**Resumo Knn**

**Vertebral Column Data Set**

Conteúdo

[1. Descrição da Técnica.........................................................................................................................................](#_Toc477629333).3

[2. Descrição do Código............................................................................................................................................](#_Toc477629338)4

[3.](#_Toc477629340)[Experimentos........................................................................................................................................................5](#_Toc477629340)

[[3.1.](#_Toc477629340)](#_Toc477629341) [Descrição da base de dados – Vertebral Column Data Set ..........................................5](#_Toc477629340)

[3.2.](#_Toc477629341) Resultados do Experimento...........................................................................................................5

[3.3.](#_Toc477629341) Análise da Variação do K...............................................................................................................7

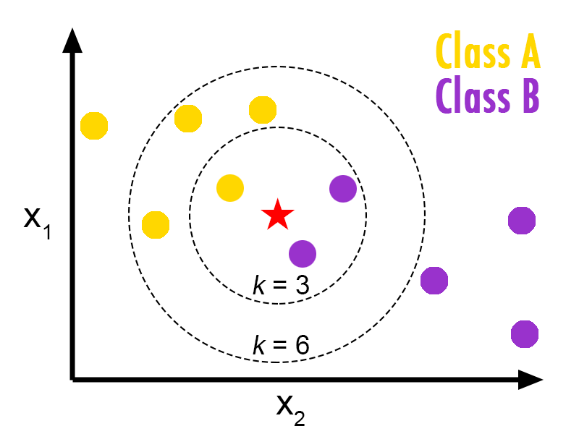
[3.4.](#_Toc477629341) Análise das métricas globais.........................................................................................................8

[4.Conclusão...............................................................................................................................................................9](#_Toc477629362)

[5. Referências..........................................................................................................................................................10](#_Toc477629367)

1. **Descrição da Técnica**

O **Knn** (*K-Nearest Neighbour*) é um algoritmo de aprendizagem de máquina muito usado para classificação. É considerado um dos mais simples algoritmos de aprendizagens de máquinas. A idéia do Knn é classificar uma amostra *x* por analisar seusk vizinhos mais próximos, sendo que estes são obtidos a partir do cálculo da distância de todos os vizinhos em relação a *x* (existem três métodos de cálculo para a distância dos vizinhos). Os k vizinhos mais próximos a x fornece cada um suas classes, logo x será classificada com base na classe mais frequente entre os seus k vizinhos mais próximos. O algoritmo Knn pode fornecer uma maior precisão de classificação em relação a algoritmos de classificação que usa funções lineares.

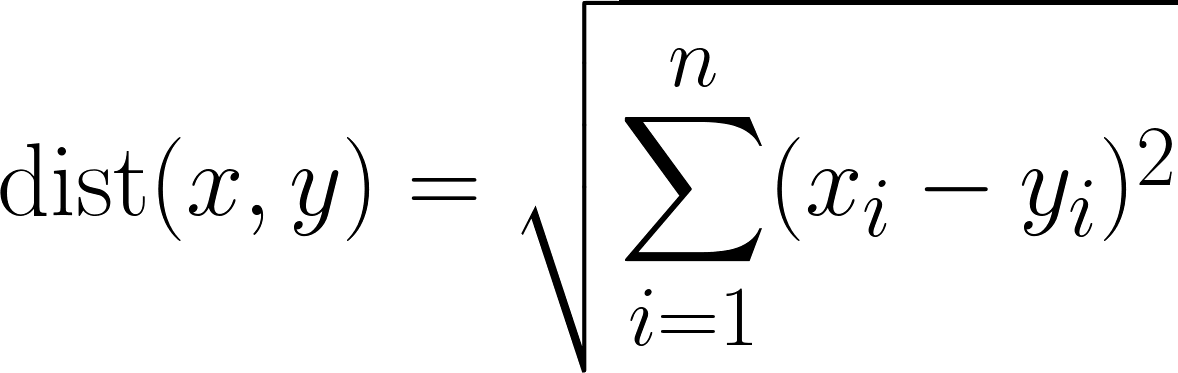


Abaixo segue os passos principais do algoritmo:

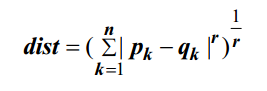
* Quando alguém passar um padrão de entrada:
  + Encontrar os k pontos mais próximo a este padrão
  + Devolver a resposta de maior ocorrência associada aos k

Em seguida, os três métodos matemáticos para calcular as distâncias dos vizinhos

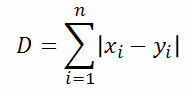
Distância Euclidiana



Distância de Mikowski



Distância de Manhattan



1. **Descrição do Código**

O código implementado em Java que realiza o algoritmo Knn lê dois arquivos de entrada, cada arquivo contém um subconjunto de amostras de um conjunto total de amostras, isto é, a base de amostras completa. O primeiro arquivo é o conjunto de amostras de **treinamento**, é um subconjunto da base de amostras que contém n amostras para cada classe. A base de treinamento é usada apenas para memorização, que será usada na fase de teste. O segundo arquivo contém a base de **teste** que corresponde as n’ amostras que serão posteriormente classificadas de acordo com seus k vizinhos mais próximos, que estão na base de treinamento. Geralmente as n’ amostras de teste são as amostras restantes da base completa, excluindo as amostras da base de treinamento (não deve haver convergência de amostras entre as duas bases)

Após a leitura dos arquivos, é realizado o cálculo das distâncias usando os três métodos matemáticos supracitados. Cada método de cálculo de distância é realizado seis vezes para cada valor ímpar de k que está num intervalo entre 1 e 11 (1,3,5,7,9 e 11). Para cada configuração de k é gerada na saída a matriz de confusão que fornece dados do quantitativo de classificação de cada classe (A matriz faz uma relação entre a **quantidade desejada** de classificações para cada classe e a **quantidade obtida** de classificações para cada classe). Com base nos dados contidos na matriz de confusão, as seguintes métricas são definidas até então:

* **Taxa de acerto**: (soma dos acertos/quantidade de amostras de teste) \* 100
* **Taxa de acerto[Classe]:** (soma dos acertos da classe/quantidade de amostras da classe) \* 100. [Taxa específica para cada classe]

Os cálculos das distâncias dos k vizinhos com suas respectivas matrizes de confusão e métricas serão realizados 18 vezes (6 para cada método de cálculo de distância). Por fim, é definido outras métricas mais abrangentes (que medem o desempenho para cada classe e para cada método de cálculo de distância, além de medir o desempenho global da máquina). As métricas globais estão definidas abaixo:

* **Taxa de acerto global:** Soma(todosAcertos)/18
* **Taxa de acerto[Classe]:** Soma(todosAcertosClasse)/18 [Taxa específica para cada classe]
* **Taxa de acerto global para Distância Euclidiana:** Soma(todosAcertosDE)/6
* **Taxa de acerto global para Distância de Minkowski:** Soma(todosAcertosDMin)/6
* **Taxa de acerto global para Distância de Manhattan:** Soma(todosAcertosDMan)/6

1. **Experimentos**
   1. **Descrição da base de dados – Vertebral Column Data Set**

A base de dados selecionada para ser usada pelo código foi extraído do site da UCI Learning Machine Repository (<http://archive.ics.uci.edu/ml/index.php>) cuja denominação é **Vertebral Column Data Set**. De acordo com o site, a base consiste num “*conjunto de dados que contém valores para seis características biomecânicas que são usadas para classificar pacientes ortopedistas em 3 classes (normal, hérnia de disco ou espondilossistese) ou 2 aulas (normal ou anormal)*”.

Em outras palavras, essa base pode ser usada para realizar diagnósticos de doenças da coluna vertebral, como a hérnia de disco e a espondilossistese, bem como classificar uma coluna vertebral como normal. Então uma máquina de aprendizagem que usa Knn pode ser usada para sugerir possíveis diagnósticos dadas as características biomecânicas da coluna vertebral de cada pessoa contido na base de dados.

A base possui ao todo 310 instâncias(amostras) e 6 características, além de 3 classes (Disk Hernia (DH), spondilolysthesis(SL) e Normal (NO)).

* 1. **Resultados do Experimento**

O Experimento foi realizado usando uma base de treinamento de 120 amostras (40 para cada classe) e uma base de teste de 190 amostras (20 para DH, 110 para SL e 60 para NO). Abaixo segue uma tabela para cada método de cálculo de distância que fornece as respectivas saídas com respeito as métricas e a matriz de confusão geradas para cada um dos 6 cálculos (pois são 6 configurações de k)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Distância Euclidiana** | | | | | |
| **K** | **Taxa de Acerto DH** | **Taxa de Acerto SL** | **Taxa de Acerto NO** | **Taxa de Acerto Total** | **Matriz de Confusão** |
| **1** | **85,00%** | **88,18%** | **65,00%** | **80,53%** | **17 0 3**  **5 97 8**  **12 9 39** |
| **3** | **75,00%** | **90,91%** | **60,00%** | **79,47%** | **15 0 5**  **8 100 2**  **22 2 36** |
| **5** | **80,00%** | **90,91%** | **55,00%** | **78,42%** | **16 0 4**  **9 100 1**  **25 2 33** |
| **7** | **85,00%** | **90,91%** | **51,67%** | **77,89%** | **17 0 3**  **9 100 1**  **27 2 31** |
| **9** | **90,00%** | **90,00%** | **58,33%** | **80,00%** | **18 0 2**  **10 99 1**  **23 2 35** |
| **11** | **85,00%** | **90,91%** | **60,00%** | **80,53%** | **17 0 3**  **9 100 1**  **21 3 36** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Distância de Minkowski (r=4)** | | | | | |
| **K** | **Taxa de Acerto DH** | **Taxa de Acerto SL** | **Taxa de Acerto NO** | **Taxa de Acerto Total** | **Matriz de Confusão** |
| **1** | **95,00%** | **87,27%** | **63,33%** | **80,53%** | **19 0 1**  **7 96 7**  **12 10 38** |
| **3** | **85,00%** | **90,91%** | **58,33%** | **80,00%** | **17 0 3**  **8 100 2**  **22 3 35** |
| **5** | **85,00%** | **90,91%** | **56,67%** | **79,47%** | **17 0 3**  **9 100 1**  **23 3 34** |
| **7** | **80,00%** | **90,91%** | **55,00%** | **78,42%** | **16 0 4**  **9 100 1**  **25 2 33** |
| **9** | **85,00%** | **90,91%** | **60,00%** | **80,53%** | **17 0 3**  **9 100 1**  **22 2 36** |
| **11** | **85,00%** | **90,00%** | **58,33%** | **79,47%** | **17 0 3**  **10 99 1**  **23 2 35** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Distância de Manhattan** | | | | | |
| **K** | **Taxa de Acerto DH** | **Taxa de Acerto SL** | **Taxa de Acerto NO** | **Taxa de Acerto Total** | **Matriz de Confusão** |
| **1** | **80,00%** | **89,09%** | **68,33%** | **81,58%** | **16 0 4**  **3 98 9**  **11 8 41** |
| **3** | **55,00%** | **93,64%** | **63,33%** | **80,00%** | **11 0 9**  **5 103 2**  **19 3 38** |
| **5** | **80,00%** | **95,45%** | **63,33%** | **83,68%** | **16 0 4**  **4 105 1**  **19 3 38** |
| **7** | **80,00%** | **92,73%** | **61,67%** | **81,58%** | **16 0 4**  **6 102 2**  **20 3 37** |
| **9** | **90,00%** | **91,82%** | **63,33%** | **82,63%** | **18 0 2**  **7 101 2**  **19 3 38** |
| **11** | **90,00%** | **90,91%** | **63,33%** | **82,11%** | **18 0 2**  **9 100 1**  **19 3 38** |

* 1. **Análise da variação do K**

Com o valor de k variando entre valores ímpares do intervalo entre 1 a 11, dada a base de treinamento e a de teste, o que se percebe é que a taxa de acerto total se mantém no intervalo entre 77,89% a 82,11%. Abaixo segue a tabela com o cálculo total de taxa de acerto para cada valor de k

|  |  |
| --- | --- |
| **K** | **Taxa** |
| **1** | **80,88%** |
| **3** | **79,91%** |
| **5** | **80,52%** |
| **7** | **79,29%** |
| **9** | **81,05%** |
| **11** | **80,70%** |

Percebe-se que a taxa de acerto não é diretamente proporcional ao valor de k, ou seja, a taxa de acerto não aumenta em relação a todos os valores de k, o que se observa é que a taxa de acerto varia, não apresentando um gradualismo ascendente da taxa de acerto, em outras palavras, a taxa de acerto ora aumenta ora reduz (Isso significa que a eficácia do Knn é uma técnica com muitos fatores de análise, não dependendo unicamente e substancialmente do valor de k em especial). Apesar disso, pelo menos para os valores sugeridos de k, a taxa de acerto aparentemente se mantém num intervalo constante entre 79% a 81%.

Com respeito aos métodos de cálculo de distância, segue abaixo a tabela que resume os valores totais das taxas de acerto de cada método de cálculo de distância para cada valor de k

|  |  |  |  |
| --- | --- | --- | --- |
| **K** | **Taxa de Acerto DE** | **Taxa de Acerto DMin** | **Taxa de Acerto DMan** |
| **1** | **80,53%** | **80,53%** | **81,58%** |
| **3** | **79,47%** | **80,00%** | **80,00%** |
| **5** | **78,42%** | **79,47%** | **83,68%** |
| **7** | **77,89%** | **78,42%** | **81,58%** |
| **9** | **80,00%** | **80,53%** | **82,63%** |
| **11** | **80,53%** | **79,47%** | **82,11%** |
| **Média** | **79,47%** | **79,73%** | **81,93%** |

Nenhum dos métodos de cálculo de distância manteve uma ascendência da taxa de acerto total com o aumento de k, porém percebe-se que a Distância de Manhattan correspondeu melhor ao aumento de k que os demais por manter a taxa de acerto não inferior a 80%. A Distância Euclidiana teve a menor taxa de acerto (77,89%) quando k=7 e também a menor média total (79,47%). A Distância de Minkowski não manteve a taxa de acerto que não fosse inferior a 80%, apesar de ser ligeiramente superior à média obtida pela Distância Euclidiana. Então, é conclusivo que a Distância de Manhattan correspondeu melhor ao aumento do valor de k em relação aos demais métodos de cálculo de distância, apesar que a diferença não possa ser considerada relevante.

Em relação as três classes da base em questão, segue abaixo a tabela que resume as taxas de acerto das três classes para cada valor de k

|  |  |  |  |
| --- | --- | --- | --- |
| **K** | **DH** | **SL** | **NO** |
| **1** | **86,66%** | **88,18%** | **65,55%** |
| **3** | **71,66%** | **91,82%** | **60,55%** |
| **5** | **81,66%** | **92,42%** | **58,33%** |
| **7** | **81,66%** | **91,51%** | **56,11%** |
| **9** | **88,33%** | **90,91%** | **60,55%** |
| **11** | **86,66%** | **90,60%** | **60,55%** |
| **Total** | **82,77%** | **90,90%** | **60,27%** |

É notável o contraste entre a precisão de acerto na classe SL (90,90%) com a classe NO (60,27%). A classe DH, apesar de manter uma boa média de 82,77% teve uma redução notável em sua taxa de acerto quando k=3 (71,66%), o que pode denotar uma certa imprevisibilidade do experimento que pode ter sido acarretado por um erro de normalização da base ou um ruído. Em geral, a classe SL teve uma boa precisão, ao passo que a classe NO, na realidade, não correspondeu de forma efetiva para nenhum dos valores de k se for considerado que uma média acima de 70% seria a mínima taxa aceitável.

* 1. **Análise das métricas globais**

Abaixo segue a tabela referente as métricas globais

|  |  |
| --- | --- |
| **Métrica (Taxa de Acerto Global)** | **Taxa** |
| **DH** | **82,78%** |
| **SL** | **90,91%** |
| **NO** | **60,28%** |
| **Distância Euclidiana** | **79,47%** |
| **Distância de Minkowski** | **79,74%** |
| **Distância de Manhattan** | **81,93%** |
| **Total** | **80,38%** |

Em geral, a máquina teve um bom desempenho de 80,38%, explicado de forma simples, isso significa que a máquina acertaria cerca de 80 de 100 amostras analisadas.

As classes DH e SL manteve suas taxas de acerto num intervalo aceitável entre 82% a 91%, entretanto, conforme já abordado, é preocupante a baixa precisão na classe NO, principalmente pelo fato de que a classe NO representa o perfil clínico normal em relação a outras duas patologias, isto significa que se essa máquina fosse usada para diagnósticos de patologias na coluna vertebral, a mesma provavelmente poderia fornecer um número considerável de falsos positivos, ou seja, a máquina poderia indicar uma patologia inexistente na coluna vertebral de um paciente (cerca de 40 % do que seria normal, foi indicado como uma patologia), Logo, essa máquina não seria considerado eficiente no que se diz a respeito a descartar patologias da coluna vertebral.

Com respeito aos métodos de cálculo de distância dos k vizinhos, todos os três métodos obtiveram uma taxa de precisão aceitável e além disso as taxas de cada método são próximos entre si (79,74% – 81,93%), logo, é evidente que a escolha do método de cálculo de distância não se mostrou decisivo ou que simplesmente não exerce ampla influência na precisão ou não das classificações.

1. **Conclusão**

Os experimentos realizados forneceram um quadro completo acerca da medida da eficiência do algoritmo em relação ao valor de k. **Percebe-se que o aumento de k não necessariamente ofereceu mais precisão, logo o valor de k, somente em si, não pode ser decisivo na classificação das amostras (k não é diretamente proporcional a precisão)**, isso é devido ao fato de que o algoritmo Knn, como qualquer algoritmo de aprendizagem de máquina, depende de um ajuste fino de todos os seus parâmetros para uma maior precisão na classificação, que inclui um apropriado quantitativo de amostras, tanto na base de treinamento como na base de teste, além da normalização completa dessas bases e também da necessária eliminação de ruídos (amostras fora do padrão da base). Entretanto não é uma tarefa trivial definir tais valores ajustados para melhorar a eficiência do algoritmo, alguns desses valores somente podem ser conhecidos apenas após várias execuções do algoritmo, afinal, é um algoritmo de aprendizagem e erros fazem parte de tal processo.

Com respeito à aplicação prática de tal algoritmo na base selecionada foi percebido através dos experimentos que o mesmo, de forma geral, é considerado **eficiente ao diagnosticar patologias na coluna vertebral, porém não atingiu um desempenho aceitável ao descartar as doenças** (Apenas 60% de eficiência). Não se pode inferir através dos dados dos experimentos a causa para baixa precisão na classe NO(Normal). Existem as hipóteses de que a causa pode ter sido ocasionada ou por ruídos ou por uma normalização incompleta das bases.

**A precisão poderia ter sido maior se houvesse uma análise mais detalhada das bases para averiguar a normalização das mesmas e a existência de ruídos**. Entretanto, para um primeiro ciclo de aprendizagem, o algoritmo teve um desempenho aceitável e é possível que a máquina de fato possa aprender apenas com poucos ciclos adicionais. Recomenda-se a realização de mais de um ciclo de aprendizagem para o algoritmo, além de uma análise mais detalhada das bases antes de sua execução. Independentemente dos resultados obtidos, ficou esclarecido que o Knn ainda pode ser considerado uma boa opção para a resolução de problemas de classificação, pelo menos a princípio (depende da extensão do problema).

1. **Referências**

IA08 – Aprendizagem de Máquina: <https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxyZnBuaWZwZXxneDozYzJkYzNhMzA1Y2IyNzZm>

K-nearest neighbors algorithm – Wikipédia (em inglês):

<https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm>

UCI Learning Machine Repository – Vertebral Column Data Set:

http://archive.ics.uci.edu/ml/datasets/vertebral+column