Estimadores de Kernel Aplicados na Modelagem e Classificação de Dados de Eficiência de Quimioterapia Neoadjuvante

Maria Fernanda B. Wanderley[1], Antônio P. Braga[1], Eduardo M. A. M. Mendes[1], René Natowicz[2], Roman Rouzier[3]

[1]Departamento de Engenharia Eletrônica
Universidade Federal de Minas Gerais, Brazil

[2]Département d'informatique
Université Paris-Est, ESIEE - Paris, France

[3]Service de gynecologie
Hôpital Tenon, France

{mfbw, apbraga}@ufmg.br, emmendes@cpdee.ufmg.br

Abstract. *On this paper we propose an application of local statistical models to the problem of identifying patients with pathologic complete response (PCR) to neoadjuvant chemotherapy. The idea of using local models is separate the input space (with data from PCR and NoPCR patients) and build a model for each partition. After the construction of the models we used bayesian classifiers and logistic regression to classify patients on the two classes.*

Resumo. *Neste trabalho propomos a aplicação de modelos estatísticos locais ao problema da identificação de pacientes com resposta patológica completa (PCR) para quimioterapia neoadjuvante. A ideia de utilizar modelos locais é particionar o espaço de entrada (com dados de pacientes PCR e NoPCR) e construir um modelo para cada partição. Após a construção dos modelos, foram utilizados classificadores bayesianos e regressão logística para classificar os pacientes em PCR e NoPCR.*

1. Introdução

Em 2008, baseado nos dados internacionais disponíveis mais recentes, foram estimados 12,4 milhões de novos casos e 7,6 milhões de mortes por câncer no mundo. Os tipos mais comuns em termos de incidência são o de pulmão (1,52 milhões de casos), de mama (1,29 milhões de casos) e de cólon (1,15 milhões de casos) [Boyle and Levin 2008].

O câncer de mama, foco do estudo desse projeto, pode ser tratado através de quimio e radioterapia e/ou cirurgia, que pode ser parcial (quadrantectomia) ou radical (com a retirada total da mama). No caso do câncer operável, pode-se fazer uso da quimioterapia antes da cirurgia (quimioterapia neoadjuvante) para diminuir o tamanho do tumor e evitar que o mesmo se espalhe por outros órgãos.

A quimioterapia, porém, apresenta muitos efeitos colaterais uma vez que age não somente nas células cancerosas mas também em outras células do corpo que possuem a mesma característica de crescimento e multiplicação acelerados que os tumores. Dentre

os efeitos colaterais estão anemia e diminuição da resistência a infecções causadas pela ação nas células produtoras dos glóbulos sangüíneos vermelhos e brancos, queda de pelos e cabelos devido à ação nas células do folículo piloso, náuseas, vômitos e diarréia, em decorrência da ação nas células do aparelho digestivo, além da dificuldade de engravidar e parada da menstruação, já que as células do sistema reprodutor também são afetadas. Tendo isso em vista, é importante não submeter a paciente ao tratamento quimioterápico neoadjuvante se o mesmo não for obter nenhum resultado.

Em geral, prever a eficiência do tratamento utilizando características clínicas dos pacientes não funciona adequadamente e, por isso, utiliza-se a informação baseada na expressão de mRNA para obter perfis de diversos tumores e assim realizar a predição [Natowicz et al. 2008b]. Os perfis obtidos através de biópsias podem ser correlacionados com características como o tamanho do tumor, o estágio em que se encontra, recorrência do tumor e sensibilidade ao tratamento. Uma resposta patológica completa (PCR), na cirurgia está correlacionada com um excelente resultado, enquanto uma resposta incompleta (NoPCR) está associada a um resultado ruim. É importante prever corretamente se uma paciente é PCR ou NoPCR pois no segundo caso outras alternativas de tratamento podem ser buscadas.

Neste trabalho propomos o uso dos dados de pacientes PCR e NoPCR para obter modelos estatísticos que nos permitam realizar uma predição satisfatória de futuros pacientes PCR.

2. Referencial Teórico

2.1. Estimação não-Paramétrica de Densidades

Em geral, problemas práticos, não são modeláveis por estimadores paramétricos que possuem densidade unimodal [Thompson and Tapia 1990]. Por isso, a estimação não-paramétrica de densidades, que pode ser usada com qualquer tipo de distribuição, é comumente utilizada na modelagem desse tipo de problema.

Em comparação com estimadores paramétricos, onde o estimador tem sua estrutura fixada e os parâmetros (estimadores deles) da função densidade são as únicas informações a serem guardadas, os estimadores não-paramétricos não possuem estrutura fixa e dependem de toda amostra para obter uma estimativa da função densidade.

Nas sub-seções a seguir apresentamos um método de modelagem de dados, o *kernel density estimation* e dois tipos de classificadores, o bayesiano e o regressor logístico.

2.2. Kernel Density Estimation

Para compreender os *kernel estimators* é necessário primeiramente compreender o conceito de histogramas, uma vez que foram suas desvantagens que motivaram a criação dos *kernel estimators*.

Os histogramas são os mais antigos e utilizados estimadores de densidade [Silverman 1986]. Dada uma origem x_0 e uma caixa de largura h, as caixas do histograma são definidas pelos intervalos $[x_0 + mh, x_0 + (m + 1)h[$ para inteiros m positivos ou negativos.

Um histograma é definido por:

$$\hat{f}(x) = \frac{1}{nh} * \{no.\ de\ X_i\ na\ mesma\ caixa\ de\ x\}$$

onde n é o número total de amostras e h é a largura da caixa.

Na figura 1 apresentamos um histograma que estima a densidade de duas classes de dados que possuem média e desvio padrão diferentes. O número total de amostras é igual a 10000 e o número de caixas utilizadas é igual a 50.

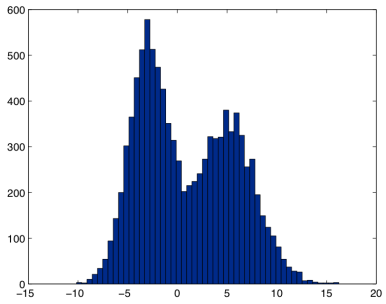

Figura 1. Histograma

Devido as suas características, histogramas não provém uma estatística suave, dependem da largura das células e dos pontos que são estabelecidos como limites, além de apresentarem descontinuidade quando $h \to 0$. Tais problemas são amenizados utilizando-se *kernel density estimation*.

Um *kernel estimator* com kernel K é definido por:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

onde h é a largura da janela, K é uma função de kernel que satisfaz a condição $\int_{-\infty}^{\infty} K(x)dx = 1$ e $x_1, x_2, \ldots, X_n$ são amostras independentes e identicamente distribuídas de uma variável aleatória. No presente trabalho utilizamos o kernel box ou uniforme, como função de kernel. O kernel box é dado por

$$K(x) = \frac{1}{2} I_{\{|x| \leq 1\}}$$

onde I é a função característica.

O método de *kernel density estimation* ou Janela de Parzen consiste num histograma contínuo, cujos blocos são centralizados em cada um dos pontos de dados de onde se quer estimar a densidade. A função de kernel utilizada define o formato dos "picos" observados nos dados, sendo o estimador uma soma dos "picos".

Na figura 2 apresentamos um estimador de kernel cuja largura da janela utilizada foi 0,4.

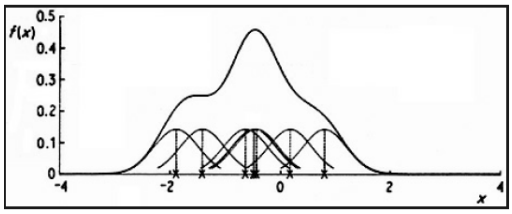

Figura 2. Estimador de kernel mostrando os picos individuais.[Silverman 1986]

2.3. Classificador Bayesiano

Dado um problema genérico de classificação de duas classes para distinguir entre duas classes C_1 e C_2, de acordo com um vetor de características x. Um classificador bayesiano é um classificador probabilístico baseado no Teorema de Bayes, que dadas a probabilidade *a priori* das classes e a verossimilhança das mesmas é capaz de separar os elementos pertence a cada uma das classes [Duda et al. 2000].

A regra de decisão do classificador bayesiano indica que um vetor de características x é atribuído a classe C_i de maior probabilidade *a posteriori* $P(C_i|x)$. Para o problema de duas classes, tal regra pode ser matematicamente escrita como:

$$Classe(x) = \begin{cases} C_1 & \text{se } P(x|C_1) * P(C_1) > P(x|C_2) * P(C_2), \\ C_2 & \text{caso contrário.} \end{cases}$$

onde $P(x|C_1)$ e $P(x|C_2)$ são as verossimilhanças das classes em relação ao vetor x e $P(C_1)$ e $P(C_2)$ são as probabilidades *a priori* pra cada classe.

2.4. Regressão Logística

Métodos de regressão são frequentemente utilizados para descrever a relação entre uma variável de saída e uma ou mais componentes de entrada. Na regressão logística, essa variável de saída é binária, ou seja, o vetor de características de entrada é classificado em duas classes opostas [Hosmer and Lemeshow 2000].

A função logística é dada por

$$f(z) = \frac{1}{1 + e^{-z}}$$

onde z é um conjunto de características e $f(z)$ é a probabilidade de uma determinada saída, dadas as entradas. A função logística aceita como entrada qualquer valor entre mais e menos infinito, porém sua saída é limitada por valor no intervalo $[0, 1]$.

3. Metodologia Proposta

3.1. Obtenção e utilização dos dados experimentais

A triagem clínica foi conduzida no Nellie B. Connaly Breast Center, do M.D. Anderson Cancer Center, da Universidade do Texas [Hess et al. 2006]. 133 pacientes com câncer de mama em estágio I - III foram incluídas na triagem realizada, sendo coletados dados de 82 pacientes em Houston, Estados Unidos, e os dados das outras 51 pacientes coletados em Villejuif, na França. Antes do início do tratamento neoadjuvante foram coletadas

amostras dos tumores utilizando o método de aspiração e, ao final do tratamento, todas foram submetidas a cirurgias para raspagem do local onde o tumor se encontrava, para verificar se houve ou não a resposta patológica completa. Os dados aqui utilizados são públicos e estão disponíveis em http://bioinformatics.mdanderson.org/pubdata.html.

De posse desses dados experimentais,propomos utilizar o método de kernel density estimation para obter uma modelagem local dos dados e, em seguida, um classificador bayesiano para separar os dados de entrada nas classes PCR e NoPCR. Os mesmos dados serão utilizados em um regressor logístico para posteriormente comparmos o desempenho dos dois métodos de classificação.

3.2. Modelo local dos dados - *Kernel Density Estimation*

Propomos a utilização do método de *Kernel Density Estimation*, utilizando um *kernel* do tipo box, para modelar cada uma das classes PCR e NoPCR. Com esse método, é possível estimar as densidades de cada uma das classes para obter a verossimilhança de cada uma e em seguida utilizar o classificador.

O método proposto por esse trabalho possui característica inovadora, uma vez que os trabalhos desenvolvidos até o presente [Natowicz et al. 2008b, Hess et al. 2006, Natowicz et al. 2008a, Natowicz et al. 2008c] baseiam-se, em geral, em modelos globais e não obtiveram resultados definitivos. É bastante provável que isso ocorra devido a características dos dados, como por exemplo o conjunto de dados americanos e franceses não estarem distribuídos estatisticamente de maneira uniforme, dificultando assim a modelagem global. Além disso, os dados disponíveis são escassos e esparsos, o que também dificulta a modelagem global. Neste caso, o tratamento através de modelos e estimadores locais pode ser uma alternativa para o tratamento individualizado por grupos de amostras.

3.3. Metodologia

Dado o conjunto de 30 sondas obtidas por [Natowicz et al. 2008b] foram separadas as três melhores sondas para um estudo e comparação com o método de regressão logística, que é amplamente utilizado na classificação de dados médicos. Propomos duas diferentes abordagens: a primeira onde cada uma das sondas é utilizada separadamente como um classificador e a segunda onde as três sondas são utilizadas juntas.

No primeiro caso, a densidade para cada uma das sondas é estimada, utilizando parte do conjunto de dados. Em seguida, um classificador bayesiano tem como tarefa classificar uma outra parte dos conjuntos de dados, utilizando a informação de verossimilhança estimada pelo estimador de kernel. O mesmo é feito para o regressor logístico.

No segundo caso, após a estimação de densidade para as três sondas com parte do conjunto de dados, o classificador bayesiano utiliza um critério de votação da maioria para decidir em qual classe cada entrada do conjunto de testes será classificada. O regressor logístico, por sua vez, utiliza as mesmas três sondas em conjunto para realizar a classificação.

4. Resultados

Utilizamos como base os dados de triagem clínica apresentados na sub-seção 3.1, que foram divididos da seguinte forma:

- Das 82 pacientes de Houston, EUA:
  - 61 pacientes são No-PCR e 21 são PCR
- Das 51 pacientes de Villejuif, na França:
  - 38 pacientes são No-PCR e 13 são PCR.

Das 22283 sondas existentes para cada paciente utilizamos as três melhores sondas das trinta selecionadas por [Natowicz et al. 2008b] e aplicamos o método de *kernel density estimation* em cada uma delas. Nas figuras 3, 4, 5 apresentamos as funções de densidade estimadas para cada sonda. No eixo das abscissas estão representados os valores das sondas e no eixo das coordenadas estão os valores de densidade de probabilidade estimados. É possível ver que a função de densidade estimada para a segunda sonda difere das demais, tendo sido esta a sonda que apresentou melhor resultado.

Para a estimação das densidades e treinamento do regressor logístico foram usados os dados das 82 pacientes de Houston. Já para o classificador bayesiano e o teste do regressor foram utilizados os dados das 51 pacientes de Villejuif.

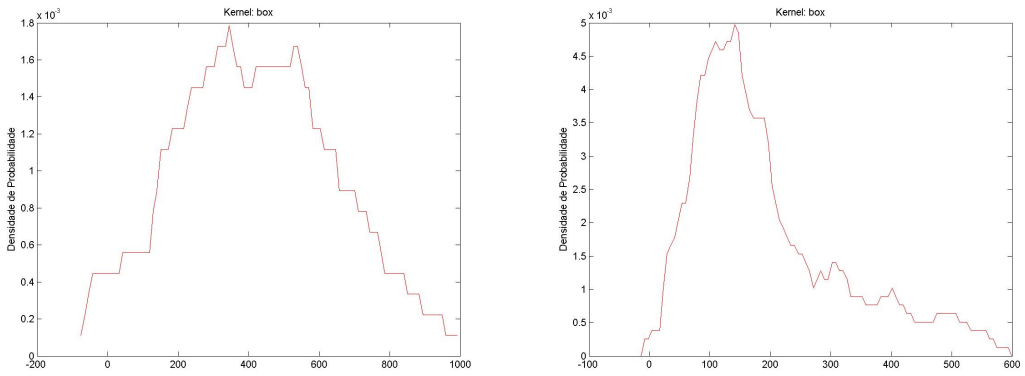

Figura 3. Função de densidade estimada das classes PCR (à esquerda) e NoPCR (à direita) da primeira sonda (213134_x_at).

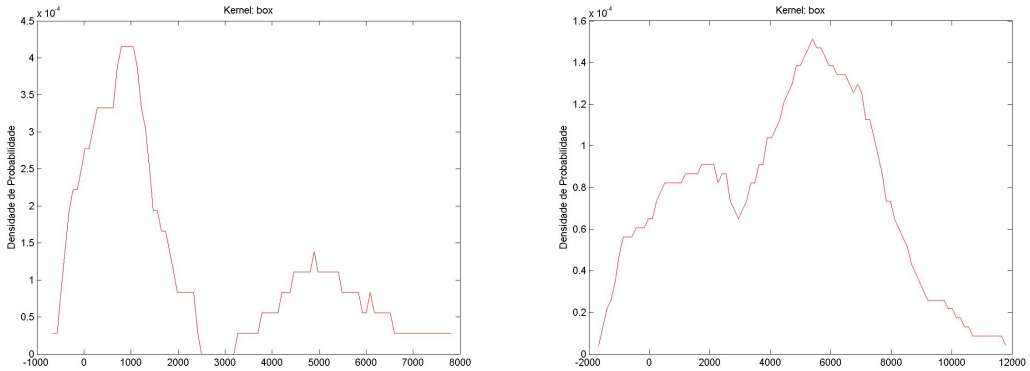

Figura 4. Função de densidade estimada das classes PCR (à esquerda) e NoPCR (à direita) da segunda sonda (205548_s_at).

O resultado, para o conjunto de treinamento de 51 pacientes, será apresentado e analisado em termos de três medidas:

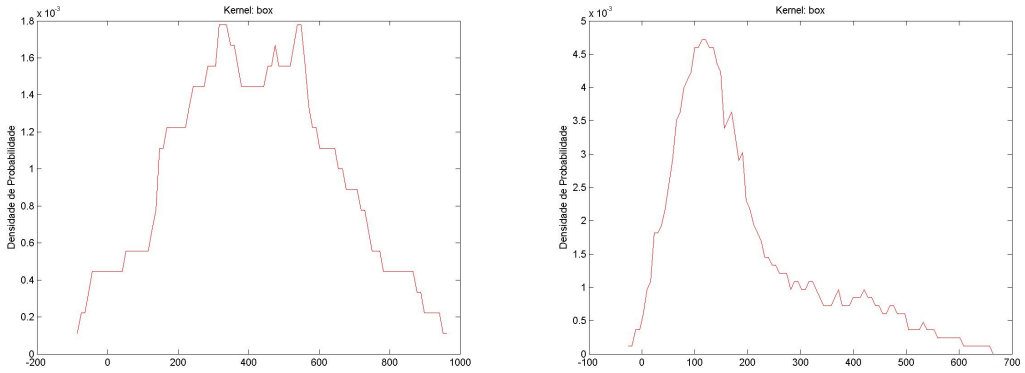

Figura 5. Função de densidade estimada das classes PCR (à esquerda) e NoPCR (à direita) da terceira sonda (209604_s_at).

Tabela 1. Resultados dos dois métodos utilizados para a sonda 1 (213134_x_at).

Método	Classificador Bayesiano	Regressor Logístico
Sonda	213134_x_at	213134_x_at
Ac	0,8627	0,7450
Se	0,6153	0,3846
Es	0,8780	0,8048

- Acurácia (Ac) - número de acertos de classificação para cada classe
- Sensibilidade (Se) - medida de capacidade do método em identificar corretamente os casos PCR
- Especificidade (Es) - medida de capacidade do método em excluir corretamente os casos NoPCR

  A sensibilidade é calculada da seguinte forma:

$$Se = \frac{VP}{VP + FN}$$

onde VP são os casos PCR corretamente preditos e FN são os casos em que a amostra era PCR e foi classificada como NoPCR.

A especificidade é calculada da seguinte forma:

$$Es = \frac{VN}{VN + FP}$$

onde VN são os casos NoPCR corretamente preditos e FP são os casos em que a amostra era NoPCR e foi classificada como PCR.

Como pode-se ver na Tabela 2, a sonda 205548_s_at foi a melhor quando utilizada como um classificador, sendo inclusive melhor do que os resultados em que foram utilizadas três sondas. É importante ressaltar que o valor de sensibilidade ser alto é importante, pois indica que o número de falsos negativos é baixo. Ou seja, pacientes que responderiam positivamente ao tratamento não deixam de ser tratados.

Embora a sonda 205548_s_at tenha sido a melhor, os outros resultados foram superiores ao método de regressão logística e podem ser considerados satisfatórios pois se

Tabela 2. Resultados dos dois métodos utilizados para a sonda 2 (205548_s_at).

Método	Classificador Bayesiano	Regressor Logístico
Sonda	205548_s_at	205548_s_at
Ac	0,9019	0,7450
Se	0,9230	0,3076
Es	0,8994	0,8994

Tabela 3. Resultados dos dois métodos utilizados para a sonda 3 (209604_s_at).

Método	Classificador Bayesiano	Regressor Logístico
Sonda	209604_s_at	209604_s_at
Ac	0,8627	0,7647
Se	0,6153	0,1538
Es	0,8780	0,9024

equiparam e em alguns casos superam os resultados mostrados na literatura como, por exemplo, em [Hess et al. 2006], para um conjunto de 30 sondas, os resultados foram 76% de acurácia, 92% de sensibilidade e 71% de especificidade. Isso indica que a metodologia aqui proposta é apropriada para o problema e, se estendida, talvez obtenha resultados melhores do que os atualmente conhecidos.

O método de regressão logística apresenta resultados ruins para todas as abordagens utilizadas. A sua acurácia fica em torno de 75% e o maior falor de sensibilidade obtido foi de 46%.

5. Conclusões

O objetivo do presente trabalho era iniciar a investigação sobre modelagem dos dados de pacientes com câncer utilizando *kernel density estimation*. A partir dos resultados aqui apresentados iniciais acredita-se ser possível utilizar este método para a modelagem, uma vez que a partir das distribuições geradas é possível distiguir de qual tipo de classe (PCR ou NoPCR) ela pertence com razoável acurácia.

No futuro propomos a utilização de outros tipos de suavizadores de kernel (normal, epanechnikov, triangular), bem como outros métodos de cálculo do kernel. Outra possível linha de investigação é o uso de mais sondas, isoladamente ou em conjunto, para aumentar a acurácia, sensibilidade e especificidade.

Referências

Boyle, P. and Levin, B. (2008). World cancer report 2008. Technical report, International Agency for Research on Cancer, Lyon.

Tabela 4. Resultados dos dois métodos utilizados para o conjunto de 3 sondas.

Método	Classificador Bayesiano	Regressor Logístico
Ac	0,8823	0,7647
Se	0,6923	0,4615
Es	0,8780	0,8048

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience, 2 edition.

Hess, K., Anderson, K., Symmans, W., Valero, V., Ibrahim, N., Mejia, J., Booser, D., Theriault, R., Buzdar, A., Dempsey, P., Rouzier, R., Sneige, N., Ross, J., Vidaurre, T., Gomez, H., Hortobagyi, G., and Pusztai, L. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, 24(26):4236–4244.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley Series in Probability ans Statistics, 2 edition.

Natowicz, R., Braga, A. P., Incitti, R., Horta, E., Rouzier, R., Rodrigues, T. S., and Costa, M. (2008a). A new method of dna probes selection and its use with multi-objective neural networks for predicting the outcome of breast cancer preoperative chemotherapy. In *ESANN'2008 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning*, pages 71–76. d-side publi.

Natowicz, R., Incitti, R., Horta, E. G., Charles, B., Guinot, P., Yan, K., Coutant, C., Andre, F., Pusztai, L., and Rouzier, R. (2008b). Prediction of the outcome of preoperative chemotherapy in breast cancer by dna probes that convey information on both complete and non complete responses. *BMC Bioinformatics*, 9:149.

Natowicz, R., Incitti, R., Rouzier, R., Çela, A., Braga, Ant o., Horta, E., Rodrigues, T., and Costa, M. (2008c). Downsizing multigenic predictors of the response to preoperative chemotherapy in breast cancer. In *KES '08: Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part II*, pages 157–164, Berlin, Heidelberg. Springer-Verlag.

Silverman, B. (1986). Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*.

Thompson, J. R. and Tapia, R. A. (1990). *Nonparametric function estimation, modeling and simulation*. Ed. Siam - Society for Industrial and Applied Mathematics, 1a edition.