

# Relatório de atividades: Uso do classificador Bayesiano

David Clifte\*

2015, v-1.0

## Resumo

Este trabalho apresenta os resultados obtidos ao aplicar o classificador bayessiano. A implementação foi feita no Matlab<sup>TM</sup>

**Palavras-chaves:** Bayes. Reconhecimento de padrões.

## Introdução

### 1 Preparação da base

#### 1.1 Base de dados da flor de íris

A base de dados da flor de íris criado por Fisher (??). Nessa base de dados as informações obtidas das flores foram o comprimento e a largura das pétalas e sépalas de 3 tipos de flor de íris, virgínica, versicolor e setosa. Cada tipo de flor possui 50 instancias.

#### 1.2 Base de dados da dermatologia

A base de dados da dermatologia foi criada por Altay, (GüVENIR; DEMİRÖZ; ILTER, 1998). Nessa base foram coletadas informações de pacientes que possuíam sintomas de doenças de pele. As doenças são psoríase, seborreica dermatite, líquen plano, pitíriase rosea, dermatite crônica, e pitíriase rubra pilaris. Foram coletadas 34 características de um total de 366 pacientes.

---

\*cliftedavid@gmail.com

### 1.3 Base de dados da coluna vertebral

Nessa base de dados as informações obtidas das flores foram o comprimento e a largura das pétalas e sépalas de 3 tipos de flor de íris, virgínica, versicolor e setosa. Cada tipo de flor possui 50 instancias.

### 1.4 Normalização e codificação

Após o carregamento da base foi realizado apenas a normalização dos dados e a codificação dos rótulos. A normalização foi realizada separadamente para cada atributo. Foi identificado o máximo e o mínimo do atributo  $p$  e todos os valores foram normalizados na faixa  $[0,1]$ .

A codificação do rótulo foi feita no modelo 1-de-k(1-of-K ou one-hot encoding), nesse modelo o rótulo é codificado em um vetor onde cada posição do vetor representa uma classe. Nesse modelo para uma quantidade  $m$  de classes temos um vetor com  $m$  posições e a classe  $k$  é representada por um vetor onde todas as outras posições diferentes de  $k$  possuem o valor zero e a posição  $k$  possui o valor 1. Dessa forma as três classes possíveis da íris foram codificadas em um vetor de 3 posições, onde a posição 1, 2 e 3 representam a classe setosa, versicolor e virgínica respectivamente. O mesmo foi repetido para as bases da dermatologia e da coluna vertebral.

### 1.5 Análise das características

#### 1.5.1 Base da íris

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

Na diagonal principal é apresentado o histograma do atributo. Devido a natureza continua dos atributos do problema, pois o comprimento e largura da sépala e pétala podem assumir qualquer valor real O histograma foi obtido após a realização da quantização dos atributos. Cada atributo foi quantizado em 15 possíveis valores com faixas de mesma largura. A largura da faixa foi obtida da seguinte forma  $(max_p - min_p)/15$ , onde max e min são os valores máximos e mínimos do atributo  $p$ .

Podemos perceber que a classe setosa pode facilmente ser separada das outras utilizando a largura ou o comprimento da pétala como variável. Nos histogramas localizados na parte inferior isso pode ser verificado, pétalas com comprimentos menores que 0,25 ou sépalas com largura menor que 0,3 de comprimento são claramente da classe setosa. 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.



Figura 1 – Matriz de características.

1.5.2 Base da dermatologia

1.5.3 Base da coluna vertebral

## 2 Classificador de Bayes

### 2.1 Introdução

Dada uma classificação entre  $M$  classes, o classificador de Bayes faz a seleção de um dado  $x$  com base na probabilidade de  $w_i$  dado um  $x$ ,  $P(w_i|x)$ . Assim temos:

$$x \in w_i \iff P(w_i|x) \geq P(w_j|x) \forall i \neq j \quad (1)$$

Gaussiana				
DataSet	média	desvio Padrão	máximo	mínimo
Íris	0.972414	0.026246	1.000000	0.931034
Vertebra	0.726344	0.057531	0.854839	0.629032
Dermatologia	0.888584	0.035382	0.972603	0.821918
Matriz de covariacia $\Sigma_i = \sigma^2 I \forall i$				
DataSet	média	desvio Padrão	máximo	mínimo
Íris	0.916092	0.050972	1.000000	0.793103
Vertebra	0.237097	0.046041	0.338710	0.145161
Dermatologia	0.509589	0.055077	0.616438	0.424658
Matriz de covariacia diagonal				
DataSet	média	desvio Padrão	máximo	mínimo
Íris	0.955172	0.035247	1.000000	0.862069
Vertebra	0.787097	0.055776	0.887097	0.661290
Dermatologia	0.936530	0.025560	0.986301	0.904110
Janela de Parzen				
DataSet	média	desvio Padrão	máximo	mínimo
Íris	0.951724	0.040091	1.000000	0.862069
Vertebra	0.823656	0.052892	0.935484	0.693548
Dermatologia	0.908219	0.026227	0.958904	0.849315

Tabela 1 – TTTT

## 2.2 Função de densidade de probabilidade

## 2.3 Metodologia

A avaliação do classificador de Bayes foi feita de várias formas ao longo deste trabalho, veja a subseção 2.4.

Na sessão 2.4 são apresentados os resultados obtidos ao aplicar o classificador de bayes utilizando como função de densidade probabilidade a Gaussiana e a janela de parzen. Para a janela de parzen foi exibido o melhor resultado para um determinada largura de janela do tipo gaussiana. O tamanho da janela foi determinado após uma pesquisa linear entre valores na faixa de 0.001 e 3 para o valor da variância.

## 2.4 Resultados obtidos

Na tabela 1 são exibidos os resultados obtidos ao realizar a classificação utilizando a regra de Bayes com funções de densidade probabilidade (PDF) diferentes. Para cada base de dados tem-se o resultado utilizando a Gaussiana como PDF e a janela de parzen.

#### 2.4.1 Busca por tamanho de janela ótima de Parzen

Abaixo, figura 2 são exibidos os resultados obtidos ao realizar a busca pelo tamanho da janela ótima de parzen para cada base de dados. Foram considerados todos os atributos disponíveis em cada base, estes foram normalizados e codificados apropriadamente assim como dito na subseção 1.4.

A linha ao centro de cada gráfico é a acurácia as outras duas são os limites. Os limites são determinados a uma distância de um  $\sigma^2$  (desvio padrão) acima e abaixo da acurácia média. Pode-se perceber que indiferentemente da base de dados o aumento considerável da janela de parzen causa uma redução da acurácia do classificador. Isso é causado por uma maior interferência dos vizinhos no cálculo da probabilidade condicional o que faz com que seja levado em conta apenas as probabilidades a priori para a determinação da classe para um dado  $x$ . Isso pode ser observado quando a variância, largura da janela de parzen, é maior que 2 na base da íris ou maior que 0.7 na base da coluna vertebral.

Os pontos máximos destas curvas nos indicam o valor ótimo da janela de parzen. Desta forma temos na tabela 1 o valor de acurácia e tamanho da janela para as três bases.

### 2.5 Análise da matriz de covariancia na definição da região de decisão

Abaixo são apresentadas as regiões de decisões

#### 2.5.1 Matrizes de covariancia distintas. $\Sigma_i \neq \Sigma_j \forall i \neq j$

Com matrizes de covariancias distintas temos uma região de decisão quadrática.

#### 2.5.2 Matrizes de covariancia iguais. $\Sigma_i \Sigma \forall i$

#### 2.5.3 Janela de Parzen

#### 2.5.4 Região de decisão

#### 2.5.5 Matriz Confusão

Na tabela ?? é exibida a matriz confusão obtida para K igual a 10. 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. Além da matriz confusão a tabela ?? traz os resultados, falso-positivo, falso-negativo, verdadeiro-positivo e verdadeiro-negativo.

## 3 Segmentação utilizando Classificador de Bayes

A segmentação de uma imagem é realizada levando em consideração a intensidade dos pixels nas três componentes RGB. A definição dos dados

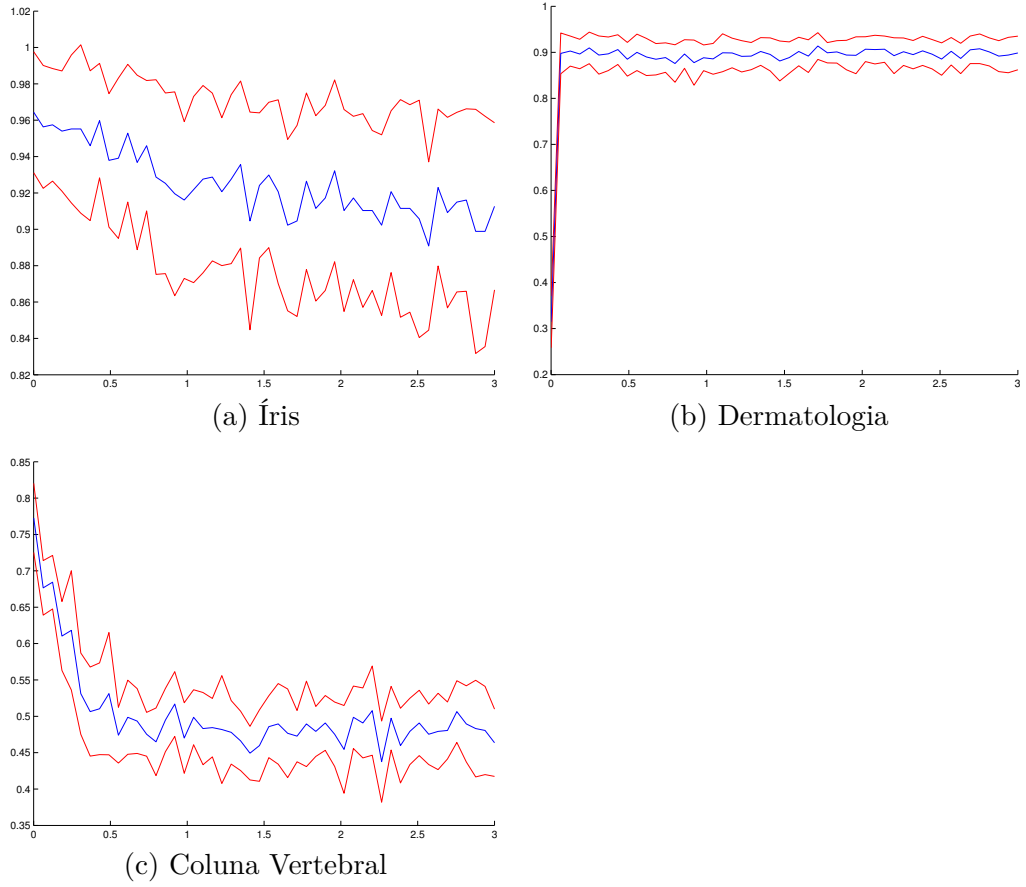


Figura 2 – Resultados da busca pela largura ótima da janela de parzen.

de de uma classe é feita a partir da seleção de  $k$  regiões da imagem, não necessariamente de mesma área  $A_i$ . A área é utilizada para calcular a probabilidade a priori por isso classes com uma diferença de área muito grande podem impactar na classificação. Assim temos:

$$P(w_i) = \frac{A_i}{\sum_{i=1}^k A_J} \quad (2)$$

### 3.1 PDF Gaussiana

Nas figuras 5 e 7 são apresentado o resultado da segmentação de duas imagens utilizando o classificador Bayessiano. Os parâmetros  $\sigma^2$  e  $\mu$  são estimados de acordo com os dados de cada classe individualmente. Assim temos na figuras 5 e 7 o resultado deste classificador sobre 2 imagens distintas.

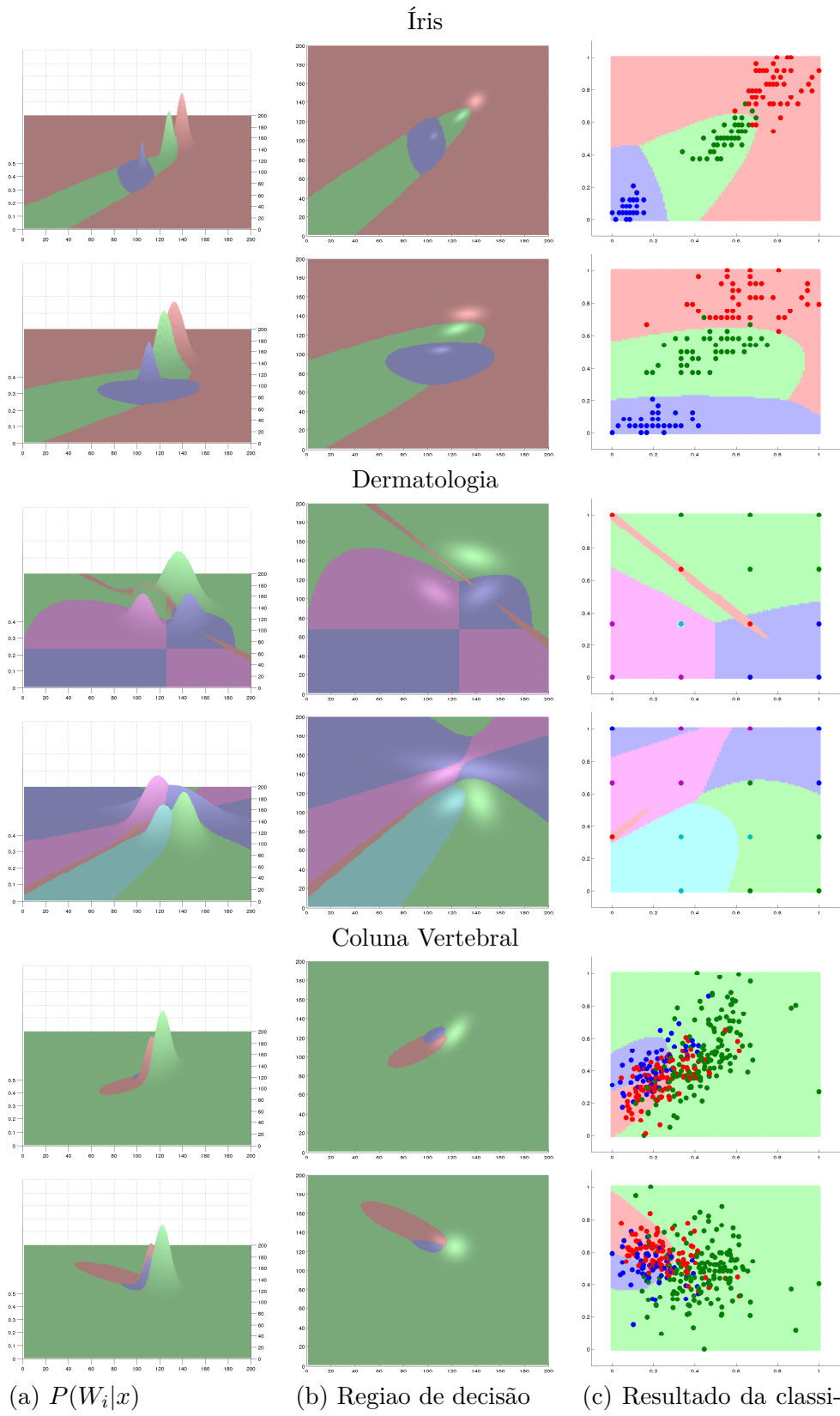


Figura 3 – Resultados da busca pela largura ótima da janela de parzen.

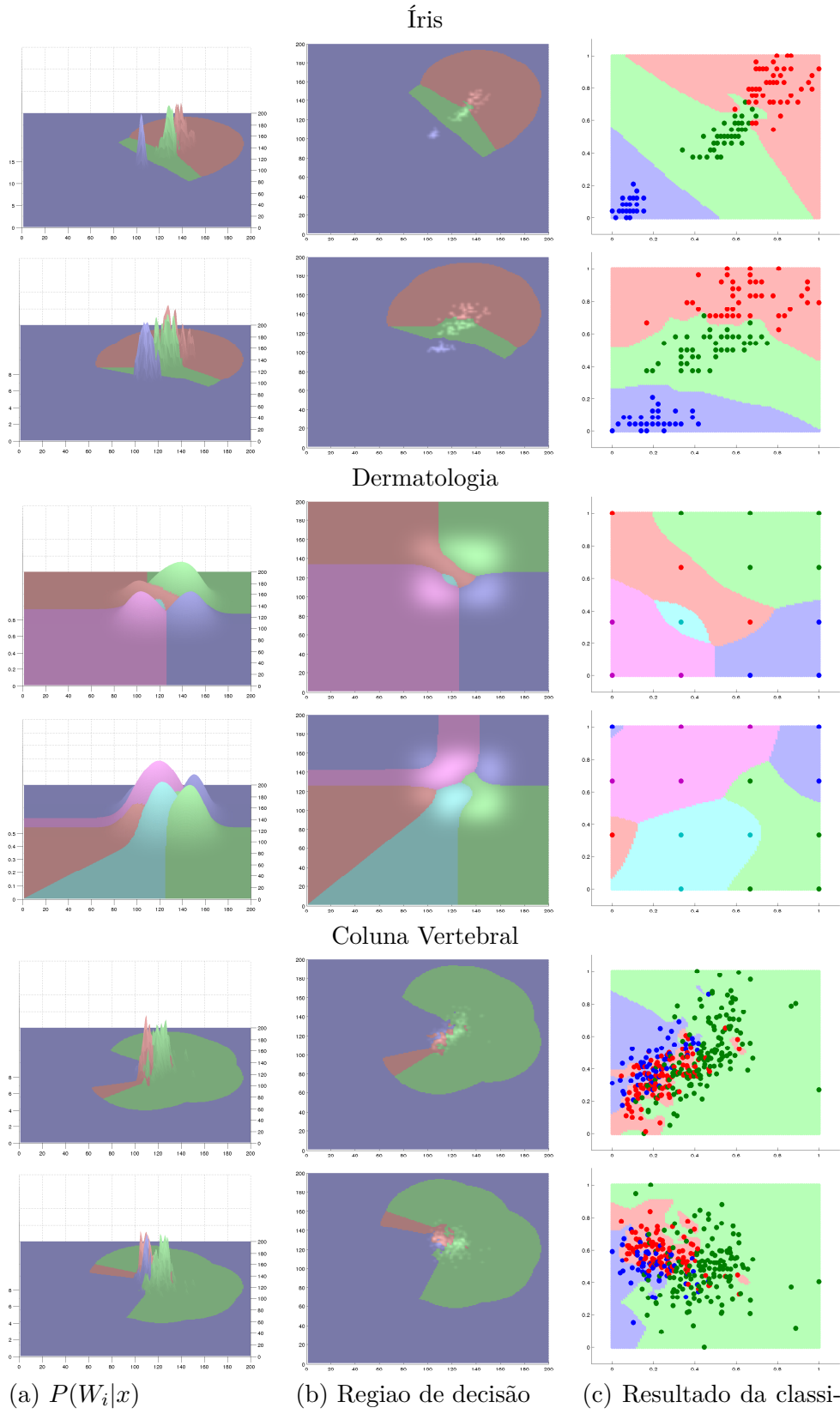


Figura 4 – Região de decisão calculada utilizando janela de parzen com largura  $\sigma^2 = 0.05I$ .



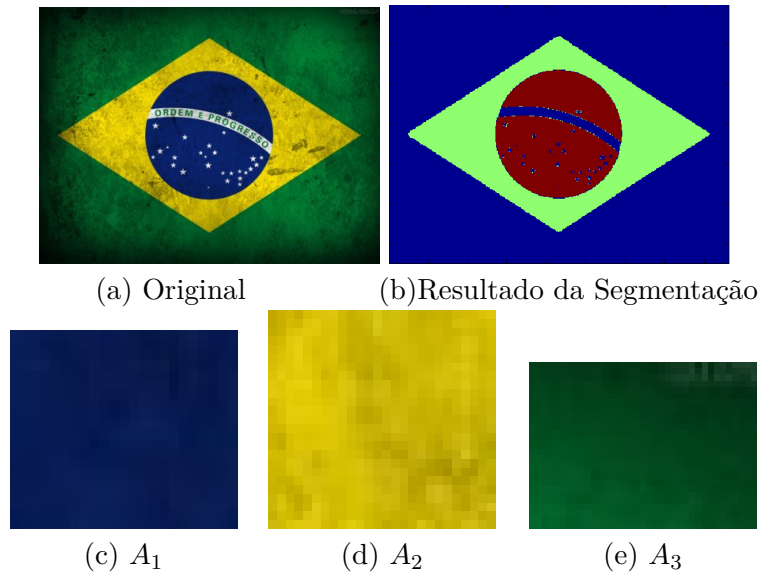


Figura 5 – Resultado da segmentação utilizando o classificador Bayessiano. Na linha inferior são apresentados as amostras utilizadas para definir as classes. As áreas utilizadas são: (c) céu, (d) areia e (e) coqueiro

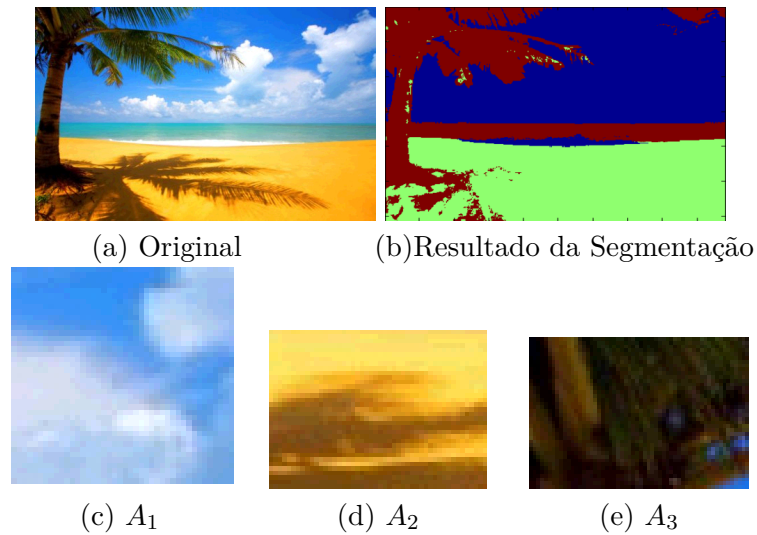


Figura 6 – Resultado da segmentação utilizando o classificador Bayessiano. Na linha inferior são apresentados as amostras utilizadas para definir as classes. As áreas utilizadas são: (c) céu, (d) areia e (e) coqueiro

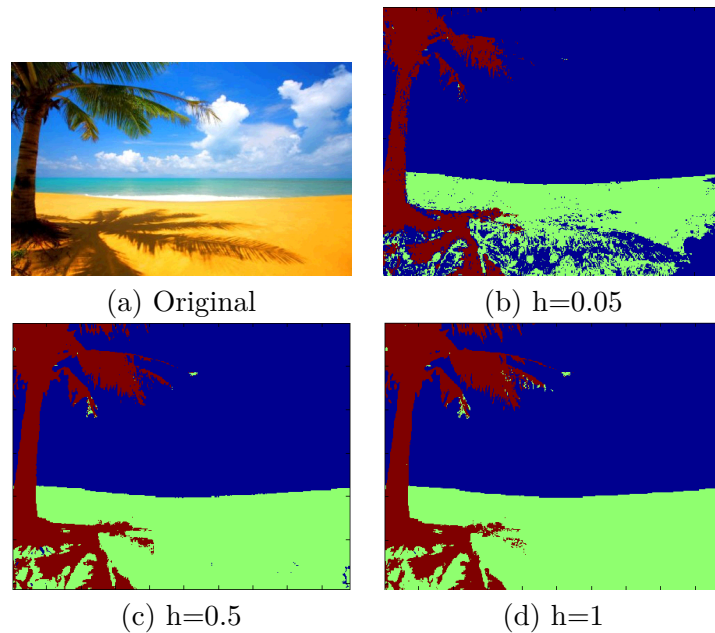


Figura 7 – Resultado da segmentação utilizando o classificador Bayessiano com janela de parzen. Com o aumento do valor de  $h$  pode-se notar que a segmentação se torna mais suave.

### 3.2 PDF Janela de Parzen

Abaixo é apresentado o resultado da segmentação de uma imagem utilizando o classificador Bayessiano, desta vez utilizando janela de parzen como função de densidade probabilidade. Foram realizados três testes com três variancias da janela, valor de  $h$ , distintas. Os valores testados foram, 0.05, 0.5 e 1. Com o aumento do valor de  $h$  houve uma maior suavização

## Considerações finais

## Referências

GÜVENİR, H. A.; DEMİRÖZ, G.; ILTER, N. Learning differential diagnosis of erythematous-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine*, v. 13, p. 147, 1998. Citado na página 1.