Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Alocação de pacientes em grupos via Boosting: uma comparação com métodos tradicionais de classificação

Érica Aparecida Pereira[1]

Tiago Martins Pereira[2]

1 Introdução

Problemas de classificação são rotineiros em nosso dia-a-dia. Desde a infância somos capazes de reconhecer dígitos e letras. Caracteres pequenos ou grandes, manuscritos ou impressos, escritos de maneira desordenada ou até mesmo parcialmente ocultos — todos são facilmente reconhecidos. Esta habilidade de reconhecer ou classificar padrões faz com que nós humanos sejamos os melhores classificadores existentes, ainda que não conheçamos totalmente o mecanismo através do qual reconhecemos padrões.

Segundo Ferreira (2007), um procedimento de classificação ou simplesmente classificador é algum método formal capaz de decidir, com base em informações fornecidas, a que grupo ou população um determinado objeto pertence. De uma maneira geral, estamos interessados em construir, com base em uma amostra de dados, uma regra de classificação e utilizá-la para classificar novos objetos.

A classificação de dados tem se tornado uma tarefa difícil devido à grande quantidade de informação disponível. Ferreira (2007) ainda lembra, que o mercado exige cada vez mais a precisão dos classificadores, a qual é necessária em várias áreas de estudo, por exemplo, na medicina, em diagnóstico de doenças através de raio X, em bancos, como tomada de decisões para crédito, nas área de entretenimento, como na classificação de gêneros musicais, tornando-se essencial para muitas atividades na vida cotidiana, entre outras.

Os métodos de classificação (ou discriminação) mais conhecidos e mais utilizados em estatística aplicada são, provavelmente, regressão logística (HOSMER e LEMESHOW, 1989) e análise discriminante (JOHNSON e WICHERN. 2002). Esses métodos, cujos fundamentos estão estabelecidos na teoria estatística há bastante tempo, são lineares, têm baixo custo computacional, e em geral produzem resultados razoáveis, apesar de não raro serem utilizados em situações que contradizem suas suposições paramétricas distribucionais. As propriedades estatísticas destes métodos são conhecidas há muito tempo, e suas interpretações são claras.

[1] DEEST – UFOP. e-mail: pereira.a.erica@gmail.com
[2] DEEST – UFOP. e-mail: tiago.martin@iceb.ufop.br

De acordo com Loesch e Hoeltgebaum (2012), a análise discriminante baseia-se na informação das variáveis independentes para se alcançar a separação ou discriminação mais clara possível entre dois ou mais grupos. A função discriminante pode ser linear ou quadrática. Para o caso da função linear assume-se que as matrizes de covariâncias são iguais, suposição de homocedasticidade. No caso dos dados não serem homocedasticos temos a função discriminante quadrática.

Um método de classificação moderno que têm apresentado bons resultados é o método conhecido como Boosting. O Algoritmo Boosting funciona como um conjunto de vários classificadores simples que gera um classificador forte, e a cada novo classificador gerado são atribuídos pesos às observações. Esses pesos são alterados a cada novo classificador gerado, onde aumenta-se o peso da observação erroneamente classificada e diminui-se o peso da observação classificada corretamente, fazendo com que um classificador atue no erro do anterior.

O presente trabalho visa comparar três técnicas de classificação, Análise Discriminante Linear e Quadrática, métodos muito conhecidas no contexto da estatística multivariada, e o Algoritmo Boosting, método de classificação com um alto potencial, flexibilidade e simplicidade para ser implementado em diferentes cenários. Para efeitos de comparação, os classificadores foram aplicado a uma base de dados real, caracterizado por seis atributos biométricos usados para classificar pacientes ortopédicos como normais ou anormais.

2 Metodologia

2.1 Dados

Foram utilizados os dados disponibilizados por UCI Machine Learning Repository (BACHE e LICHMAN, 2013). Os dados são referentes a 294 pacientes com presença ou não de doença da coluna vertebral e essa condição está em função de seis variáveis independentes. Uma descrição completa dessas variáveis independentes pode ser vista em (BACHE e LICHMAN, 2013).

2.2 Análise Discriminante

A Análise Discriminante é uma técnica estatística multivariada que atua na identificação das variáveis que pertencem a cada grupo e quantas dessas são necessárias para

obter a melhor classificação. Tem como característica básica a utilização de um conjunto de informações obtidas acerca de variáveis consideradas independentes para conseguir um valor de uma variável dependente que possibilite a classificação desejada. Na análise discriminante, a variável dependente é de natureza qualitativa (não métrica), ou seja, categórica ou discreta, já que seu valor representa uma classificação estabelecida. Podendo assim ser observado que ela funciona mais como um rótulo do que um valor em si. Com relação às variáveis independentes, são geralmente métricas com valores contínuos, mas também podem assumir valores que representem categorias. Uma regra de classificação que deve ser considerada é se as variâncias das populações são iguais ou não. Quando a regra de classificação assume que as variâncias das populações são iguais, isto é, homocedásticas as funções discriminantes são ditas lineares e quando não são, funções discriminantes quadráticas.

2.3 Boosting

O Boosting é um algoritmo flexível e simples de ser implementado em diversos cenários, além do seu potencial de classificação. Ele pode ser usado juntamente com outros algoritmos para melhorar o desempenho de tais classificadores. Ele efetua repetidas execuções de um classificador fraco produzindo assim uma combinação de vários classificadores insatisfatórios combinando os resultados gerando um eficiente. Em cada novo formado são atribuídos pesos, que são ajustados a cada classificador, fazendo com que cada um atue na falha do anterior. Os pesos atribuídos inicialmente na distribuição são uniformes. Portanto quando há um erro, o peso é elevado ou diminuído caso contrário. O cálculo do erro total é o somatório dos pesos ponderados dos objetos classificados incorretamente e é definido na equação (1).

$$\varepsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i] = \sum_{i:h_t(x_i)\neq y_i} D_t(i) \quad (1)$$

O cálculo da importância associada do classificador, calculada a partir do erro. É apresentado na equação (2).

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \quad (2)$$

Ao final de cada rodada os pesos são atribuídos de acordo com o índice da importância associada do classificador (α_t) é dada pela equação (3).

$$D'_{t+1} = D_t(i) \times e^{-\alpha \times h_t(x_i)} \quad (3)$$

Normaliza-se a distribuição de pesos na equação (4), onde Z_t é o fator de normalização.

$$D_{t+1} = \frac{D'_{t+1}}{Z_t} \quad (4)$$

Tendo, portanto, uma combinação linear de todos os coeficientes obtidos pela equação (5).

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t \times h_t(x)\right) \quad (5)$$

2.4 Métodos para comparação dos classificadores

Uma matriz de tabulação cruzada entre a classe prevista pelo modelo e a classe real dos exemplos é denominada de matriz de confusão. É uma maneira de apresentar as estatísticas para a avaliação de um classificador, onde temos uma tabela com as frequências absolutas (contagem). Considerando as classes como positivas e negativas temos, quando uma observação positiva é classificada como positiva, dizemos que ela é um verdadeiro positivo, um negativo é classificado como positivo, dizemos que ele é um falso positivo, um negativo classificado como negativo, verdadeiro negativo e um negativo classificado como positivo, falso negativo. E temos nas linhas os valores reais e nas colunas os valores preditos. Como podemos ver na Tabela 1.

Tabela 1: Matriz de confusão para um classificador

	Predito		
Real	Verdadeiro positivo	Falso negativo	Positivos reais
	Falso positivo	Verdadeiro negativo	Negativos reais
	Preditos positivo	Predito negativo	

Para medir o desempenho do classificador, foi utilizado o índice kappa, que é uma medida de concordância entre as classes preditas e observadas, que deduz o número esperado de acerto do classificador. É dado por:

$$kappa = \frac{(probabilidade\ de\ concordância\ - probabilidade\ de\ não\ concordância)}{(1 - probabilidade\ de\ não\ concordância)}$$

Considera-se, em geral, quando kappa for igual a zero não existe concordância, kappa menor que 0,4 concordância fraca, kappa entre [0,4; 0,8) moderada, kapa maior igual a 0,8 forte e kappa igual a um perfeita.

Outra importante medida utilizada na avaliação de classificadores é a acurácia, ou a proporção de observações que foram classificadas corretamente pelo classificador. É dado pela razão dos objetos classificados corretamente pelo total de objetos classificados. A taxa de erro aparente (TEA) é dada pela proporção de observações incorretamente classificadas pelo método de classificação.

Todas as análises foram feitas usando o software R. Primeiramente verificamos a pressuposição de homocedasticidade (variância comum) pelo teste qui-quadrado, onde a hipótese nula de que as populações são homocedásticas foi rejeitada, justificando a análise utilizando funções discriminantes quadráticas.

3 Resultados

A base original é composta por 294 pacientes. A qual foi dividida em duas, uma de tamanho 205, (70% da base original) para o treinamento, e a outra, de tamanho 89, (30% da base) para teste. Primeiramente, usamos a amostra de treinamento para os algoritmos. Utilizando a amostra de treinamento, ajustamos inicialmente modelos de análise discriminante linear e quadrática. As duas análises deram resultados bem próximos, com a análise discriminante quadrática um pouco melhor. O algoritmo boosting foi igual à análise discriminante quadrática como podemos ver na tabela (2).

Tabela 2. Taxas de erro, acurácia e índice kappa para os três classificadores analisados

Classificador	TEA	Acurácia	Kappa
Análise Discriminante Linear	0.1910112	0.8089888	0.606502
Análise Discriminante Quadrática	0.1685393	0.8314607	0.6602189
Boosting	0.1685393	0.8314607	0.6410325

Com apresentado na Tabela 2, podemos ver que os três classificadores tem uma taxa de erro baixa e sem muita diferença de um modelo para o outro. As acurácias foram bem elevadas todas acima de 80%. Com este exemplo não podemos definir qual classificador é mais eficiente devido à proximidade dos erros.

4 Conclusão

Na aplicação apresentada, o algoritmo boosting apresentou resultados muito bons e comparáveis aos encontrados pelos métodos estatísticos tradicionais de classificação. A utilização de metodologias modernas e eficientes, como boosting, pode trazer muitos benefícios em diversas áreas onde o problema de classificação aparece, principalmente quando as suposições dos métodos paramétricos não são satisfeitas.

5 Referências

[1] BACHE, K. & LICHMAN, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] FERREIRA, M. R. P., Análise discriminante clássica e de núcleo : avaliações e algumas contribuições relativas aos métodos Boosting e Bootstrap. Dissertação (mestrado) – Universidade Federal de Pernambuco. CCEN. Estatística, 2007

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

[4] JOHNSON, R. A.; WICHERN, D. W. Applied Multivariate Statistical Analysis. New York: Prentice Hall, 2002.

[5] LOESCH,C; HOELTGEBAUM, M. Métodos Estatísticos Multivariados, Ed. Saraiva, 2012.