Narizes Electrónicos para Identificação de Plásticos

Reconhecimento de Padrões

Leissi Castañeda León, Jorge Guevara Díaz

Instituto de Matemática e Estatística
Universidade de São Paulo,São Paulo-Brasil
leissicl@vision.ime.usp.br, jorjasso@vision.ime.usp.br

Abstract

Este trabalho descreve o processo de extracção de características e classificação de padrões usando a teoria de decisão bayesiana para o reconhecimento de plásticos pelo seu cheiro usando narizes electrónicas. Para testar o classificador de maneira experimental foi usado a técnica de k-fold cross validation.

Categories and Subject Descriptors D.3.2 [*Programming languages*]: Language Classifications—Object-oriented languages; D.2.2 [*Software Engineering*]: Design Tools and Techniques—Object-oriented design methods; D.2.2 [*Software Engineering*]: Design Tools and Techniques—Petri nets

General Terms k-fold cross validation, teoria de decisão bayesiana.

1. Introduction

A presente resenha descreve o processo de reconhecimento de cheiros de plásticos utilizando a teoria de decisão bayesiana. O projecto foi iniciado pelo professor Dr. Jonas Gruber do Instituto de Química da USP. O trabalho é baseado nos dados proporcionados por ele e sua equipe de trabalho.

Este trabalho foi focalizado em dois aspectos, o primeiro aspecto foi o mal calculo dos valores das extracção das características[1], para isto usamos algumas técnicas de processamento digital de sinais. O outro aspecto foi o desenho do classificador e a avaliação experimental deste com a técnica de k-fold cross validation.

O informe esta organizado da seguinte maneira: A secção 2 descreve o processo de extracção de características, algumas técnicas de processamento digital de sinais foram usadas. A secção 3 faz un análise do espaço de características em 1 e 2 dimensões, presentado alguns gráficos obtidos. A secção 4 descreve o desenho do classificador usando a teoria de decisão bayesiana. A secção 5 presenta os resultados obtidos no processo de classificação onde o calculo do error médio foi feito usando a técnica de k-fold cross validation. Finalmente secção 6 presenta a discussão dos resultados obtidos.

[1] comentado pelo professor Jonas em aula, onde falou que alguns valores de Ra eram negativos

Reprinted from , [Unknown Proceedings], , pp. 1–4.

2. Extracção de características

O vector de características foi gerado pelo cálculo da resposta relativa R_a (Fig. 1) nos dados definida como:

$$R_a = \frac{G_2 - G_1}{G_1} \tag{1}$$

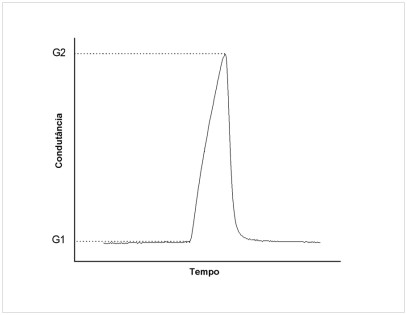

Figure 1. cálculo da resposta relativa R_a.

Seja

$$X^i = (x_1^i, x_2^i, ...x_T^i) \tag{2}$$

o vector que representa os dados, onde $0 \leq i \leq 4$. representa os quatro sensores que foram usados na leitura do cheiro dos plásticos (Fig 2). Cada x_t corresponde a medida da condutância no tempo t, para $1 \leq t \leq T$.

Para encontrar as posições corretas dos valores G_1 e G_2 é necessário fazer um processamento das sinais dos quatro sensores. Para isto fizemos o seguinte

1. Filtrado.
2. Analise da segunda derivada.
3. Etiquetado.

2.1 Filtrado

Cada sinal correspondente ao sensor X^i tem muitos componentes de alta frequência (Fig. 3) associados, isto faz um pouco dificultoso encontrar as corretas posições dos valores G_1 e G_2, Então fizemos uma filtragem usando em primer lugar um filtro da media no domínio do tempo.

$$X_{fil}^i = X^i * h_m \tag{3}$$

Onde $*$ é o operador convolução e h_m é a mascara de convolução de tamanho m, neste caso o filtro de media $h_m = (1/m, ...1/m)$

Logo fizemos uma filtragem usando um valor umbral θ no domínio da frequência para eliminar os componentes de alta frequência.

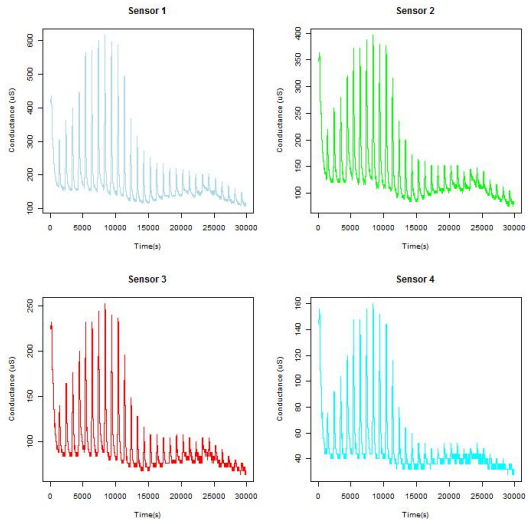

Figure 2. Os quatro sensores.

Figure 3. Componentes de alta frequência presentes na medida da condutância de cada sensor.

$$H^i(k) = \sum_{t=0}^{N-1} x^i(t) e^{-j2\pi kt/N}, 0 \leq k \leq (N-1) \qquad (4)$$

Onde H^i é a transformada de Fourier do sensor i, $x^i(t)$ é a componente t de X^i_{fil} . As frequências acima de uma frequência determinada pelo valor $|k| \geq \theta$ são estabelecidas com valor zero.

Finalmente a sinal filtrada é a inversa da transformada de Fourier do H^i

$$Y^i_t = \frac{1}{N} \sum_{k=0}^{N-1} H^i(k) e^{j2\pi kt/N}, 0 \leq t \leq (N-1) \qquad (5)$$

2.2 Análise da segunda derivada

O análise da segunda derivada é necessário para determinar os valores *candidatos* correspondentes aos mínimos locales e os valores dos máximos locales da sinal de cada sensor. que corresponderam as medidas G_1 e G_2 respectivamente. Seja IndMin(j) a posição da sinal filtrada com un candidato de mínimo local e IndMax(j) a posição da sinal filtrada com un candidato de máximo local. Estas posições são calculadas tendo-se em conta os valores onde a primeira derivada é zero.

$$\text{IndMin(j)=t, se } Y^i(t)'' > 0 \text{ e } Y^i(t)' = 0 \qquad (6)$$

$$\text{IndMax(j)=t, se } Y^i(t)'' < 0 \text{ e } Y^i(t)' = 0 \qquad (7)$$

Finalmente os mínimos locales son calculados na *sinal original* X^i a partir das posições candidatas obtidas na sinal filtrada Y^i da seguinte maneira:

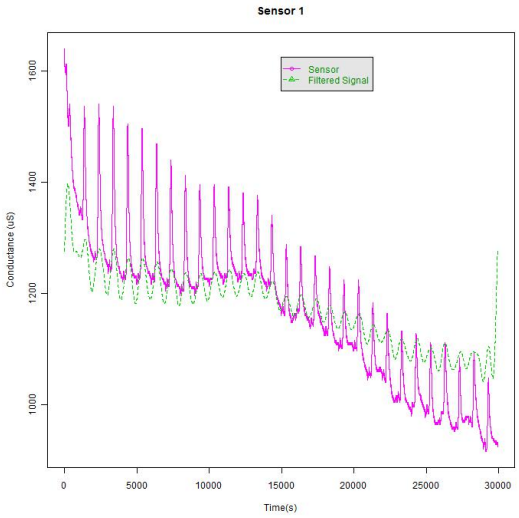

Figure 4. Sinal original e sinal filtrada.

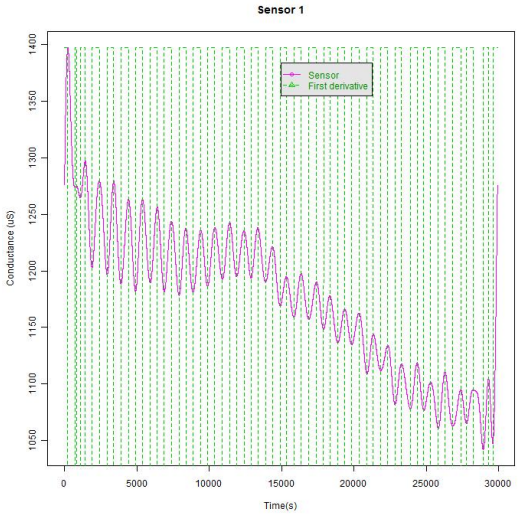

Figure 5. Primeira derivada.

$$\min\{(x^i_{IndMax(j)}, ..., x^i_{IndMax(j+1)}\} \qquad (8)$$

e os máximos locales da seguinte maneira:

$$\max\{x^i_{IndMin(j)}, ..., x^i_{IndMin(j+1)}\} \qquad (9)$$

para todas as posições j em IndMax e IndMin.

2.3 Etiquetado

Para fazer o etiquetado final de valores G_1 e G_2 tivemos en conta os seguintes casos:

1. Descartar valores mínimos (máximos) consecutivos. A estratégia foi eliminar o primer de esses valores. (Fig. 7.a).
2. Procurar sempre o valor mínimo entre a metade da distancia de máximos e o seguinte máximo. para evitar mínimos muito distanciado do máximo (Fig. 7.b).

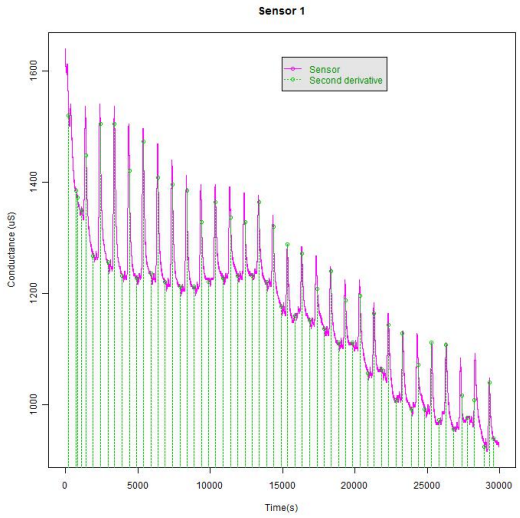

Figure 6. Segunda derivada.

3. Verificar que sempre se commence com un mínimo, para fazer o calculo correto do valor Ra.

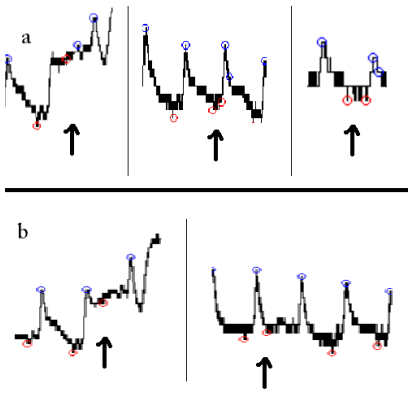

Figure 7. a) Descartar valores máximos o mínimos consecutivos. b) Mínimos muito distanciados de um maximo

O calculo dos valores G_1 e G_2 usando todo este processamento é bastante *robusto*, por exemplo não foram encontrados valores negativos para os Ra's como foi discutido em aula. Finalmente os valores Ra são calculados usando a equação 1.

3. Gráficos dos Ra's

Para estudar o comportamento das características se fizeram as gráficas dos Ra's para os sete plásticos.[1]. O gráfico da Fig. 9 corresponde as comparações das densidades dos histogramas do primeiro Ra[2] de todos os sete plásticos. No caso do gráfico da Fig. 10 corresponde as comparações das densidades normais para dados

[1] O arquivo graficos.pdf contem todos os gráficos gerados os quais não são incluídos neste articulo por causa do espaço limitado

[2] O arquivo graficos.pdf contem os gráficos gerados para as comparações entre os segundos, terceiros e quartos Ra's

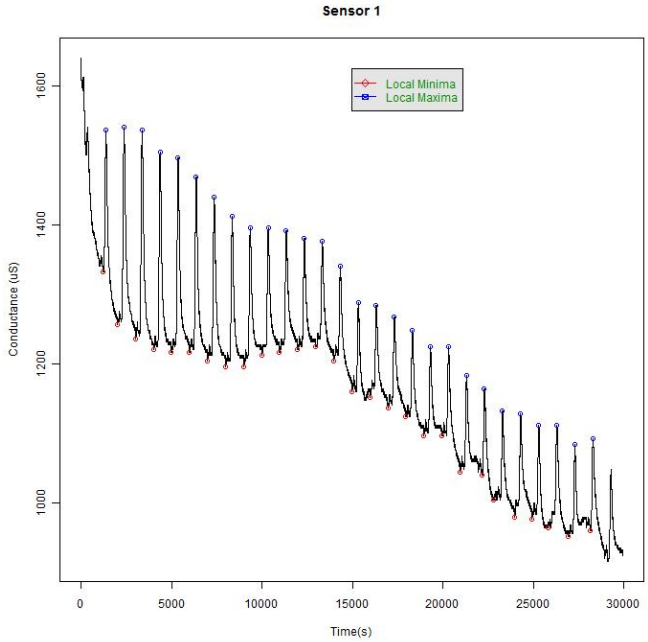

Figure 8. Máximos e mínimos locais correspondentes aos valores G_1 e G_2.

bivariates para o primeiro e quarto dos Ra's dos sete plásticos[3]. E possível observar a partir dos gráficos (Fig. 9 e Fig 10) que

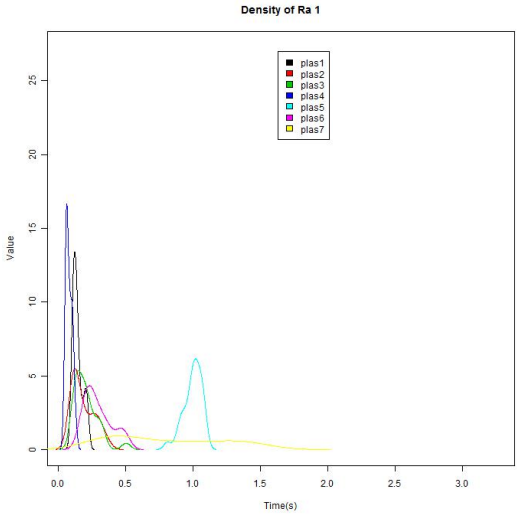

Figure 9. Comparação das densidades de primer Ra dos sete plásticos

o plástico 5 se diferencia muito das outras classes de plástico. É possível observar nos resultados obtidos (Secc. 5) que o error de classificação para o plástico 5 é 0.00%. Outro plástico que também se diferencia muito é o plastico 7 mais ele tem alguns erros na

[3] O arquivo graficos.pdf contem os gráficos gerados para todas as passiveis comparações dois a dois dos Ra's dos sete plásticos

classificação (Secc. 5) pois possivelmente algum de seus valores estão sobrepostos nas regiões dos outros plásticos.

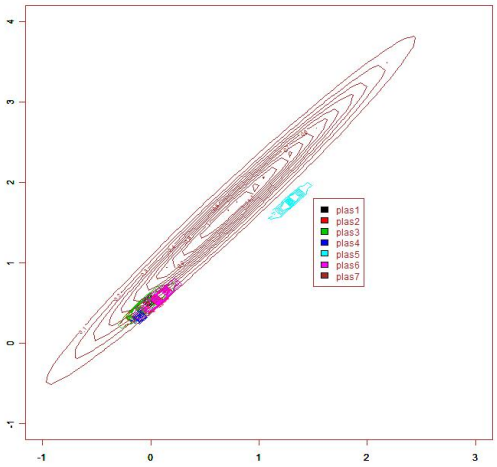

Figure 10. Comparação das densidades de primer Ra e o quarto Ra dos sete plásticos

4. Classificador bayesiano

Foi construído um classificador bayesiano e foi testado experimentalmente usando k-fold cross validation.

$$g(x) = \max(g_1(x,), ..., g_7(x,)) \tag{10}$$

O classificador $g(x)$ faz a assinação do vector de características x, a uma das sete classes de plástico (a classe com maior probabilidade a posteriori).

Os valores dos $g_i(x)$ foram calculados usando a formula de Bayes, onde foi assumido valores de probabilidade a priori iguais para todas as sete classes de plástico, e o valor do likelihood foi calculado usando uma função de densidade normal multivariate, a qual foi desenhada com o valor $\sum_i$=arbitrario [1].

O classificador bayesiano foi construído com base nas distribuições normais calculadas para cada classe, calculando a media e matriz de covarianza para cada classe (14 parâmetros).

Para o k-fold fizemos a divisão dos dados com $k = 10$, na qual uma partição foi estabelecida como dados de test e o resto dos dados foram estabelecidos como dados de entrenamento, fizemos isto para as 10 possíveis partições, finalmente a media dos erros foram calculados.

5. Resultados

A Avaliação do classificador foi feito usando k-fold cross validation com $k = 10$. A tabela 1 descreve a media do error do classificador por iteração por cada classe de plástico, onde $E1$ é media do error por iteração da classe 1, E2 é media do error por iteração da classe 2 y assim de maneira sucessiva. ET corresponde a media do error total do classificador.

Iteração	1	2	3	4	5
E1(%)	0.00	33.33	22.22	16.67	13.33
E2(%)	100.00	50.00	44.44	33.33	26.67
E3(%)	100.00	83.33	55.56	41.67	40.00
E4(%)	0.00	33.33	22.22	16.67	20.00
E5(%)	0.00	0.00	0.00	0.00	0.00
E6(%)	0.00	0.00	0.00	0.00	0.00
E7(%)	0.00	0.00	0.00	0.00	0.00
ET(%)	28.57	28.57	20.63	15.48	14.29
Iteração	6	7	8	9	10
E1(%)	11.11	9.52	8.33	7.41	11.67
E2(%)	33.33	38.10	33.33	35.19	41.67
E3(%)	44.44	42.86	45.83	51.85	56.67
E4(%)	16.67	14.29	12.50	11.11	10.00
E5(%)	0.00	0.00	0.00	0.00	0.00
E6(%)	11.11	19.05	25.00	29.63	36.67
E7(%)	0.00	0.00	0.00	5.56	10.00
ET(%)	16.67	17.69	17.98	20.15	23.85

Tabela 1. Media do error por iteração no 10-fold cross validation

A tabela 2 descreve a media do erro total do classificador, usando os valores de Ra1,Ra2,Ra3,Ra4 para cada componente x_t do vector de características no caso de ETI, para o caso do ETII foram usados os valores Ra1,Ra2,Ra3 para cada componente x_t, para o caso de ETIII foram usados os valores Ra1,Ra2 para cada componente x_t.

Iteração	1	2	3	4	5
ETI(%)	28.57	28.57	20.63	15.48	14.29
ETII(%)	35.71	36.91	26.19	19.64	16.67
ETII(%)	35.71	39.29	26.19	22.02	17.62
Iteração	6	7	8	9	10
ETI(%)	16.67	17.69	17.98	20.15	23.85
ETII(%)	19.44	22.11	23.72	26.64	30.40
ETII(%)	19.44	20.07	24.43	27.28	30.98

Tabela 2. Media do error por iteração no 10-fold cross validation com o vector de características em 2 3 e 4 dimensões

6. Discussão

Pode-se observar que o classificador tem um error de 23.85% quando alcança a iteração 10. Também a boa separação da classe 5 das demais classes. Isto foi possível observar nos gráficos das figuras 9 e 10, além disso é possível observar que a classe 7 só tem erros nas iterações 9 e 10, pode-se observar das figuras 9 e 10 que a aproximação da distribuição dos dados desta classe usando uma normal multivariate é grande com respeito as demais distribuições e estas ultimas ficam dentro da distribuição para a classe 7 (excepto a distribuição de la classe 5). Uma possível solução poderia ser a eliminação de outliers no processo de entrenamento do classificador ou algum outro processamento dos dados. Na tabela 2 pode-se notar que o classificador tem um maior desempenho quando são usados todos os valores dos Ra.

References

[1] Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern classification. Wiley, 2 edition, November 2001.

[2] Broughton, S., Allen;Bryan, Kurt, M. Discrete Fourier analysis and wavelets. Applications to signal and image processing. (2009)

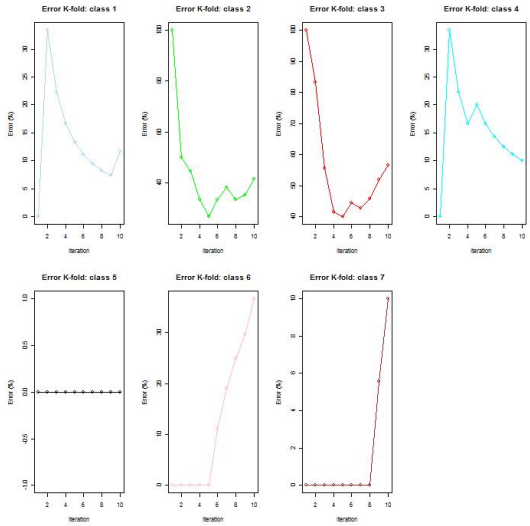

Figure 11. Media do error por iteração da validação 10-fold por classe, grafico correspondente aos valores da tabela 1 excepto o valor ET.

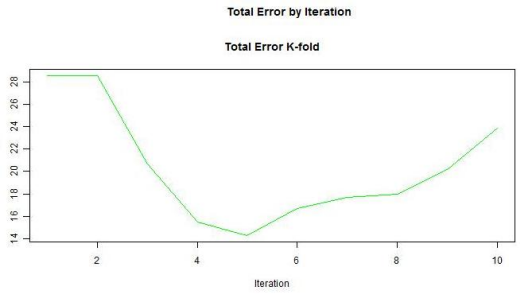

Figure 12. Media do error por iteração da validação de 10-fold, grafico corresponde ao valor ET da tabela 1