração: processador Intel(R) Core(TM) i7-6500U a 2.5 GHz com 8 GB de RAM e executando Windowns 10. Além disso, foi utilizado a linguagem de programação MATLAB. Todos os testes foram feitos com 50 realizações em cada base.

No classificador KNN foi utilizado o busca em grade com validação cruzada *k-fold* para encontrar o melhor k, o intervalo variou de 1 até 31, incrementando de 2 em 2. A Tabela 1 mostra os resultados do KNN em todas as bases de dados, levando em consideração as métricas já mencionadas.

	Bases de Dados					
Métricas	Íris	Coluna (3C)	Coluna (2C)	Dermatologia	Câncer	Artificial
Acurácia	96,40	76,84	100,00	97,03	100,00	100,00
Taxa Mínima	86,67	67,29	100,00	94,59	100,00	100,00
Taxa Máxima	100,00	87,10	100,00	100,00	100,00	100,00
Desvio Padrão	03,42	04,34	00,00	01,87	00,00	00,00
Sensibilidade	96,57	73,38	100,00	96,63	100,00	100,00
Especificidade	98,19	84,41	100,00	99,43	100,00	100,00
Precisão	96,36	74,04	100,00	96,82	100,00	100,00
Tempo (s)	40,89	81,21	89,77	115,36	262,39	20,92
Intervalo de k	[3-11]	[13-27]	[1]	[3-13]	[1]	[1]

Tabela 1. Resultados do KNN.

A Tabela 2 mostra os resultados do DMC em todas as bases de dados, levando em consideração as métricas já mencionadas.

	Bases de Dados					
Métricas	Íris	Coluna (3C)	Coluna (2C)	Dermatologia	Câncer	Artificial
Acurácia	94,00	75,19	100,00	96,49	99,03	100,00
Taxa Mínima	83,33	66,13	100,00	90,54	97,86	100,00
Taxa Máxima	100,00	88,71	100,00	100,00	100,00	100,00
Desvio Padrão	04,47	05,27	00,00	02,02	00,64	00,00
Sensibilidade	94,23	72,50	100,00	95,96	99,62	100,00
Especificidade	97,03	87,00	100,00	99,30	97,90	100,00
Precisão	94,19	74,07	100,00	96,34	98,92	100,00
Tempo (s)	00,24	00,29	00,20	00,36	00,21	00,15

Tabela 2. Resultados do DMC.

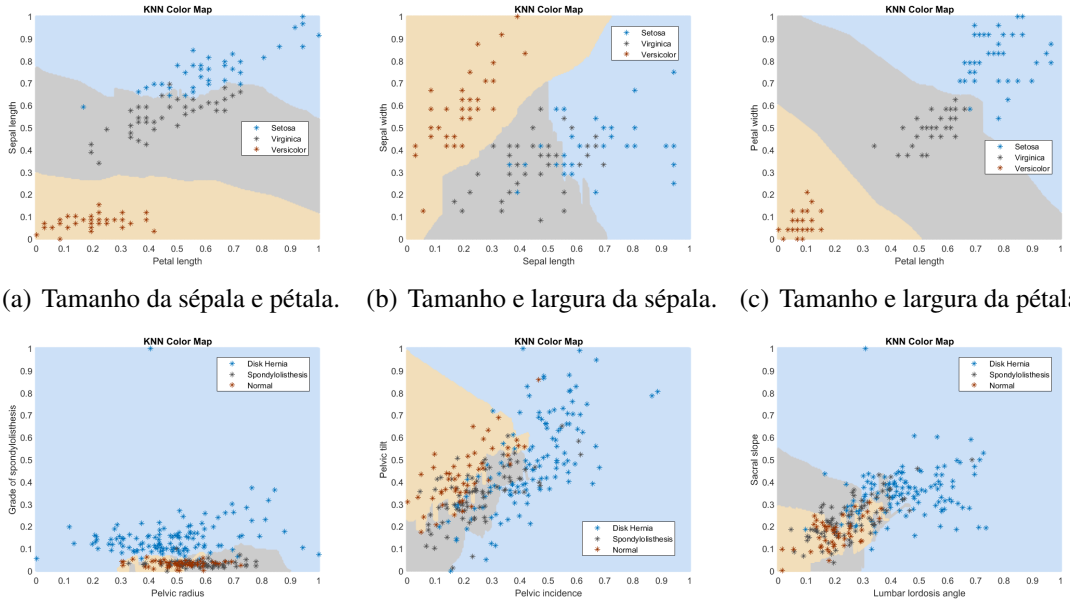
A Figura 2 apresenta uma matriz de confusão de cada base de dados. Essa matriz é a matriz que ficou mais perto da acurácia.

(a) Íris. (b) Coluna (2C). (c) Coluna (3C).

(d) Dermatologia. (e) Câncer. (f) Artificial (2C).

Figura 2. Matriz de Confusão para o KNN.

A Figura 3 apresenta a superfície de decisão para a base de dados da Íris e da Coluna com o classificador KNN.

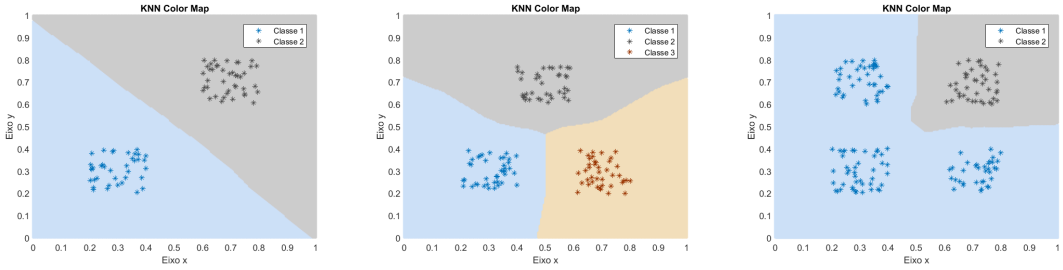


(a) Tamanho da sépala e pétala. (b) Tamanho e largura da sépala. (c) Tamanho e largura da pétala.

(d) Raio pélvico e Grau de espondilolistese. (e) Incidência e inclinação pélvica. (f) Ângulo de lordose lombar e Inclinação sacral.

Figura 3. Superfície de decisão para Íris e Coluna do classificador KNN.

Na Figura 4 é apresentada a superfície de decisão para três bases artificiais com o classificador KNN.



(a) Base artificial com 2 classes. (b) Base artificial com 3 classes. (c) Base AND.

Figura 4. Superfície de decisão do classificador KNN.

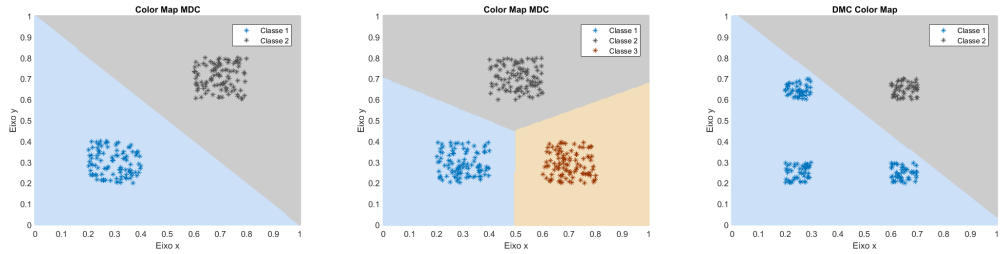
A Figura 5 apresenta uma matriz de confusão de cada base de dados. Essa matriz é a matriz que ficou mais perto da acurácia.

(a) Íris. (b) Coluna (2C). (c) Coluna (3C).

(d) Dermatologia. (e) Câncer. (f) Artificial (2C).

Figura 5. Matriz de Confusão para o DMC.

Na Figura 6 é apresentada o superfície de decisão para três bases com o classificador DMC.



(a) Base artificial com 2 classes. (b) Base artificial com 3 classes. (c) Base AND.

Figura 6. Superfície de decisão do classificador DMC.

5. Resultados

Analisando os experimentos, pode-se visualizar que em geral, os dois classificadores se sai muito bem tanto na classificação de padrões, chegando a taxas de acerto de 100% quase em sua totalidade em algumas bases de dados. A base Iris tem as classes bastante separadas, tornando fácil encontrar retas que separe-as. Já em bases como a Coluna Vertebral e a Dermatologia, onde dados de diferente classes sobrepõem-se, pode-se ver que o desempenho da taxa de acerto decai. Na base do Câncer foi obtido ótimo desempenho. Já com a base de dados gerada artificialmente, o resultado já era esperado obter 100% de acerto, visto que as classes são separáveis uma das outras.

Analisando o tempo computacional, o classificador KNN demorou mais em relação ao DMC, isso pode se dá porque foi utilizada a busca em grade para obter o melhor k, e também dependendo do valor de k, o custo poderá ser ainda maior.

Referências

Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml`. Acesso em março de 2019.