Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Seleção de Características com kNN-SMOTE em Atributos do Mal de Alzheimer

Yuri Elias Rodrigues[1]
Evandro Manica[2]
Departamento de Matemática, UFRGS, Porto Alegre, RS
Eduardo Rigon Zimmer[3]
Departamento de Bioquímica, UFRGS, Porto Alegre, RS
Pedro Rosa Neto[4]
Tharick Ali Pascoal[5]
Translational Neuroimaging Laboratory, McGill University, Montréal, QC, Canada

Resumo.

Combinações de biomarcadores permitem um diagnóstico mais preciso e auxiliam no entendimento do progresso do mal de Alzheimer. No entanto, dependendo do conjunto observado, algumas características podem ser pouco relevantes ou não adicionarem vantagem significativa, gerando modelos mais complexos que requerem maior volume de dados para serem ajustados. Aqui avaliamos modelos de predição da doença em um problema de classificação com três classes (controle normal, comprometimento cognitivo leve, mal de Alzheimer) utilizando seleção de características. Para isso aplicamos o método dos k-vizinhos mais próximos *leaves-one-out* (kNN-LOOCV) simultaneamente com *Synthetic Minority Over-sampling Technique* (SMOTE) e comparamos os resultados de algoritmos indutivos de seleção de características com o *ranking* global. Para validação, definimos uma métrica para matrizes de confusão combinada com a validação cruzada 10-*fold*. Os algoritmos de indução encontraram o valor ótimo na fase de treino. Nossos resultados sugerem que algumas variáveis são irrelevantes em relação ao conjunto analisado.

Palavras-chave. reconhecimento de padrões, seleção de características, k-vizinhos mais próximos, biomarcadores para Alzheimer, dados desbalanceados.

1 Estratégias do classificador e comentários

O mal de Alzheimer pode ser inferido como resultado entre disfunções já que mudanças patológicas em relação ao cérebro resultam em complexas conformações, influenciando na

[1]yuri.rodrigues@ufrgs.br, yuri.rodrigues@acad.pucrs.br
[2]evandro.manica@ufrgs.br
[3]eduardo.rigonzimmer@mail.mcgill.ca
[4]pedro.rosa.neto@gmail.com
[5]tharick.alipascoal@mail.mcgill.ca

variabilidade de padrões da mesma classe. Em um conjunto de características que assumem valores reais disponibilizados pelo ADNI (*Alzheimer's Disease Neuroimaging Initiative*), cinco testes neuropsicológicos e três biomarcadores proteômicos, aplicamos algoritmos indutivos para incluir ou excluir características ao modelo de acordo com sua relevância. Comparamos os 247 modelos possíveis com os resultados dos algoritmos de seleção, direta, inversa e passo-a-passo para avaliarmos sua eficiência, pois estão suscetíveis a parar em máximos locais. Bases de dados médicos em geral possuem classes desbalanceadas, dado que quadros clínicos ocorrem em diferentes proporções. Optamos por usar o algoritmo kNN, implementado na linguagem *R*, pois tem revelado sucesso em diversas aplicações e sua formulação matemática faz parte de estratégias de pré-processamento de dados, como o SMOTE [1] por exemplo. Para lidar com as classes desequilibradas aplicamos SMOTE. Esta técnica produz dados sintéticos na região da classe minoritária da seguinte forma: dado um padrão da classe minoritária tomamos um outro padrão aleatório não sintético localizado em uma k-vizinhança do ponto em questão e adicionamos randomicamente o novo padrão entre eles, repetimos o procedimento até obter a proporção desejada. Com as classes equilibradas utilizamos o kNN-LOOCV para avaliar a sensibilidade do algoritmo à remoção da informação e identificar qual parâmetro k que maximiza a precisão. Este algoritmo remove um elemento do conjunto de treino por vez e toma a média entre todas as remoções. Para definir o kNN consideremos o conjunto de treino, $T = \{(x_i, y_i)\}$ com $i = 1, ..., n$, em que $x_i \in \mathbb{R}^m$ são os padrões e $y_i \in \Omega := \{c_1, ..., c_m\}$ suas classes correspondentes. Nosso objetivo é encontrar a classe $y_* \in \Omega$ à qual o padrão desconhecido x_* pertence. Logo,

$$\hat{y_*} = \underset{c \in \Omega}{\operatorname{argmax}} \sum_{x_j \in N\{x_*, k\}} \delta(y_j, c), \quad \text{para} \quad j = 1...m,$$

em que m é o número de classes, $N(x_*, k)$ é o conjunto dos k-vizinhos mais próximos do padrão a ser classificado, $\hat{y_*}$ é a classe calculada por votação pelo kNN e $\delta(c_i, c)$ é o delta de Kronecker. Os resultados estão representados na Figura 1.

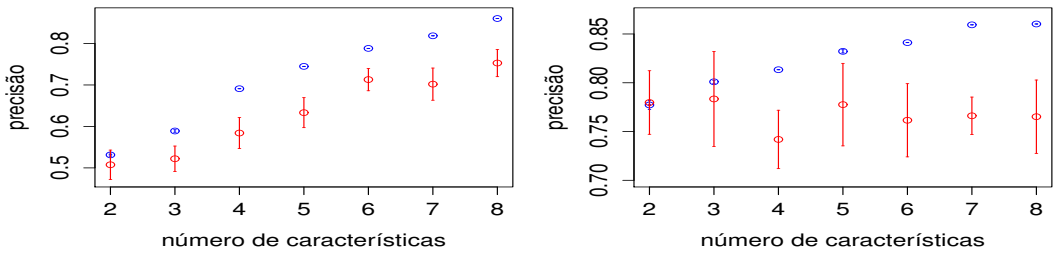

Figura 1: Combinações com menor (direita) e maior (esquerda) precisão pelo número de características. Em azul o resultado do treinamento e em vermelho a validação usando 20% da amostra.

Referências

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, Universität Trier, vol. 16, 321–357, (2002).

DOI: 10.5540/03.2015.003.02.0093