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.

Espectros de infravermelho próximo como ferramenta para a discriminação de espécies de Rhodniini

Erica Castilho Rodrigues [1]
Alexandre S. Paula [2]
Jaime Rodríguez-Fernández [3]
Hélcio Gil Santana [4] **Celio Pasquini** [5]
Cleber Galvão [4] [6]

1 Introdução

A espectroscopia de infravermelho próximo tem sido utilizada em vários campos de pesquisas e, mais recentemente, vem sendo incorporada nos estudos de sistemática biológica. A utilização crescente da espectroscopia NIR, em diferentes áreas de estudos, é o resultado de vários fatores, tais como: (1) aplicação universal em moléculas que têm ligações C-H, N-H, O-H, C-C ou S-H, (2) é um método rápido - um espectro pode ser obtido em menos de 60 segundos, (3) não há geração de resíduos sólidos, líquidos ou gasosos na espectroscopia NIR, (4) pequenas amostras podem ser utilizadas, e podem incluir material vivo ou amostras in situ, (5) as amostras não necessitam de tratamento prévio e, em princípio, é um método não invasivo e não destrutivo ([5], [7]).

Neste estudo aplicamos cinco modelos: *Penalized Linear Discriminant Analysis, Sparse Discriminant Analysis, Regularized Discriminant Analysis, Sparse Partial Least Square e Support Vector Machine* para a discriminação de oito espécies do gênero *Rhodnius Stal* (*Rhodnius brethesi Matta, R. milesi Carcavallo, Rocha, Galvão, Jurberg, R. nasutus Stal, R. neglectus Lent, R. pallescens Barber, R. pictipes Stål, R. robustus Larroussee R. stali Lent, Jurberg Galvão*) e comparamos os resultados obtidos para evidenciar qual destes métodos seria mais adequado para utilização em estudos de sistemática de Rhodniini. A escolha dos métodos aqui aplicados foi feita com base na performance que os mesmos apresentaram para diversos tipos de aplicação, como, por exemplo, classificação de alimentos e de óleos combustíveis [1], tipos de câncer e imagens [2].

[1]Departamento de Estatística - Universidade Federal de Ouro Preto e-mail: *ericarodrigues@iceb.ufop.br*
[2]Departamento de Biodiversidade, Evolução e Meio Ambiente - Universidade Federal de Ouro Preto *hetalex@terra.com.br*
[3]Ecossistema Consultoria Ambiental, Curitiba, Paraná *formycusub@yahoo.com.br*
[4]Laboratório Nacional e Internacional de Referência em Taxonomia de Triatomíneos do Instituto Oswaldo Cruz, Fundação Oswaldo Cruz *galvao@ioc.fiocruz.br*
[5]Instituto de Química, Universidade Estadual de Campinas

[6]Agradecemos ao CNPq, CAPES e FAPEMIG pelo apoio financeiro.

O método *Penalized Linear Discriminant Analysis* foi proposto por Witten et al. [8] e consiste em uma variação do método de Análise Discriminante tradicional. O método tradicional busca a projeção das variáveis que tornam a variação entre os grupos a maior possível, ao mesmo tempo em que minimiza a variância dentro de cada grupo. Este resultado é obtido através de uma técnica de maximização com restrição. A proposta de Witten et al. [8] foi modificar essa maximização, incluindo um termo de penalização que faz com que vários elementos da nova projeção fiquem iguais a zero.

A metodologia denominada *Sparse Discriminant Analysis* foi proposta por Clevense et al. [2]. Sua ideia central é utilizar, no processo de classificação, apenas aquelas variáveis que são, de fato, relevantes e eliminar os ruídos. No caso específico dos dados do infravermelho próximo, isso consiste em selecionar apenas aqueles comprimentos de ondas que realmente diferenciam as espécies. Essa seleção é feita associando uma penalização aos parâmetros das variáveis e forçando alguns deles a ficarem muito próximos de zero. Esse processo é feito simultaneamente à classificação, o que garante sua eficiência computacional.

Yiagian et al. [9] desenvolveram a técnica chamada de *Regularized Discriminant Analysis* que foi inicialmente aplicada para dados de Microarranjo. Porém, pode ser facilmente adaptada para o tipo de problema que estamos lidando neste trabalho. Para resolver o problema da singularidade da matriz de covariância eles a substituem por uma combinação linear com a matriz identidade. O processo de classificação também é alterado. Sabese que grande parte da diferença entre os grupos é apenas ruído aleatório. Dessa maneira, Yiagian et al. [9] propuseram a remoção desse ruído fixando um limiar para as variáveis, eliminando aquelas que não possuem poder discriminatório.

A técnica de *Sparse Partial Least Square* (SPLS) é uma proposta de Dongjun et al. [10] e consiste em uma adaptação do método *Partial Least Squares* para o caso em que o número de variáveis observadas é muito elevado, se comparado com o tamanho da amostra. Essa metodologia busca fazer seleção de variáveis e redução da dimensão de maneira simultânea. No processo classificação, a variável resposta é tratada inicialmente como se fosse contínua e não categórica. Em seguida um método de codificação é utilizado para encontrar os grupos.

O método *Support Vector Machine* (SVM) foi proposto originalmente por Vladimir et al. [3] e consiste em encontrar o hiperplano que melhor separa os grupos. Esse hiperplano é aquele que faz com que a distância entre eles seja a maior possível. Essa distância é chamada de margem e quanto maior essa margem, melhor é o classficador. Tal metodologia é capaz de lidar com situações em que os grupos não são linearmente separáveis. Neste caso basta mudar o núcleo do classificador para alguma forma mais genérica.

2 Material e Métodos

Os espectros foram obtidos com um espectrofotômetro FT-NIR BOMEM MB-160 (ABB, Canadá) com uma fonte de luz tungstênio-halogênio InGaAs e um detector de leitura para a

Penalized LDA	Regularized LDA	Sparse LDA	SPLS	SVM
25,83	67,50	31,25	64,16	58,75

Tabela 1: Porcentagem de acertos obtidas para os cinco métodos comparados.

região 800-2500 nm, com resolução de aproximadamente 0,7 nm. Os exemplares adultos (machos e fêmeas) foram colocadas sobre um acessório para refletância difusa. Foram utilizados cinco modelos *Penalized Linear Discriminant Analysis, Sparse Discriminant Analysis, Regularized Discriminant Analysis, Sparse Partial Least Square e Support Vector Machine* para verificar quais deles seria mais adequado para discriminação das espécies de *Rhodnius* com os dados de espectroscopia NIR (Near-infrared spectroscopy). A performance dos métodos foi comparada através da técnica de validação cruzada. Isso significa que dois terços da base de dados foram utilizados para ajustar os modelos. Esse conjunto compreendeu o conjunto de treinamento. O restante dos dados foram utilizados para testar a porcentagem de vezes que o método classificou corretamente os exemplares das espécies estudadas. Esse conjunto de dado compreendeu o conjunto de teste. Os elementos que fazem parte dos conjuntos de teste e treinamento foram selecionados de maneira aleatória. Para evitar que o resultado fosse sensível à seleção dos elementos, esse sorteio aleatório foi realizado 10 vezes e a taxa de acerto global calculada como a média dos resultados obtidos. Todas as análises foram realizadas com o software R e alguns pacotes adicionais que serão descritos ao longo do texto.

3 Resultados

A Figura 1 apresenta as taxas de acertos para cada uma delas para os dez diferentes grupos de treinamento selecionados. Os métodos que apresentaram melhores resultados foram o Regularized, Sparse Partial Least Square e o Support Vector Machine.

A Tabela 1 mostra a taxa média de acerto de cada um dos métodos aplicados. As três metodologias indicadas mencionadas atingiram um nível alto de acerto - acima 60% em quase todos os casos. Por se tratar de um problema de discriminação de espécies com um número elevado de categorias (oito), este resultado superou muito as expectativas. Se a classificação fosse feita de maneira totalmente arbitrária, esperaríamos uma taxa de acerto em torno de 12,5%, que está muito abaixo dos valores encontrados para todos os métodos utilizados neste estudo.

4 Conclusão e Discussão

De Lima, M. G. et al. [7] utilizaram espectroscopia NIR para a discriminação de espécies de Coenosiinae (Muscidae: Diptera), utilizando Análise de Componentes Principais e Análise Discriminante. Apesar da Análise Discriminante ser uma das metodologias usuais propostas por Fisher para classificações, ela não adequada para ser aplicada aos dados do infravermelho

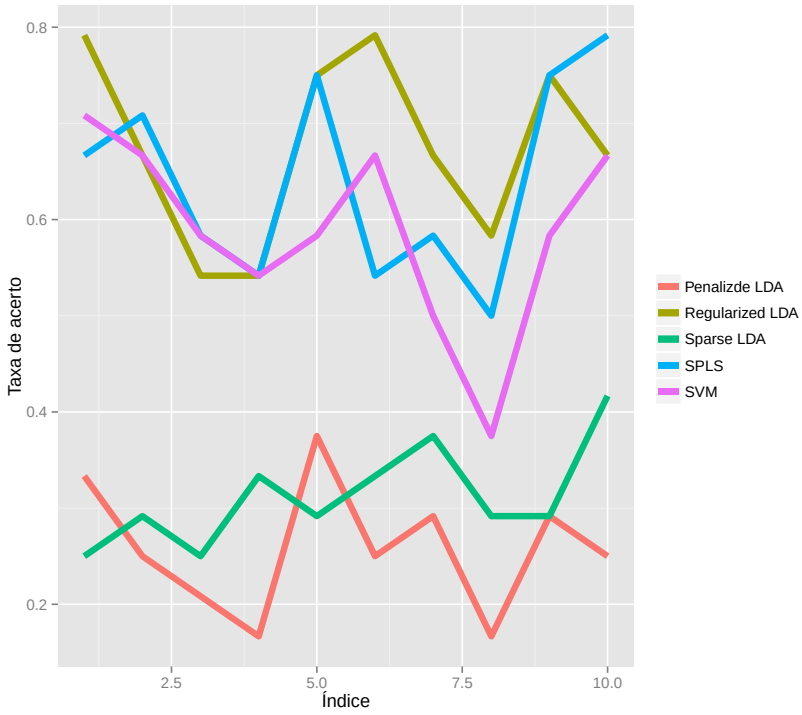

Figura 1: Comparação das taxas de acertos para os cinco métodos analisados: Penalized LDA, Regularized LDA, Sparse LDA, SPLSS e SVM.

próximo. Em geral, nesse tipo de aplicação o número de variáveis observadas (número de comprimentos de ondas coletados) é muito superior ao número de indivíduos observados. Assim, a matriz de correlações entre as variáveis é singular, o que leva a diversos problemas numéricos para estimação dos parâmetros. Em áreas como genética, epidemiologia e computação, esse tipo de situação também é comum. Os resultados encontrados neste estudo demonstram o grande potencial da utilização das técnicas aplicadas na classificação (discriminação) de espécies de *Rhdoniini*. As taxas de acerto obtidas foram elevadas, principalmente se considerarmos a complexidade do problema em termos de número de variáveis e categorias para um tamanho de amostra relativamente pequeno. Esse tipo de situação, em que o número de variáveis ou atributos é muito maior do que o número de indivíduos observados, têm se tornado recorrente nas mais diversas áreas. Portanto, técnicas adequadas devem ser aplicadas e avaliadas com cautela, principalmente por se tratarem de metodologias muito recentes.

Referências

[1] BALABIM, R. M.; SAFIEZA, R. Z. and LOKAMINA, E. I. Nearinfrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines. **Microchemical Journal**. Elselvier. v. 98, n. 1, p. 121-128, 2011.

[2] CLEMMENSEN, L.; HASTIE, T. and WITTEN D. and ERSBbaØll, B. Sparse discrimi-

nant analysis **Technometrics**. v. 53, n. 4, 2011.

[3] CORINA, C. and VLADIMIR V. Supportvector networks **Machine learning**. Springer. v. 20, n. 3, p. 273-297, 1995.

[4] AMI, D; NATALELLO, A.; ZULLINI, A. and DOGLIA, S. M. Fourier transform infrared microspectroscopy as a new tool for nematode studies **FEBS letters**. v. 576, n. 3, p. 297-300, 2004.

[5] PASQUINI, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications **Journal of the Brazilian Chemical Society**. v. 14, n. 2, p. 198-219, 2003.

[6] BENEDICT, A. A. Group classification of virus preparations by infrared spectroscopy **Journal of bacteriology**. v. 69, n. 3, p. 264-269, 1995.

[7] DE LIMA, M. G.; MOURA, M. O. and ARÍZAGA, G. C. G. Barcoding without DNA? Species identification using near infrared spectroscopy **Zootaxa**. v. 2933,p. 46-54, 2011.

[8] WITTEN, D. M. and TIBSHIRANI, R. Penalized classification using Fisher's linear discriminant **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**. v. 73, p. 753-772, 2011.

[9] GUO, Y.; HASTIE, T. and TIBSHIRANI, R. Regularized discriminant analysis and its application in microarrays **Biostatistics**. v. 1, p. 86-100, 2005.

[10] C HUNG, D. and KELES, S. Sparse partial least squares classification for high dimensional data **Statistical applications in genetics and molecular biology**. v. 9, 2010.

[11] ASSOCIAÇÃO BRASILEIRA DE HEREFORD E BRAFORD. **Hereford - Carne de qualidade tipo exportação**. Disponível em: http://www.hereford.com.br/. Acesso em: 13 de fevereiro de 2011.

[12] BARBIN, D. **Planejamento e análise estatística de experimentos agronômicos**. Arapongas: Editora Midas Ltda. 2003. 194 p.

[13] HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: models and estimation. **Computational Statistics & Data Analysis**. Elselvier. v. 27, n. 2, p. 151-170, 1998.

[14] PESCIM, R. R. **A distribuição beta generalizada semi-normal**. 2009. 124 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica), Universidade de São Paulo, Piracicaba, 2009.