Implementação dos classificadores KNN e DMC utilizando o MATLAB

Gustavo Siebra Lopes[1]

Resumo: O relatório apresenta os resultados obtidos para dois tipos de classificadores o KNN e DMC para a os padrões da flor de Iris, a implementação foi feita no Matlab.

Introdução

1 Preparação da base

1.1 Base de dados da flor de íris

Foi utilizada uma base de dados disponibilizada pela UCI Machine Learning Repository [UCI15] onde são definidas 3 classes (Iris Setosa, Iris Versicolour, Iris Virginica) e 4 parametros por classe (comprimento e largura da sépala e pétala), a base possui 150 padrões diferentes de Iris dividido em 50 padrões para cada classe.

A base foi dividida usando o modelo *holdout*, este método consiste em dividir o conjunto total de dados em dois subconjuntos mutuamente exclusivos, um para treinamento (estimação dos parâmetros) e outro para teste (validação). O conjunto de dados pode ser separado em quantidades iguais ou não. Uma proporção muito comum é considerar 2/3 dos dados para treinamento e o 1/3 restante para teste. (Kohavi, 1995)

Após o carregamento da base foi realizado a normalização dos dados separadamente para cada atributo. Identificando o mínimo e o máximo que foram normalizados na faixa [0,1].

1.2 Análise das características

Na **figura 1** é apresentada a matriz de características. Essa matriz consiste de gráficos formados pelos pares de características combinadas.

[1] gustavosiebra@gmail.com

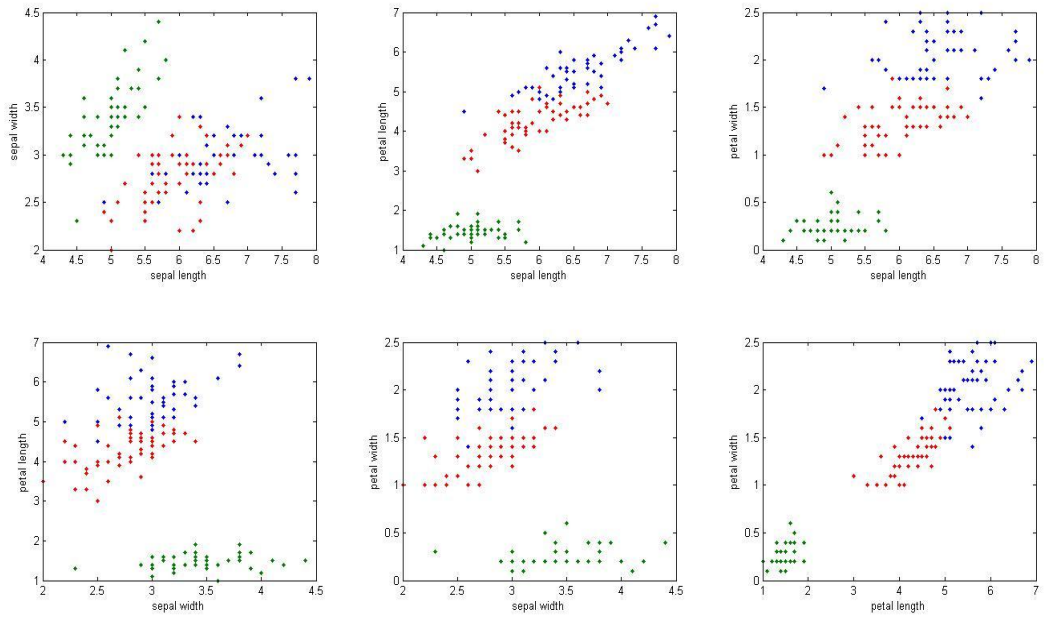


Figura 1 – Matriz de características.

Podemos perceber que a classe setosa pode facilmente ser separada das outras utilizando a largura ou o comprimento da pétala como variável Já para as duas outras características essa separação não é tão simples, perceba a sobreposição dos histogramas em todos os atributos bem como a mistura das classes nos gráficos de dispersão.

2 KNN

2.1 Introdução

Este é um método de classificação não paramétrico, que estima o valor da função densidade de probabilidade. Em reconhecimento de padrões , o algoritmo $k$nn é utilizado como método de classificação de objectos (elementos) com base na formação dos exemplos próximos no espaço dos elementos. $k$nn é um tipo de "Lazy Learning", onde a função é aproximada apenas localmente e toda computação é adiada para a classificação.

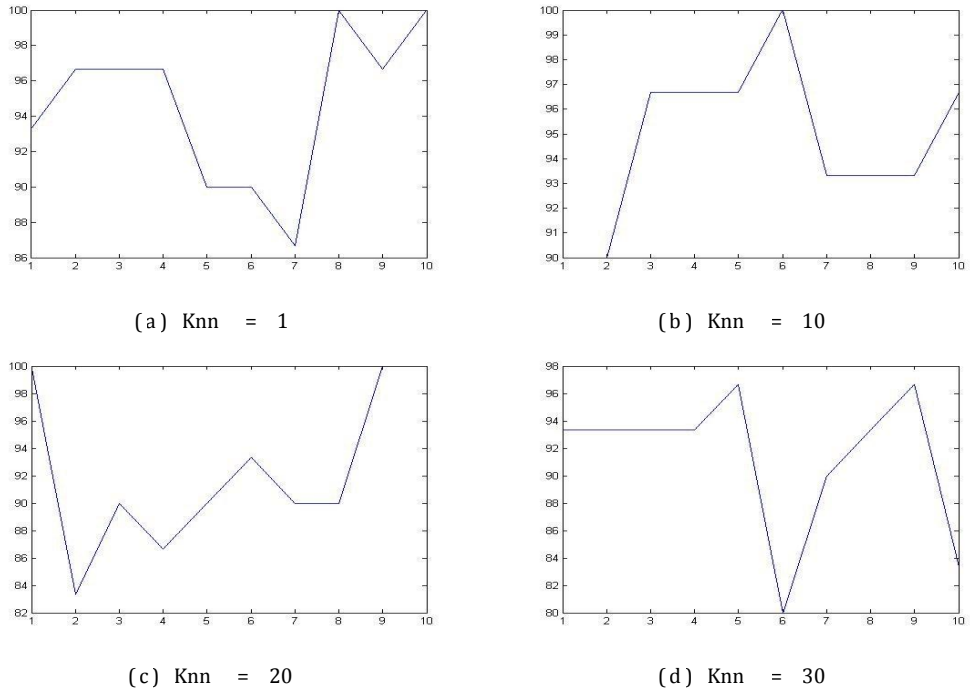
O 1-NN é um caso específico do KNN para este classificador o número de vizinhos que deve ser levado em conta é apenas 1. Uma importante vantagem do KNN frente ao 1-NN é a capacidade de controlar a interferência a ruídos presentes na base de dados. Quanto maior o valor de K mais informação será levada em

consideração para a identificação da label da amostra e considerando também que a quantidade de ruído presente próximo a amostra é suficientemente pequena frente ao valor de K este pode ser contornado.

2.2 Resultados obtidos

2.2.1 Variação do valor de K

Foram feitos uma série de teste variando o valor de K bem como testes de generalização onde foi realizado a minimização da quantidade de dados necessários na base para uma correta classificação. Na **figura 2**, o gráfico mostra o resultado da acurácia para 10 repetições do classificador para o valores de k = 1, k = 10, k = 20 e k = 30.



(a) Knn = 1 (b) Knn = 10

(c) Knn = 20 (d) Knn = 30

Figura 2: Análise do algoritmo KNN para diferentes K

Na **figura 3** o grafico mostra a média das acurácias para 120 repetições de cada valor de K. É percebido que partir de K = 70 a média começa a dimunuir.

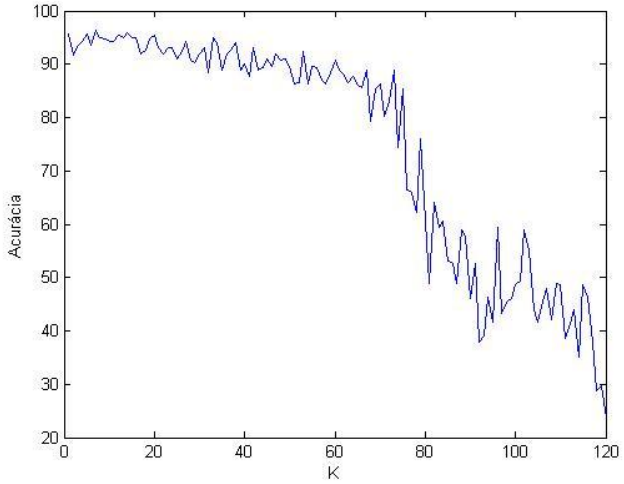


Figura 3 – Análise da Acurácia Média em função do valor de k.

A **figura 4** mostra o desvio padrão para 120 repetições de cada valor de K. É percebido que para o problema da íris a variação do valor de k não representa diferença significativa para um valor de k até 15. Na faixa [60 a 120] a variação da acurácia é grande, o que indica baixa confiabilidade no classificador.

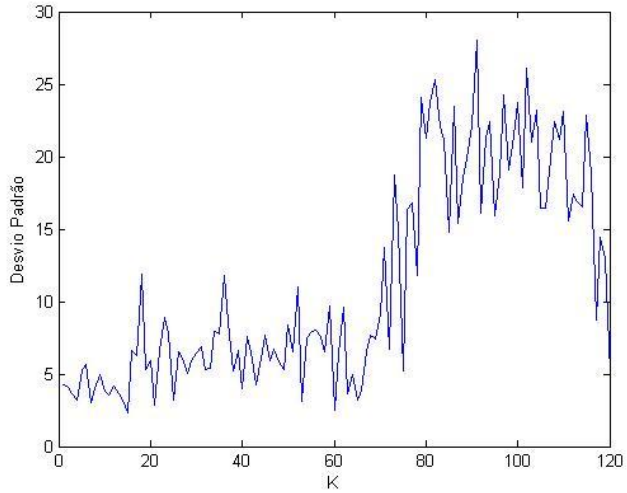


Figura 4 – Análise do Desvio Padrão em função do valor de k.

2.2.2 Matriz Confusão

Na **tabela 1** é exibida a matriz confusão obtida para K igual a 15. Esse valor de k foi escolhido devido aos testes de acurácia em função de k mostrarem que com este valor é obtida a melhor acurácia.

Classe	Setosa	Versicolor	Virgínica
Setosa	12	0	0
Versicolor	0	11	0
Virgínica	0	0	7

Tabela 1 – Matriz confusão obtida após execução do KNN para K = 15.

3 DMC

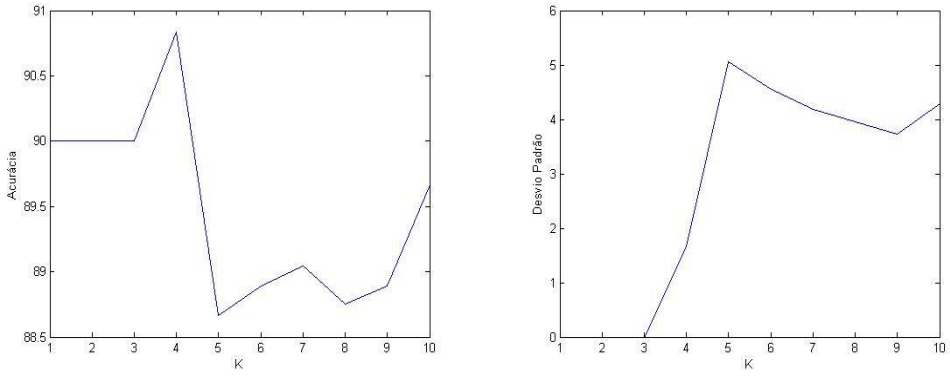
3.1 Introdução

Para cada classe é assumido um centro de massa (também conhecido como centróide). Um objeto pertence a essa classe quando a distância entre ele e o centróide for menor que todas as distâncias entre os outros centróides restantes do espaço de características. O primeiro passo do processo de classificação por distancia mínima é o calculo dos vetores médios (centróides) que representam cada classe por padrões [Scaranti, Bernardi e Plotze, 2010].

3.2 Resultados obtidos

3.2.1 Acurácia Média e Desvio Padrão

Devido a metodologia adotada a posição do centróide varia ao longo do cálculo da acurácia média, , no entanto deve-se ficar atento em atualizar a matriz média para cada padrão classificado. Na **figura 5** são exibidos os resultados para o cálculo do DMC para 10 repetições.



(a) Análise da Acurácia Média. (b) Análise do Desvio Padrão.

Figura 5: Análise do algoritmo DMC.

3.2.2 Matriz confusão

Na **tabela 2** é exibida a matriz confusão após a execução do DMC.

Classe	Setosa	Versicolor	Virgínica
Setosa	7	0	0
Versicolor	0	13	0
Virgínica	3	0	7

Tabela 2 – Matriz confusão obtida após execução do DMC.

Considerações finais

Nesse relatório foi apresentado dois classifcadores, o KNN e DMC. É percebido que o KNN indiferentemente da forma como os dados estão dispostos no espaço, é possível separar as regiões por mais irregulares que sejam. Para o DMC, temos um resultado menos satisfatório, pois não foi atingido uma acurácia tão alta, porém em termos de desempenho o classificador teve um ganho considerável.

Referências

[UCI15] Uci machine learning repository, 2015. Disponível em: <http:// archive.ics.uci.edu/ml/>.

FRUTUOSO, R. L. Identificação de Órgãos foliares utilizando as wavelets de daubechies. ***XIV Workshop de Informática Médica*, FCT/UNESP**, v. 1, n. 1, p. 211–126, 2013. ISSN none. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/wvc/2010/0037.pdf>. Citado na página 8.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **International joint Conference on artificial intelligence**. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

Scaranti, A. ; Bernardi, R. ; Plotze, R. O.. (2010) “Identificação de Órgãos Foliares Utilizando as Wavelets de Daubechies”. In: WVC'2010 - **VI Workshop de Visão Computacional.**