Aplicación de clasificadores lineales y no lineales a espectroscopía de ruptura inducida por láser (LIBS)

F.R. Checozzi[a], J.Vorobioff[a,b], N. Boggio[a,c,d] , C. Rinaldi[a,c,d]

[a]CNEA, Av. Gral Paz 1499, San Martín, Bs.As. Arg.
[b]Universidad Tecnológica Nacional, Sarmiento 440, C.A.B.A., Arg.
[c]CONICET, Godoy Cruz 2290, C.A.B.A., Arg.
[d]Universidad Nacional de San Martín, 25 de Mayo y Francia, San Martín, Bs.As. Arg.
E-mail: federicochecozzi@cnea.gob.ar

Resumen

En este trabajo se comparan diferentes algoritmos para reconocer muestras de suelos no identificadas. Se aplicaron variadas técnicas de quimiometría para la clasificación de cuatro muestras diferentes de suelos sin etiquetar, extraídas de distintos lugares. Los espectros químicos obtenidos a partir de esas muestras mediante ensayos de espectroscopía de plasma inducida por láser (LIBS) se convirtieron a un formato tabular estándar para su procesamiento posterior. Luego se identificó posibles outliers (esto es, observaciones atípicas para un tipo específico de suelo) mediante el análisis de componentes principales robusto. Por último, se entrenaron y compararon modelos de clasificación en combinación con métodos de reducción dimensional. Se utilizaron algoritmos lineales (análisis discriminante lineal, análisis de componentes principales) y no lineales (análisis discriminante cuadrático, gradient boosting, mapas autoorganizados, aproximación y proyección uniforme de variedad). Se empleó la precisión (en inglés, accuracy) como métrica de evaluación de los modelos por validación cruzada, esto es, la proporción de observaciones correctamente clasificadas. Se concluye que los espectros por tipo de suelo son separables mediante modelos con fronteras de decisión lineales y que el uso de modelos y transformaciones más complejos son contraproductivos para esta aplicación porque intercambian sesgo, el cual genera errores mínimos en un modelo simple, por varianza. Esto no necesariamente se extiende a conjuntos de datos de estructura más compleja.

Materiales y métodos

El conjunto de datos consiste de 156 archivos de valores separados por comas (CSV) correspondientes a espectros químicos extraídos de cuatro tipos diferentes de suelo obtenidas mediante espectroscopía LIBS.

Se empleó una versión robusta del análisis de componentes principales (PCAR) y la metodología propuesta por Hubert[] para detectar outliers en las observaciones. Este método reduce la dimensionalidad los espectros tratando de encontrar las direcciones de mayor dispersión de los datos y realiza tests de hipótesis tanto sobre los espectros proyectados al espacio de menor dimensión como al residuo resultante de esa proyección.

Luego se realizó una comparación de algoritmos de clasificación lineales como el análisis discriminante lineal (LDA) y no lineales como el análisis discriminante cuadrático (QDA) y LightGBM. Estos métodos particionan el espacio de observaciones por clase de diferentes formas (empleando hiperplanos en el primero, superficies cuadráticas en el segundo y particiones binarias de cada dimensión en la última). En combinación con estos también se utilizó algoritmos de reducción dimensional para simplificar el entrenamiento de esos clasificadores, como el análisis de componentes principales (PCA y PCAR), mapas autoorganizados (SOM) y proyección uniforme de variedad (UMAP). Los mapas autoorganizados son redes neuronales que aprenden un mapa bidimensional de los datos identificando cúmulos de observaciones de mayor densidad, mientras que la proyección uniforme de variedad aprende la estructura de variedad presente en los datos mediante

un grafo y lo mapea a un grafo en otro espacio de menor cantidad de dimensiones. Se utilizó una estrategia de validación cruzada de 5 iteraciones para evaluar la precisión los distintos modelos.

Resultados

Detección de outliers

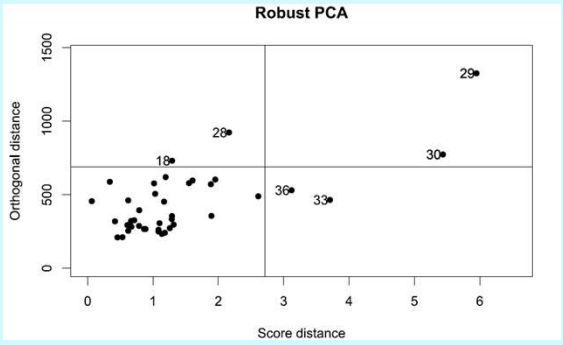

Fig. 1. Gráfico de diagnóstico de outliers para uno de los tipos de suelo.

Modelos de clasificación

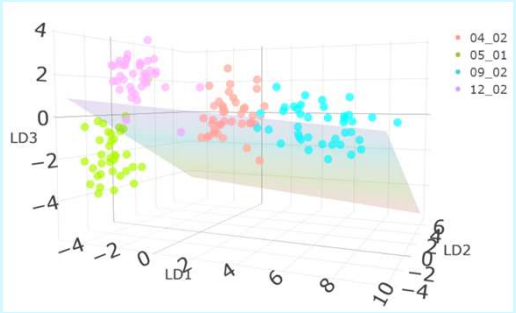

Fig. 2. Frontera de decisión para dos grupos en el espacio de variables canónicas.

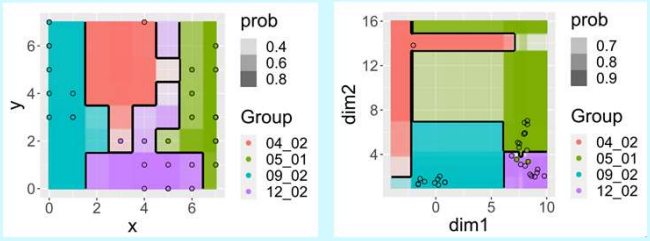

Fig. 3. Fronteras de decisión para: Izq.)LightGBM + SOM, Der.)LightGBM + UMAP. Los puntos corresponden a muestras del conjunto de test y la intensidad de los colores a la probabilidad de que una región sea de un determinado grupo.

Algoritmo	Tiempo promedio[s]	Notas
PCA	1.364744	90 componentes
PCAR	8,676014	90 componentes
UMAP	2.841912	Hiperparámetros por defecto
SOM	280.322036	-

Algoritmo	Exactitud	Tiempo promedio[s]	Notas
LDA	0,99375	32,114018	-
LDA	0,96875	0.064339	PCA
LDA	0,96875	0,042671	PCAR
QDA	0.90625	0.015281	PCA (23 componentes)
QDA	0,91250	0,023746	PCAR (23 componentes)
LightGBM	0,96250	4,238983	Hiperparámetros por defecto
LightGBM	0,96875	1400,645764	Optimizado
LightGBM	0.96250	0.102803	PCA con hiperparámetros por defecto
LightGBM	0.96250	57.473760	PCA Optimizado
LightGBM	0,93750	0,1091893	PCAR con hiperparámetros por defecto
LightGBM	0,94375	54,161651	PCAR Optimizado
LightGBM	0,93750	0,072268	UMAP con hiperparámetros por defecto
LightGBM	0,94375	133,904209	UMAP Optimizado
LightGBM	0,96250	0,068559	SOM con hiperparámetros por defecto
LightGBM	0,96250	122,837646	SOM Optimizado

Fig. 4. Tiempos de ejecución y precisión para algoritmos de clasificación y reducción dimensional.

Conclusiones

Los algoritmos lineales de reducción dimensional y clasificación funcionaron mejor que algoritmos no lineales. Esto es debido a que se intercambió sesgo por varianza. La cantidad de datos es baja y su estructura es sencilla por lo que modelos simples funcionan mejor. No es el caso necesariamente para datos más complejos.

Referencias

[1] Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics, 47(1), 64–79.