I Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

O uso de Espectroscopia de Emissão Ótica com Plasma induzido por Laser (LIBS) para a identificação da composição do solo

Heloisa do Vale Guilherme [1]
ICMC - USP

Paulino Ribeiro Villas Boas [2]
Embrapa Instrumentação
ICMC - USP

1 Introdução

A Química Analítica é um campo de grande destaque nos últimos anos e representa uma promissora área de pesquisa. Dentre as diversas vertentes passíveis de aplicações destacam-se a agricultura, a área alimentícia e farmacêutica. No que concerne ao agronegócio, as técnicas analíticas tornaram-se excelentes aliadas em várias aplicações, em especial no estudo de composições do solo. A Espectroscopia de Emissão Ótica com Plasma induzido por Laser (LIBS) é uma técnica analítica que permite a avaliação de materiais, independente de sua natureza, de forma qualitativa e quantitativa, possibilitando a determinação de múltiplos elementos e moléculas a partir do plasma produzido pela emissão de pulsos de laser de alta energia na superfície da amostra. A transição dos níveis de energia dos átomos e íons no plasma emite radiação em diferentes comprimentos de onda conforme cada elemento. Sendo assim, as linhas de emissão presentes nos espectros coletados pela técnica LIBS podem ser utilizadas para determinar a composição da amostra [1, 3].

No entanto, os espectros decorrentes do uso desta técnica são complexos e sujeitos a uma série de fatores que podem dificultar a análise destes dados, tais como: flutuações de energia do laser, taxa de ablação, características do próprio plasma e o efeito de matriz. Tais razões podem impactar as características espectrais das linhas de emissão e desta forma, impedir o uso mais amplo da LIBS em aplicações e estudos. Como forma de mitigar estas dificuldades, este trabalho tem por objetivo promover a combinação da LIBS à técnicas de aprendizado de máquina, através do desenvolvimento de um modelo para a classificação e identificação de amostras a partir das

[1] heloisa.guilherme@usp.br
[2] paulino.villas-boas@embrapa.br

linhas de emissão para que, posteriormente, modelos de quantificação de elementos possam ser desenvolvidos para cada classe de amostra.

Nas próximas seções deste trabalho, serão abordados em detalhes a descrição dos dados e metodologia empregada, bem como os resultados da classificação de amostras presentes no conjunto de dados.

2 Materiais e métodos

Para o desenvolvimento deste trabalho, foram utilizados espectros LIBS de amostras de cinco tipos de solos, de dois conjuntos de folhas de plantas e de três amostras sintéticas: H_3BO_3, KBr e NaCl. O número de amostras variou entre os conjuntos, e para cada amostra foram coletados aproximadamente 100 espectros. O conjunto inteiro de dados possui 16.335 registros e 13.746 variáveis, dos quais cada registro corresponde a um espectro e cada variável, à intensidade das linhas de emissão na faixa de 190 nm a 980 nm. A variável alvo é a classe a que o espectro pertence e a partir desta, nota-se que os dados estão desbalanceados - com proporções de espectros muito distintas entre as classes, fator este que foi considerado durante a análise dos dados.

Espera-se testar e comparar o desempenho de diferentes técnicas de Aprendizado de Máquina para a construção do classificador. Neste sentido, fez-se uso dos seguintes algoritmos: Regressão Logística, Árvore de Decisão e *K-Nearest Neighbors* (KNN). A regressão Logística é um modelo baseado em uma função denominada distribuição logística cumulativa, cujos parâmetros são ajustados através do processo de treinamento. A partir desta função são estimadas de forma probabilística as relações entre a variável resposta categórica e as demais variáveis independentes. O modelo de Árvore de Decisão utiliza uma estrutura em árvores para definir os possíveis caminhos de decisão e os respectivos resultados. Diferentes algoritmos podem ser utilizados para sua construção, tais como o CART, ID3 e C4.5. O algoritmo *K-Nearest Neighbors*, baseado em distância, tem como processo de classificação o cálculo da distância entre o objeto desconhecido e um ou mais objetos rotulados. Sendo assim, a classe deste objeto será equivalente à classe da maioria dos objetos vizinhos mais próximos [2, 4, 5].

Para avaliar os resultados dos classificadores, tendo em vista se tratar de dados desbalanceados, foram aplicadas as seguintes métricas: *Precision* - confiabilidade positiva do modelo classificador, *Recall* - representa a eficácia do modelo classificador e *F1 score* - média harmônica entre a precisão (*Precision*) e a sensibilidade (*Recall*).

3 Resultados

Para cada algoritmo avaliado foi aplicada uma validação cruzada estratificada com 10 partições, otimização de hiper parâmetros Bayesiana e testes com um conjunto de validação. Para elegermos o melhor modelo para o objetivo proposto, as métricas foram avaliadas de forma macro (média) para inferir o comportamento geral dos modelos e por classe para observar o desempenho dos classificadores – em especial nas classes minoritárias. Ao estudar o conjunto de dados, observa-se que as classes H_3BO_3, KBr e NaCl são as de menor recorrência, com proporções de 1,99%, 1,99% e 1,83%, respectivamente.

A Tabela 1 apresenta a métricas gerais de avaliação para todos os modelos testados. Ao analisarmos os resultados, nota-se que os modelos selecionados cumpriram a tarefa de forma satisfatória, com bons resultados de *Precision* e *Recall* – o que demonstra uma boa confiabilidade ao classificar os verdadeiros positivos associado à uma alta eficácia. Entre os três modelos construídos, destacam-se a Árvore de Decisão e a Regressão Logística com F1 score médio de 91% e 96%, respectivamente, resultados bem superiores se comparados aos do modelo KNN.

Tabela 1: Resultados gerais de classificação por modelo

Modelo	Média Precision	Desvio Precision	Média Recall	Desvio Recall	Média F1 score	Desvio F1 score
Árvore de Decisão	0,91	0,02	0,91	0,02	0,91	0,02
Regressão Logística	0,97	0,01	0,96	0,02	0,96	0,02
KNN	0,82	0,02	0,82	0,02	0,81	0,02

As Tabelas 2, 3 e 4 mostram os resultados para cada modelo, segmentados por classe. Como os dados estão desbalanceados, nota-se que os resultados das métricas *Precision* e *Recall* apresentaram grande variação entre as classes – em especial para as classes *Plant Leaves 2*, *Soil 3* e *Soil 4* com métricas bem inferiores se comparadas às demais. Tratando-se de um fator muito evidente nos modelos de Árvore de Decisão e KNN. A classe *Plant Leaves 2*, dentre todas, foi a que apresentou a maior variabilidade durante as predições em todos os cenários testados, fato este que pode ser observado a partir do desvio padrão considerável obtido em ambas as métricas. Um dos motivos para este resultado é o menor número de instâncias desta classe – apenas 3,64% do conjunto de dados. Em contrapartida, para as demais classes minoritárias todos os modelos atingiram resultados ótimos de predição com métricas superiores a 96%. Este resultado era esperado, uma vez que as linhas de emissão para as amostras sintéticas são bem divergentes das demais, apresentando padrões característicos que contribuem positivamente na identificação de amostras pertencentes a estas classes. Para as classes *Soil 3* e *Soil 4*, a *Precision* e o *Recall* têm resultados muito próximos entre si e, estudando a matriz de confusão para ambas as classes detalhadamente, percebe-se a ocorrência de falsos positivos complementares. O perfil espectral destas classes são muito semelhantes, induzindo ao erro durante as predições, o que pode justificar a baixa performance das classificações apesar do maior número de instâncias.

Tabela 2: Resultados de classificação por classes para Árvore de Decisão

Classe	Média/Desvio Precision	Média/Desvio Recall	Média/Desvio F1 score	Classe	Média/Desvio Precision	Média/Desvio Recall	Média/Desvio F1 score
H_3BO_3	0,96 ± 0,11	0,97 ± 0,05	0,96 ± 0,07	**Soil 1**	1,00 ± 0,01	1,00 ± 0,01	1,00 ± 0,00
KBr	1,00 ± 0,00	1,00 ± 0,01	1,00 ± 0,01	**Soil 2**	0,99 ± 0,03	0,99 ± 0,02	0,99 ± 0,02
NaCl	1,00 ± 0,01	1,00 ± 0,00	1,00 ± 0,01	**Soil 3**	0,68 ± 0,09	0,65 ± 0,16	0,66 ± 0,12
Plant Leaves 1	0,99 ± 0,01	0,99 ± 0,03	0,99 ± 0,01	**Soil 4**	0,67 ± 0,10	0,70 ± 0,08	0,68 ± 0,07
Plant Leaves 2	0,59 ± 0,51	0,56 ± 0,49	0,58 ± 0,50	**Soil 5**	1,00 ± 0,00	0,99 ± 0,02	1,00 ± 0,01

Como critério para escolha do melhor modelo, decidiu-se priorizar a métrica F1 score por refletir de forma numérica o equilíbrio entre as métricas de confiabilidade (*Precision*) e eficácia (*Recall*). Levando em consideração as informações explicitadas, considerou-se a Regressão Logística como melhor modelo, uma vez que, mesmo nas classes de performance inferior, o resultado atingido foi superior aos demais modelos – demonstrando maior capacidade de adequação aos dados.

Tabela 3: Resultados de classificação por classes para Regressão Logística

Classe	Média/Desvio Precision	Média/Desvio Recall	Média/Desvio F1 score	Classe	Média/Desvio Precision	Média/Desvio Recall	Média/Desvio F1 score
H_3BO_3	1,00 ± 0,00	1,00 ± 0,02	1,00 ± 0,01	**Soil 1**	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00
KBr	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00	**Soil 2**	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00
NaCl	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00	**Soil 3**	0,91 ± 0,13	0,88 ± 0,14	0,88 ± 0,11
Plant Leaves 1	1,00 ± 0,00	0,97 ± 0,04	0,98 ± 0,02	**Soil 4**	0,87 ± 0,10	0,92 ± 0,15	0,89 ± 0,10
Plant Leaves 2	0,90 ± 0,32	0,89 ± 0,31	0,89 ± 0,31	**Soil 5**	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00

Tabela 4: Resultados de classificação por classes para *K-Nearest Neighbors*

Classe	Média/Desvio Precision	Média/Desvio Recall	Média/Desvio F1 score	Classe	Média/Desvio Precision	Média/Desvio Recall	Média/Desvio F1 score
H_3BO_3	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00	**Soil 1**	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00
KBr	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00	**Soil 2**	1,00 ± 0,00	0,64 ± 0,11	0,78 ± 0,09
NaCl	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00	**Soil 3**	0,45 ± 0,06	0,48 ± 0,19	0,49 ± 0,06
Plant Leaves 1	0,97 ± 0,04	0,99 ± 0,01	0,98 ± 0,02	**Soil 4**	0,46 ± 0,05	0,58 ± 0,06	0,51 ± 0,05
Plant Leaves 2	0,70 ± 0,48	0,60 ± 0,43	0,64 ± 0,45	**Soil 5**	1,00 ± 0,00	1,00 ± 0,00	1,00 ± 0,00

4 Conclusões

Neste trabalho, foram analisadas as principais características das linhas de emissão geradas a partir da técnica LIBS para tarefas de classificação. Para predição, foram empregados três algoritmos classificadores. Diferentes métricas foram utilizadas para análise de desempenho dos modelos, contudo aquele com maior equilíbrio entre as métricas (maior *F1 score*) foi selecionado como o melhor, favorecendo a performance de classificação das classes mais deficitárias. Assim, o modelo de melhor performance foi a Regressão Logística. Futuros trabalhos poderão adotar técnicas de amostragem para o balanceamento dos dados, assim como métodos de *Feature Engineering* e explicabilidade de modelos, visando ganho de desempenho dos modelos, bem como a identificação das variáveis de maior importância para a classificação e identificação de amostras.

Referências

[1] Costa, Vinicius C., et al. ”Laser induced-breakdown spectroscopy (LIBS): histórico, fundamentos, aplicações e potencialidades.”Química Nova 42.5: 527-545, 2019.

[2] Cover, T.; Hart, P. Nearest neighbor pattern classification. IEEE transactions on information theory, IEEE, v. 13, n. 1, p. 21–27, 1967.

[3] David A. Cremers and Leon J. Radziemski. Handbook of Laser-Induced Breakdown Spectroscopy. John Wiley & Sons, Ltd, second edition, 2013. ISBN 9781118567371.

[4] Fawcett, T.; Provost, F. Data Science para Negócios. [S.l.]: Alta Books, 2016.

[5] Kotsiantis, S. Supervised machine learning: A review of classification techniques. Informatica 31, p. 249–268, 2007.