Anais do XII Encontro Mineiro de Estatística - MGEST 2013. Uberlândia - 05 e 06 de setembro de 2013.
Revista Matemática e Estatística em Foco - ISSN:2318-0552

Classificação de dados em modelos com resposta binária via algoritmo boosting e regressão logística

Gilberto Rodrigues Liska [1] [5]
Fortunato Silva de Menezes [2] [5]
Marcelo Ângelo Cirillo [3] [5]
Mario Javier Ferrua Vivanco [4] [5]

Resumo: Classificar algo é uma tarefa natural do ser humano, mas existem situações em que o mesmo não é o mais indicado para desempenhar tal função, mostrando, portanto, a necessidade de métodos automáticos de classificação. Devido a importância e aumento da complexidade de problemas do tipo, existe ainda a necessidade de métodos que forneçam maior precisão e interpretabilidade dos resultados. Entre eles os métodos de Boosting. Atualmente os modelos de regressão logística com seus parâmetros estimados via máxima verossimilhança (MRLMV) são muito utilizados para esse tipo de situação. Nesse sentido, o presente trabalho consistiu em comparar o modelo MRLMV e o modelo de regressão logística estimado via algoritmo Boosting, mais especificamente algoritmo Binomial Boosting (MRLBB), e selecionar o modelo com melhor ajuste/desempenho na situação de presença/ausência de doença cardíaca coronariana (CHD) em pacientes. O MRLBB apresentou melhores valores de AIC, BIC, AUC, sensibilidade, especificidade, acurácia, taxa de falsos positivos e falsos negativos e pelo teste de Hosmer-Lemeshow esse modelo não apresenta evidências de mal ajuste. Diante dos resultados obtidos, o MRLBB é o mais adequado para descrever o problema de presença/ausência de doença cardíaca coronariana em pacientes, pois fornece informações mais precisas acerca do problema exposto.

Palavras-chave: *Algoritmo Boosting, Critério de Informação de Akaike (AIC), Regressão Logística, seleção de modelos*.

1 Introdução

Em inúmeras situações o pesquisador se depara com a necessidade de realizar uma classificação nos dados, sobretudo, mediante ao tamanho amostral a ser considerado, bem como outras causas, as

[1]DEX - Universidade Federal de Lavras. Email: *gilbertoliska@hotmail.com*
[2]DEX - Universidade Federal de Lavras. Email: *fmenezes@dex.ufla.br*
[3]DEX - Universidade Federal de Lavras. Email: *marcelocirillo@gmail.com*
[4]DEX - Universidade Federal de Lavras. Email: *ferrua@dex.ufla.br*
[5]Agradecimento à FAPEMIG pelo apoio financeiro.

quais o modelo proposto ou os dados apresentarem algum tipo de perturbação, os métodos estatísticos convencionais poderão apresentar taxas de erros de classificação incoerentes.

Convém ressaltar que a técnica estatística da Regressão Logística, que é utilizada em situações que envolvam classificação, a resposta a um determinado fenômeno não configura uma situação contínua, ou seja, admite-se a existência de categorias, podendo assumir dois ou mais valores. Nestes casos a Regressão Logística, cuja estimação de parâmetros é feita via máxima verossimilhança, tem sido aplicado com frequência e sua utilização permite obter a probabilidade de um determinado evento ocorrer, estimada por meio de um modelo logístico. Contudo, a Regressão Logística, a priori pressupõem da criação de regras bastante interpretáveis, mas com formas restritivas para a relação entre as respostas e as variáveis preditoras.

Mediante a conjectura de aprimorar a interpretabilidade e desempenho do uso de métodos classificadores aplicados em uma variedade de problemas, tem-se os algoritmos de Boosting, originados na área de computação, funcionam aplicando-se sequencialmente um algoritmo de classificação a versões reponderadas do conjunto de dados de treinamento, dando maior peso às observações classificadas erroneamente no passo anterior. Eles foram introduzidos por [7] e desde então, várias versões de algoritmos Boosting têm sido criadas. Recentemente, [3] mostrou que Boosting pode ainda ser visto como um método para estimação funcional e pode ser utilizado para estimar um modelo de regressão logística.

Diante do exposto, esse trabalho objetiva estudar o desempenho de algoritmo Boosting , mais especificamente o algoritmo Binomial Boosting (MRLBB), em problemas de classificação que envolvam respostas binárias em comparação com o modelo de regressão logística estimada pelo método da máxima verossimilhança (MRLMV), a fim de fornecer o melhor modelo no sentido de ajuste/discriminação em situações binárias.

2 Metodologia

2.1 Dados

Foram utilizados os dados disponibilizados por *UCI Machine Learning Repository* [2]. Os dados são referentes a 270 pacientes com presença ou não de doença coronariana cardíaca (Coronary Heart Disease - CHD) e essa condição está em função de 13 variáveis independentes. Uma descrição completa dessas variáveis independentes pode ser vista em [2].

2.2 Ajuste dos Modelos de Regressão Logística

Para estimar os parâmetros do modelo de regressão logística via algoritmo Boosting será utilizado o algoritmo Binomial Boosting (MRLBB), proposto por [3]. Para executar o algoritmo Binomial Boosting

é necessário que sejam definidas duas componentes, sendo uma função perda e um procedimento base. O algoritmo Binomial Boosting utiliza a função perda binomial e o procedimento base mínimos quadrados componente a componente, uma vez que a resposta DIS configura uma situação binária e estamos interessados em ajustar um modelo linear generalizado.

O algoritmo Binomial Boosting, durante o processo de estimação paramétrica, já realiza seleção de variáveis, retornando, portanto, aquelas variáveis independentes que minimizam a função perda utilizada, levando ao modelo com as variáveis independentes que contribuem significativamente no modelo. O procedimento é descrito em [1].

Para estimar os parâmetros do modelo de regressão logística via máxima verossimilhança (MRLMV) foi utilizado o método como descrito [5]. Em seguida foi utilizado o método *stepwise* de seleção de variáveis via AIC, com o objetivo de eliminar as variáveis independentes que não contribuem de forma significativa para a probabilidade de ocorrência de doença cardíaca coronariana em pacientes.

2.3 Comparação dos Modelos MRLBB e MRLMV

Para avaliar o desempenho dos modelos obtidos pelos dois métodos, o conjunto de dados foi separado em duas partes, sendo uma parte de treinamento, que será destinada à estimação os parâmetros dos modelos MRLBB e MRLMV, e a parte de teste, que será destinada à validação dos mesmos. O conjunto de treinamento corresponderá a 70% da amostra original. O complementar das partições constituirá o conjunto de teste. A validação será feita comparando-se os critérios de informação de Akaike (AIC) e Bayesiano (BIC) dos modelos obtidos após processo de seleção de variáveis e o modelo preferido será aquele cujos critérios são menores. Será utilizado o Teste de Hosmer-Lemeshow [5] para verificar a existência de problemas de ajuste dos modelos ajustados.

Para determinar o limiar adequado a fim de classificar um paciente quanto à presença ou não de CHD, será utilizada a curva ROC, ([4]) em ambos modelos MRLBB e MRLMV. Em seguida, para os modelos ajustados com a partição ideal, serão calculados a sensibilidade, especificidade, acurácia, taxa de falsos negativos, taxa de falsos positivos e AUC. Será escolhido o modelo que apresentar os melhores valores para essas quantidades.

Finalizando a metodologia proposta, para obtenção dos resultados serão utilizados os pacotes estatísticos *mboost*, *ROCR* e *MKmisc* do Sistema Computacional Estatístico R [6], para realização das análises.

3 Resultados e Discussão

Os resultados de AIC, BIC e teste de Hosmer-Lemeshow para os modelos MRLBB e MRLMV foram os seguintes: para o MRLBB o AIC foi de 144,1786, o BIC foi de 160,4283 e o teste de Hosmer-Lemeshow foi não significativo ($valor-p = 0,1596$); para o MRLMV o AIC foi de 154,4300, o BIC foi

Tabela 1: Estimativas dos parâmetros referentes ao modelo logístico ajustado aos dados sobre doença coronariana.

Variável	Parâmetro	MRLBB	MRLMV	
		Estimativa	Estimativa	Erro padrão
Constante	0	-4,6268	-10,7509	2,5400
SEX	22	0,7979	1,5066	0,5991
PAIN	33	-	2,0030	1,0524
	34	1,5284	3,9293	1,0390
PRESS	4	0,0107	0,0336	0,0131
COL	5	0,0028	-	-
ELE	73	0,2730	-	-
HEART	8	-0,0053	-	-
EXE	92	0,5502	-	-
ST	10	0,3762	-	-
SLOPE	112	0,6459	1,7693	0,4991
	113	-	1,9773	1,0305
VES	12	0,6967	1,0290	0,3270
THAL7	133	1,1746	1,4205	0,4984

de 180,7688 e o teste de Hosmer-Lemeshow foi não significativo ($valorp = 0,0506$). Como os valores de AIC e BIC do modelo MRLBB são menores do que os do modelo MRLMV, o modelo MRLBB é preferível. A tabela 1 apresenta as estimativas dos parâmetros dos modelos MRLBB e MRLMV.

Uma vez obtido o modelo para explicar a ocorrência de CHD, pode-se verificar o poder de discriminação desse modelo, ou seja, a capacidade do modelo em classificar corretamente indivíduos que tem CHD e os que não tem. A figura 3 mostra a curva ROC do modelo MRLBB e MRLMV e observa-se que os dois modelos apresentam alto poder de discriminação, uma vez que a área abaixo de cada curva ROC é de $0,947u.a.$ e $0,905u.a.$, respectivamente.

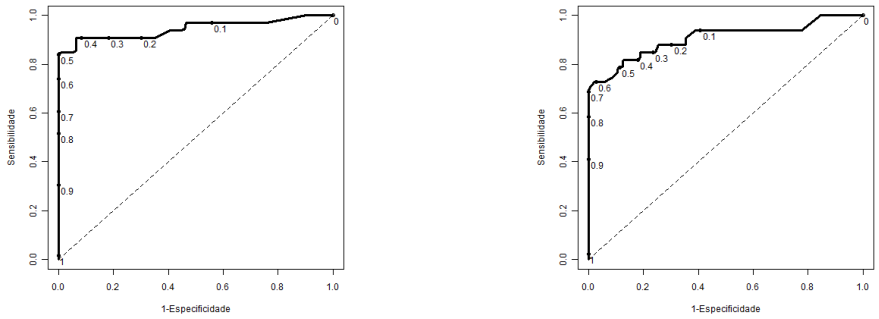

Figura 1: Curva ROC do modelo MRLBB (esquerda) e do modelo MRLMV (direita).

Uma outra vantagem da curva ROC é a possibilidade de escolher um limiar adequado para a classificação de pacientes quanto a presença ou não de CHD. Pela figura 3 um limiar adequado seria $0,5$ em ambos modelos. Portanto, para avaliar a *sensibilidade* e *especificidade* do modelo, será utilizado o seguinte critério para classificar um paciente como positivo para presença de CHD ($Y = 1$): se a probabilidade de ocorrência de CHD for maior do que 0,5 (50%). Caso contrário será classificado como ausente para CHD ($Y = 0$).

Observa-se que a *sensibilidade* do modelo MRLBB é de 82%, ou seja, 82% dos pacientes que tem

CHD o modelo os classificaram como positivo para essa característica. A taxa de falsos negativos do modelo foi de 18%, ou seja, 18% das pessoas que tem CHD o modelo acusou como falso para essa característica. A taxa de falsos positivos foi de 0%, logo, dos pacientes que não tem CHD o modelo não classificou nenhum paciente como positivo para CHD e, como consequência, a *especificidade* do modelo foi de 100%. A acurácia do modelo foi de 92,59% (Tabela 2).

Tabela 2: Tabela de confusão dos modelos MRLBB e MRLMV ajustado aos dados sobre doença arterial coronariana.

Observado	**Modelo**				**Acurácia**	
	presença		**ausência**			
	MRLBB	**MRLMV**	**MRLBB**	**MRLMV**	**MRLBB**	**MRLMV**
presença	81.82%	78.79%	18.18%	21.21%	92.59%	85.18%
ausência	0.00%	10.42%	100.00%	89.58%		

De maneira análoga, observa-se que a *sensibilidade* do modelo MRLMV foi de 79%. A taxa de falsos negativos e de falsos positivos do modelo foram 21% e 10%, respectivamente. A *especificidade* do modelo foi de 90%, ou seja, dos pacientes que não tem CHD, 90% foram classificados nessa condição. A acurácia do modelo foi de 85,18%. (Tabela 2).

4 Conclusões

Os modelos de regressão logística estimados via algoritmo Binomial Boosting (MRLBB) e pelo método da máxima verossimilhança (MRLMV) apresentaram ajuste satisfatório ao problema presença/ausência de doença cardíaca coronariana (CHD). O método de Boosting, mais especificamente o algoritmo Binomial Boosting, ajustou um modelo com melhor adequabilidade na situação presença/ausência de CHD, uma vez que a acurácia, sensibilidade, especificidade, taxa de falsos positivos e taxa de falsos negativos desse modelo foram melhores. O algoritmo Binomial Boosting constitui-se, portanto, numa alternativa poderosa para a análise de situações cuja resposta é binária.

Referências

[1] BUHLMANN, P.; HOTHORN, T. Boosting Algorithms: Regularization, Prediction and Model Fitting, *Statistical Science, v. 22, n. 4, p. 477-505, 2007.*

[2] FRANK, A.; ASUNCION, A. Machine Learning Repository, Irvine, CA: University of California, *School of Information and Computer Science, [http://archive.ics.uci.edu/ml], 2010.*

[3] FRIEDMAN, J. Greedy function aproximation: A gradient boosting machine, *The Annals of Statistics, v. 29, p. 1189 - 1232, 2001.*

[4] HANLEY, J. A. Receiver operating characteristic (ROC) methodology: the state of the art, *Critical Reviews in Diagnostic Imaging, v. 29(3), p. 307 - 335, 1989.*

[5] HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression, 2^o ed., John Wiley, New York, 1989.*

[6] R DEVELOPMENT CORE TEAM, *An Introduction to R: Version: 2.15 (2013)*. In: <http://www.r-project.org>. Acesso em 20 jun de 2013.

[7] SCHAPIRE, R. E. The strength of weak learnability, *Machine learning, v. 5, p. 197-227, 1990.*