Verificação de Assinaturas Offline Baseada em Estatística Tipo Escore

Hélio G. de Souza Júnior, Abraão D. C. Nascimento, Manoel R. Sena Jr.
Departamento de Estatística, Universidade Federal de Pernambuco
Cidade Universitária, 50740-540 Recife – PE, Brasil
jr1souza@hotmail.com, abraao.susej@gmail.com, manoel@de.ufpe.br

30 de abril de 2010

Resumo

O uso de estatística escore em melhoramento de métodos de classificação de padrões tem recebido considerável atenção. A problemática nesta metodologia é a garantia da distribuição assintótica das estatísticas de teste utilizadas a partir de um certo tamanho de amostra. Nelson *et al.* [4] propuseram um método de classificação eficiente para diferentes tamanhos de amostra via estatística escore tendo a qui-quadrado como distribuição limite. Motivado pelo problema de verificação de assinaturas, este trabalho propõe um novo teste de hipótese baseado na distribuição beta. Além disso, nós utilizamos simulação de Monte Carlo para comparar o depempenho desta nova metodologia à dois métodos clássicos baseados nas distribuições qui-quadrado e T^2 de Hotelling.

1 Introdução

Devido ao aumento dos recursos de captação de identificação, o termo *biometria* tem sido tema de muitas discussões. Em classificação estatística de padrões, esta expressão se refere ao reconhecimento individual tendo por base as características de pessoas [3]. A assinatura manuscrita ocupa um lugar especial no amplo conjunto de métodos da biometria [1]. Este recurso pode ser entendido como o resultado de um complexo processo dependente do *estado psicofísico do assinante* e das *codições sob as quais os elementos da assinatura ocorrem*. No tocante aos métodos de aquisição, um importante sistema de verificação de assinatura (VA) é nominado por *estático ou offline* [3]. Conforme imagens da Figura 1, o sitema offline representa a assinatura como uma image de nível cinza $\{\delta(x,y)\}_{0\leq x\leq X,\, 0\leq y\leq Y}$, onde $\delta(x,y)$ denota o nível cinza na posição (x,y) da imagem.

Devido ao avanço conjuntamente com a internet, a VA offline tem exigido refinadas técnicas de modelagem. Neste campo, uma questão importante é o tratamento de uma nova assinatura a partir de um banco inicial. A partir do ponto de vista de reconhecimento estatístico de padrões, os parâmetros das distribuições são geralmente estimados a partir de vetores de amostra de treinamento. No entanto, é necessário que o número de observações seja suficiente para obtermos bons estimadores.

Seguindo esta linha de atuação, o estudo feito por Wilson (1978) [5] sugerem o uso de uma função proporcional da distância Mahalanobis como discriminante, quando as observações seguem uma distribuição normal multivariada; ou a utilização da regressão logística, quando a suposição de normalidade é violada. Nelson *et al.* (1994) [4] desenvolveram um método de classificação com base na pontuação estatística, que se aproximam da distribuição qui-quadrado, assumindo normalidade para o vetor de observações. Obtendo excelentes resultados, em termos de taxas de erros de classificação, dependendo do tamanho da amostra selecionada. Gnanadesikan e Kettenring (1972) [2] mostraram que a estatística $b(x_i) = \left(\frac{n}{(n-1)(n-1)}\right) d_i$ tem distribuição beta$\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$, onde d_i representa i-ésima distância amostral segundo Mahalanobis e x_i um vetor de observações. Sena-Jr. (1997) [6] mostrou que $F(x) = \frac{n-p}{p(n-1)} \frac{n}{n+1} d(x)$ tem distribuição F$(p, n-p)$, onde d representa a distância amostral de Mahalanobis da nova observação.

No presente trabalho, nós comparamos, tendo por base as taxas de rejeição, o desempenho de três métodos de classificação em diversos tamanhos de amostra via simulação de Monte Carlo.

2 Metodologia

Para esta modelagem de classificação, considere um vetor aleatório $x = \{x(1), x(2), \ldots, x(p)\}$ distribuído segundo uma normal p-variada pertencente a uma classe paramétrica $C_i \in \{(\mu_i, \Sigma_i) : -\infty < \mu_i < \infty, \Sigma_i \geq 0\}$ $(i = 1, 2)$. O objetivo é encontrar subespaços do suporte da normal p-variada $R_i \subset \mathbb{R}^p$ $(i = 1, 2)$ tais que sejam satisfeitas as condições: (i) $R_1 \cup R_2 = \mathbb{R}^p$ e (ii) $R_1 \cap R_2 = \emptyset$. Na prática, tanto μ_i quanto Σ_i são desconhecidos.

Assumindo que o vetor de características indexado a uma dada assinatura tenha distribuição normal multivariada com parâmetros μ e Σ. Nelson *et al.* [4] considera a substituição dos parâmetros pelos seus respectivos estimadores de maxima verossimilhança (MV), $\mu_{MV} = \overline{x}$ e $\hat{\Sigma}_{ML} = S$, resultando na estatística:

$$d(x) = (x - \overline{x})^T S^{-1} (x - \overline{x}), \tag{1}$$

onde x é o novo vetor de observação, $\overline{x}$ o vetor de média e S a matriz de covariância das observações que não incluem o vetor x. É possivel mostrar que "$d(x)$" tem distribuição qui-quadrado assintótica com p graus de liberdade. Além disso, devido centralidade assintotica dos estimadores MV, tem-se que: $(\overline{x}, S) \xrightarrow[n\to\infty]{} (\mu, \Sigma)$. Esta estatística pode ser bastante viesada quando o tamanho da amostra é pequeno.

Gnanadesikan e Kettenring [2] sugeriu uma estatística corrigida pelo tamanho da amostra. Esta metodologia resultou na distância estocástica (calssificador) dada por

$$b(x_i) = \frac{n}{(n-1)^2} d(x_i), \tag{2}$$

onde $d(x_i)$ representa i-ésima distância amostral de Mahalanobis. Neste, mostrou-se que $b(x)$ tem uma distribuição beta$(\frac{p}{2}, \frac{n-p-1}{2})$. Nos procedimentos seguintes nominaremos $b(x_i)$ por "$d(beta)$".

Sena-Jr [6] propôs uma modificação na estatística $d(x)$ baseada nos números de observações n e de características do vetor x p. Ficando definida desta forma:

$$F(x) = \frac{n-p}{p(n-1)} \frac{n}{n+1} d(x). \tag{3}$$

Esta estatistica, a qual nominaremos por "$d(T^2)$", distribui-se segundo a lei probabilística $F(p, n-p)$.

3 Simulação e Resultados

Nossos resultados baseados em simulação de Monte Carlo tem o objetivo de compara três metodologia de classificação estatísticas baseadas nas medidas estocásticas: $d(x)$ - Nelson *et al.* [4], d(beta) - Gnanadesikan e Kettenring [2], e d(t2) - Sena-Jr. [6]. Isto foi feito considerando como critério comparativo a taxa de rejeição ao nível nominal de $\alpha \in \{1\%, 5\%, 10\%\}$. Considerou-se dados gerados a partir de uma nomal multivariada onde o tamanho de amostra $n \in \{5, 10, 20, 50\}$ e a quantidade de características $p \in \{2, 4\}$.

Afim de obter os parâmetros bases para a geração de número aleatórios, considerou-se um banco de assinaturas resultante de uma pesquisa feita no Dept. de comunicações (DECOM) da faculdade de Eng. Eletrica e da Computação (FEEC) situada na Universidade estadual de Campinas (UNICAMP). Este banco contém 42 características (para descrição das variávies, ver Sena Jr. [6]) relativas a um autor em 1000 situações e 825 falsificações. A Figura 1 é uma ilustração das assinaturas, sendo cinco genuínas e cinco falsificadas. Com base numa análise descritiva deste banco, considerou-se as variáveis com menores coeficientes de variação para os diferentes valores de p fixados. A partir de então se seguiu as etapas abaixo:

a. considerando uma matrix H com dimensão $10 \times p$ resultante da amostragem do banco de assinaturas genuínas, obtem-se $\hat{\mu}_H = \bar{H}$ (media amostral de H) e $\hat{\Sigma}_H = S_H$ (variancia amostral de H);

b. a fim de se definir uma *amostra de treinamento*, gera-se 100 realizações de uma normal mulivariada com parâmetros $\hat{\mu}_H$ e $\hat{\Sigma}_H$, resultando numa matrix C com dimensão $100 \times p$, e se considerou as seguintes estimativas $\hat{\mu}_C = \bar{C}$ e $\hat{\Sigma}_C = S_C$;

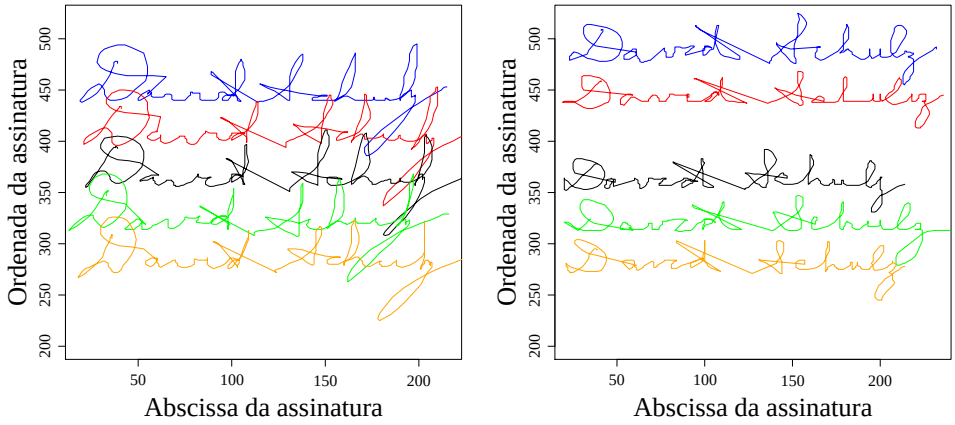
c. por fim, considerando a *amostra de teste*, gera-se 1000 realizações de uma normal mulivariada com parâmetros $\hat{\mu}_H$ e $\hat{\Sigma}_H$, $x = \{x_1, x_2, \ldots, x_{1000}\}$. E, tomando $d(x) = (x - \hat{\mu}_C)^T \hat{\Sigma}_C^{-1}(x - \hat{\mu}_C)$ nas fórmulas(1)-(3), obteve-se

$$\begin{aligned} \alpha_d &= \Pr\left(d(\mathbf{x}) \geq \chi^2_{(p,\alpha)}\right), \\ \alpha_{d(Beta)} &= \Pr\left(d_{Beta}(\mathbf{x}) \geq Beta\left(\frac{p}{2}, \frac{n-p-1}{2}, \alpha\right)\right), \\ \alpha_{d(T^2)} &= \Pr\left(d_{T^2}(\mathbf{x}) \geq F(p, n-p, \alpha)\right). \end{aligned}$$

Estes passos são repetidos em 100 replicações de Monte Carlo e se obtiveram as estimativas destas taxas, a saber $\widehat{\alpha}_d$, $\widehat{\alpha}_{d(Beta)}$ e $\widehat{\alpha}_{d(T^2)}$. A Tablea 1 e a Figura 2 apresenta os resultados de simulação. Nas Figuras podemos visualizar melhor o quanto as taxas de rejeição dos testes aumentam com o tamanho de amostra.

4 Conclusão

Embora bastante usual na prática, verificamos que o teste baseado na estatística d é o menos recomendável. Suas taxas de rejeição são distantes do nível nominal de significância pré-



(a) Imagens Verdadeiras (b) Imagens Falsas

Figura 1: Photo and HH polarization of San Francisco image.

Tabela 1: Porcentagem média das observações que excerderam o valor crítico.

	p=2 e n=5			p=4 e n=10		
α	$\widehat{\alpha}_d$	$\widehat{\alpha}_{d(Beta)}$	$\widehat{\alpha}_{d(T^2)}$	$\widehat{\alpha}_d$	$\widehat{\alpha}_{d(Beta)}$	$\widehat{\alpha}_{d(T^2)}$
1	0.00	0.48	0.40	0.00	0.9 0	0.48
5	0.00	2.63	7.89	0.00	2.12	3.14
10	0.00	13.73	15.82	0.03	11.60	10.36
	p=2 e n=10			p=4 e n=20		
1	0.00	0.77	0.67	0.00	0.96	0.95
5	0.23	2.61	3.13	2.28	5.90	5.95
10	4.59	8.52	7.20	6.68	10.56	10.55
	p=2 e n=20			p=4 e n=50		
1	0.21	1.03	1.13	0.81	1.63	1.68
5	3.56	5.25	5.13	3.73	5.07	5.06
10	6.68	8.61	8.92	8.85	10.20	10.27

fixado para os dois números de características do vetor de assinatura. Como esperado assintoticamente, os resultados são optimalizados (isto é, estão mais próximos dos níveis nominais) quando aumentamos os fatores de simulação, n e p. Por outro lado, os testes com estatísticas $d(beta)$ e $d(T^2)$ apresentam os melhores comportamentos para todos os diferentes pares (n, p). Eles apresentam taxas de rejeição semelhantes quando aumentamos o tamanho da amostra nas simulações.

Referências

[1] M.C. Fairhurst. Signature verification revisited: promoting practical exploitation of biometric technology. *Electronics & Communications Engineering Journal*, 9(6):273–280, 1997.

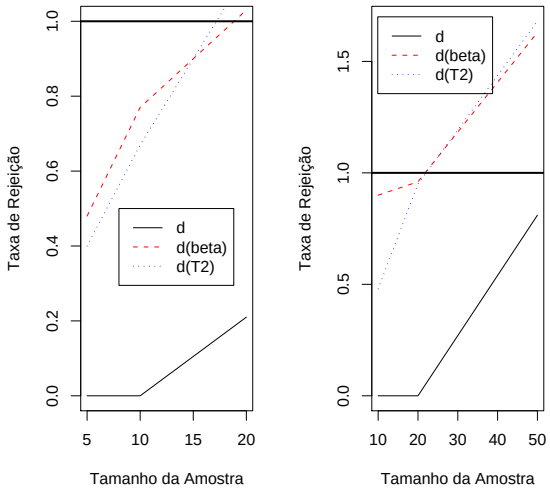


Figura 2: Taxas de rejeição de testes baseados nas estatísticas d,$d(beta)$ e $d(T^2)$, para $p = 2$ (à esquerda) e $p = 4$ (à direita) com $\alpha = 0.01$.

[2] R. Gnanadesikan and J.R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28:81–124, 1972.

[3] D. Impedovo and G. Pirlo. Automatic signature verification: The state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C - Applications and Reviews*, 38(5):609–635, 2008.

[4] W.L. Nelson, W.L. Turin, and T. Hastie. Statistical methods for on-line signature verification. *Journal of Pattern Recognition and Artificial Inteligence*, 8(3):749–770, 1994.

[5] S. J. Press and S. L. Wilson. Choosing between regression and discriminant analysis. *Journal of the American Statistic Association*, 73:699–705, 1978.

[6] M. R. Sena Jr. *Aplicações Estatísticas em Reconhecimento de Padrões em Ênfase em Verificação de Assinaturas*. PhD thesis, University Federal of Campinas, 1997.