XXIII CONIC
VII CONITI
IV ENIC

CLASSIFICAÇÃO DE PADRÃO ENVOLVENDO SIMULAÇÃO PARA PADRÕES NÃO OBSERVADOS

João Eudes Miquéias Maciel Torres[1]; Manoel Raimundo de Sena Júnior[2]

[1]Estudante do Curso de Estatística - CCEN – UFPE; E-mail: jemmt1@de.ufpe.br,
[2]Docente/pesquisador do Depto de Estatística – CCEN – UFPE. E-mail: manoel@de.ufpe.br.

Sumário: Problemas que envolvem classifição de padrões vem sendo amplamente abordados por diferentes técnicas. A utilização de distâncias estocásticas na construção de um limiar de classificação, valendo-se da distância amostral proposta por Mahalanobis, é especialmente viável sobre suposição de normalidade das observações. A idéia é associar um pequena área da distribuição a priori quando essa distância for grande. A técnica da Razão de Verossimilhanças é usada na classificação entre duas populações conhecidas e não pode ser aplicada quando não se tem acesso à duas populações distintas, ou o acesso a uma delas é restrito. Através da decomposição espectral da matriz de covariâncias pode-se estabelecer a transformação dos dados afim de simular uma distribuição com características similares a original, com mesma média, mas com matriz de covariâncias diferente. O problema da verificação de assinaturas *offline* pode ser assim pensado, queremos classificar uma nova assinatura com genuína ou feita por falsificador, quando temos acesso apenas à assinaturas verdadeiras.

Palavras–chave: classificação de padrões; razão de verossimilhanças; verificação de assinaturas

INTRODUÇÃO

Em alguns cenários práticos, estamos interessados em verificar se uma nova observação pertence a uma classe ou população já conhecida. O teste da Razão de verossimilhanças (RV) é um método amplamente utilizado em problemas desse tipo envolvendo duas populações. Entretanto, nem sempre temos acesso às duas populações, ou o acesso a uma delas é restrito, para a construção de uma função discriminante. Nesse contexto, a distância de Mahalanobis tem sido usada como referência, por ser a medida mais eficiente em termos de estabelecer um quantil da distribuição qui-quadrado, pois servirá como limiar de classificação, quando utilizamos apenas uma população, Mardia *et al*. (1979).

Quanto a classificação, deseja-se minimizar o erro cometido ao classificar uma observação na classe a qual não pertence (erro do tipo II) e ao não classificar, quando, de fato, a observação pertence a classe (erro do tipo I).

As técnicas de classificação que utilizam a Distância de Mahalanobis amostral (DM) partem do princípio que as observações seguem uma distribuição normal p-variada, e associam uma pequena área da distribuição aos pontos com alta distância amostral. Na prática os parâmetros da distribuição condicional de **x** pertencer a *i*-ésima classe são desconhecidos. Nelson *et al.* (1994) propuseram substituir os parâmetros pelos seus repectivos estimadores de máxima verossimilhança, supondo que a distribuição do vetor ***x*** é Normal *p*-variada, definindo assim a estatística ***d(x₀)*** denotada por: $d(x_0) = (x_0 \quad \bar{x}_i)' S_i^{-1} (x_0 \quad \bar{x}_i)$, onde x_0 é representa uma nova observação e $\bar{x}_i$ e S_i são, respectivamente, os estimadores de máxima verossimilhança de μ_i e Σ_i .

Sena Jr. (1997) mostrou que $F(x_0) = \left(\frac{p \quad n}{p(n \quad 1)} \frac{n}{n+1} \right) d_0$ tem distribuição $F(p, n \quad p)$, onde d_0 também representa o quadrado da distância amostral de Mahalonobis de x_0 ao centro da distribuição. Essa distância aqui denominou-se Distância de Mahalanobis Corrigida (DMC).

Em situações onde o centro de massa das populações são próximos geometricamente, as técnicas de usam distâncias estocásticas são comprometidas pela falsa associação, erro do tipo II. Esse trabalho propõe-se utilizar a decomposição espectral da matriz de covariâncias, para estabelecer a transformação dos dados afim de simular a população que não se tem acesso, com características similares a original. Em seguida estimaremos os parâmetros das duas populações para usar a RV na classificação e, através dos erros observados, comparar com método distância de Mahalanobis. Uma aplicação no problema de verificações de assinaturas foi explorada, imaginando que o falsificador consiga imitar a média, mas não a variância, de uma assinatura, Sena Jr. (1997).

MATERIAIS E MÉTODOS

Considere **X** a representação matricial de uma população de *m* indivíduos com *q* características de interresse, em que a i-ésima linha de segue uma distribuição normal *q*-variada. A matriz de covariâncias de **X** pode ser decomposta como:

$$\Sigma = PDP'$$

onde **P** é a matriz de autovetores e **D** é a matriz diagonal de autovalores.

A transformação foi feita multiplicando o maior autovalor de Σ por uma constante *c*, maior que 1, obtendo uma nova matriz $D^r = diag(c\lambda_1, \lambda_2, ..., \lambda_q)$, onde λ_i é o i-ésimo autovalor de Σ, $i = 1,...,q$. Substituição **D** na decomposição acima por D^r, gera-se uma outra matriz de covariâncias Σ_r. As *m* observações que irão compor a população simulada, denotada pela matriz **Y**(*m* x *q*), são obtidas pela seguinte transformação:

$$y_i = \Sigma_r^{1/2} z_i + \mu, i = 1,...,m$$

onde z_i é um vetor normal padão *q*-variado.

Como a combinação linear de uma variável normalmente distribuida também segue uma distribuição normal, a matriz **Y** é tal que a i-ésima linha representa uma observação normalmente distribuida com média $E(y_i) = \mu$ e matriz de covariâncias $Var(y_i) = \Sigma_i$.

De posse de duas populações seguindo a mesma distribuição de prababilidade, pode-se utilizar a RV na classificação. Decide-se por associar uma nova observação a classe que obter maior verossimilhança, Casella & Berger (2010). Com a suposição de normalidade a razão de verossimilhanças fica definida como:

$$\lambda(x_0) = \frac{L(x_0, \mu, \Sigma)}{L(x_0, \mu, \Sigma_r)} = \frac{|\Sigma|^{1/2} \exp\{(x_0 \quad \mu)' \Sigma^{1} (x_0 \quad \mu)\}}{|\Sigma_r|^{1/2} \exp\{(x_0 \quad \mu)' \Sigma_r^{1} (x_0 \quad \mu)\}}$$

Aplicando logaritmo natural e desenvolvendo essa razão tem-se o seguinte resultado:

$$T(x_0) = \quad 2\ln(\lambda(x_0)) = \ln\left(\frac{|\Sigma|}{|\Sigma_r|} \right) + (x_0 \quad \mu)' (\Sigma^{1} \quad \Sigma_r^{1}) (x_0 \quad \mu)$$

Na prática os parâmetros μ e Σ são desconhecidos. Dessa forma, é razoável substituir esse parâmetros por seus respectivos estimadores de máxima verossimilhança $\bar{x} = n^{1} X' 1$ e $S = n^{1} X' HX$, em que $1' = [1,...,1]$ e $H = I \quad n^{1} 11'$. Assim, decide-se por classificar um novo objeto como pertencente a população **X** se $T(x_0) > 0$.

RESULTADOS

O quadro 1 mostra o resultado das taxas dos erros obtidas via simulação no *software* estatístico *R*. Gerou-se uma matriz com 1000 linhas e 50 colunas, cada linha seguindo uma distribuição normal multivariada. Foi extraído um número *n ={20, 30, 50, 200}* de linhas para compor a amostra de treino e foi selecionada as p = {5,10,15} colunas com menores coeficiente de variação. Em seguida, obteve-se dessa amostra com *p* colunas, as estimativas de $\bar{x}$ e **S**, a qual foi aplicada a transformação, com *c= {2, 5}*, para obter S_r e, com esse parâmetro, gerou-se outra matriz com 1000 linhas e *p* colunas, representando observações falsas.

Quadro 1:Erros simulados para os métodos RV e DMC com diferentes valores de *n, p* e *c.*

		c = 2				c =5			
		RV		DMC		RV		DMC	
Tipo de erro		Tipo I	Tipo II	Tipo I	Tipo II	Tipo I	Tipo II	Tipo I	Tipo II
p = 5	n = 20	0.184	0.578	0.14	0.893	0.113	0.468	0.14	0.741
	n = 30	0.245	0.56	0.109	0.907	0.161	0.459	0.109	0.729
	n = 50	0.193	0.561	0.085	0.891	0.117	0.455	0.085	0.723
	n = 200	0.224	0.597	0.071	0.906	0.144	0.461	0.071	0.738
p = 10	n = 20	0.091	0.574	0.546	0.914	0.038	0.436	0.546	0.789
	n = 30	0.147	0.588	0.346	0.931	0.092	0.474	0.346	0.792
	n = 50	0.171	0.597	0.142	0.913	0.1	0.468	0.142	0.795
	n = 200	0.224	0.617	0.081	0.924	0.145	0.494	0.081	0.807
p = 15	n = 20	0.059	0.602	0.957	0.926	0.024	0.472	0.957	0.826
	n = 30	0.1	0.562	0.561	0.927	0.054	0.449	0.561	0.81
	n = 50	0.129	0.58	0.292	0.918	0.066	0.472	0.292	0.787
	n = 200	0.214	0.82	0.083	0.924	0.137	0.448	0.083	0.829

O quadro 2 apresenta as taxas dos erros para assinaturas. Os dados para análise dividiam-se em dois grupos, um grupo composto por 1000 assinaturas verdadeiras e outro composto por 825 assinaturas falsas, com 42 características de interesse. Esse conjunto de dados foi fornecido pelo Departamento de Engenharia Elétrica e Computação da Universidade Estadual de Campinas (UNICAMP). Á esses dados aplicou-se o mesmo precedimento realizado na simulação com a adição de um valor para *c*.

Quadro 2: Taxas dos erros observados na verificação de assinaturas pelos metódos RV e DMC, para diferentes valores de *n, p* e *c*.

		c = 2		c = 5		c = 10			
		RV		RV		RV		DMC	
Tipos erros		Tipo I	Tipo II	Tipo I	Tipo II	Tipo I	Tipo II	Tipo I	Tipo II
p = 5	n = 20	0.43	0.0	0.348	0.0	0.282	0.0	0.748	0.0
	n = 30	0.634	0.0	0.541	0.0	0.482	0.0	0.635	0.0
	n = 50	0.362	0.048	0.277	0.053	0.232	0.056	0.459	0.0

p = 10	n = 200	0.352	0.0	0.262	0.0	0.212	0.0	0.341	0.0
	n = 20	0.512	0.0	0.408	0.0	0.339	0.0	0.902	0.0
	n = 30	0.545	0.002	0.439	0.004	0.378	0.004	0.801	0.0
	n = 50	0.592	0.0	0.506	0.0	0.435	0.0	0.533	0.0
	n = 200	0.254	0.0	0.161	0.0	0.114	0.0	0.311	0.0
p = 15	n = 20	0.5	0.0	0.406	0.0	0.333	0.0	0.977	0.0
	n = 30	0.541	0.0	0.447	0.0	0.384	0.0	0.879	0.0
	n = 50	0.554	0.0	0.468	0.0	0.405	0.0	0.637	0.0
	n = 200	0.455	0.0	0.38	0.0	0.309	0.0	0.285	0.0

DISCUSSÃO

Com resultado da simulação confirmou-se que o método de distância estocástica DMC classifica como "verdadeiras" pseudo-observações com grande frequência, isso deve-se a suposição que o centro de massa das populações são iguias. Porém, o metodo RV não teve um resultado satisfatório, também apresentando altas taxas de erros.

Na classificação das assinaturas, o erro do tipo II assume valores baixos ou nulos devido o fato de que os centros de massa dos conjuntos de assinaturas verdadeiras e falsas serem geometricamente distantes. Para o erro tipo I, nota-se que apenas para tamanhos de amostras grande, o DMC tem melhores resultados, enquanto o RV tem resultados razoáveis para pequenas amostras, especialmente para $c = 10$. É importante salientar que, mesmo tendo melhor desempenho para amostras pequenas, o RV apresenta taxas de erros muito elevadas no contexto de verificação de assinaturas.

CONCLUSÕES

O método RV proposto mostra-se melhor em relação a técnica de distância estocástica DMC quando a amostra de treino é pequena, mas ainda necessita de ajustes. No cenário onde as médias das populações são próximas geometricamente, o RV é capaz de diminuir a falsa aceitação de observações, erro do tipo II. No entanto, as taxas dos erros ainda são altas, abrindo espaço para aprimoramentos da técnica. Também pode ser realizado um estudo acerca do valor da constante multiplicativa c que minimiza os erros.

AGRADECIMENTOS

Ao professor orientador, Dr. Manoel R. de Sena Jr. pela orientação, apoio e paciência. Á PROPESQ pelo apoio financeiro.

REFERÊNCIAS

M. R. Sena Jr., A. D. C. Nascimento, G. M. Cordeiro & L. P. Barroso - *Score-type statistics in pattern classification*. Brazilian Journal of Probability and Statistics, 2013.

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

M. R. Sena Jr. & L. L. Ling - *Verificação de assinaturas Baseada em Intervalos de Confiança*, SBT 95 - Simpósio Brasileiro de Telecomunicações, Vol. I, pp 269-274, 1995

K. V. Mardia, J. T. Kent & J. M. Bibby - *Multivariate Analysis*, Academic Press, London, 1979.

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

G. Casella & R. L. Berger - *Inferência Estatística*. Tradução Solange A. Visconte. São Paulo: Cengage Learning, 2010.