Clasificación de semilla de trigo utilizando Random Forest

Pedro Pablo Silva Antilef[1] [0000−1111−2222−3333]

Universidad de Santiago de Chile, Facultad de ingeniería, Departamento de ingeniería informática, Estación Central, Santiago de Chile pedro.silva@usach.cl

Abstract. En el presente trabajo se presenta un análisis de una base de datos con información geométricas de semillas de trigo a través de un modelo de clasificación utilizando la ampliamente utilizada técnica de *RandomForest*. Para la selección de modelos, se utiliza la métrica de *Out-of-bag error*, un índice de capa para medir la presición de la clasificación y finalmente la curva *Receiver Operating Characteristic*. Se encontró que *Random Forest* es muy preciso para la caracterización del *dataset* utilizado, pudiendo obtener altos valores de precisión con varias de las combinaciones de variables y parámetros presentados en el texto.

Keywords: Random Forest · Wheat Seeds · Data Minning.

1 Introducción

El reconocimiento y autentificación son tareas esenciales para garantizar la calidad de las semillas en la cadena productiva industrial, con énfasis en el proceso de certificación. Estas tareas se siguen haciendo de manera manual en varias industrias, por lo que la aplicación de métodos computacionales y automatizados podría significar una gran ayuda para diferentes empresas [1].

El reconocimiento de variedades también es una tarea crítica para limitar cualitativa y cuantitativamente las posibles pérdidas en el campo de cultivo, así como también para predecir la cosecha productiva asincrónicas [2].

En el presente texto se analiza la clasificación de semillas de trigo utilizando el algoritmo *Random forest*, el cual es una combinación de árboles predictores (ensamble de modelos) para tareas de clasificación y regresión. Este clasificador fue propuesto por [4] y luego mejorado por [5] en el año 2001 al agregar la idea de bagging al algoritmo, donde se particiona el dataset en varias ocasiones de manera aleatoria.

2 Estado del arte

Random forest ha sido ampliamente aplicado para distintas tareas de clasificación en variados campos de estudio. Un ejemplo es la aplicación en ecología, donde ha sido utilizado como la clasificación de especies de plantas invasoras o especies raras [6]. Otra campo de aplicación es la clasificación de uso de suelo

a partir de imágenes satelitales, donde a través de la recolección de puntos de control en terreno o por interpretación, se pueden clasificar matrices raster, obteniendo los diferentes usos de suelo de un territorio particular [7].

Random Forest también ha sido utilizado previamente para la clasificación de variedades de semillas de trigo. En [8] se utiliza *Fuzzy Cluster Random Forest* como método de clasificación para generar un proceso automatizado de clasificación de semillas, permitiendo ahorrar trabajo manual, también ayudando en el proceso de control de calidad y de selección de semillas dañadas, con una precisión promedio del 97.7%. Otro acercamiento a la clasificación de variedades de semillas de trigo es el desarrollado en [9], donde se aplican distintos clasificadores; *K-nearest neigbour, Support vector machine, Decision trees, Random Forest, Naiva bayes. Random Forest* fue el clasificador con mayor un 93%, seguido por *Decisión Trees* con un 92% de precisión.

3 Datos

El *dataset* utilizado es demoninado *seeds*, fue creado por el Instituto de Agrofísica de la Academia Polaca de Ciencias en Lublin, pero extraídos desde el repositorio *UCI Machine Learning Repository* de la Universidad de California Irvine [3]. Contiene 210 registros donde se describen 7 características geométricas del núcleo de 3 variedades de semillas de trigo: Kama, Rosa y Canadian, 70 elementos para cada variedad. Para esto se utilizaron técnicas de rayos X débil. Las variables contenidas en el *dataset* son: Área A, perímetro P, compactibilidad $C = 4\pi A/P^2$, largo del núcleo *l_kernel*, ancho del núcleo *w_kernel*, coeficiente de asimetría *assimemtry* y largo de la ranura del núcleo *l_groove*. Todas las variables son numéricas continuas y no existen datos con valores nulos o faltantes.

4 Metodología

Random Forest es un algoritmo de aprendizaje de maquina que utiliza como base el algoritmo de Árbol de decisión, ensamblando múltiples de estos árboles en la fase de entrenamiento y votando por las clases más seleccionadas de los árboles del bosque. Como métricas para la selección de modelos se utiliza la métrica *Out Of Bag Error*, el cual considera una validación cruzada dividiendo la muestra en un *dataset* de entrenamiento y otro de validación.

Para la selección de modelos, en primera instancia se realizan pruebas con 5 tipos de combinaciones de variables, variando los parámetros de cantidad de árboles desde 1000 a 3000 y de *mtry* desde 1 hasta n-1, siendo n la cantidad de variables que considera el experimento. Se seleccionan dos de los mejores resultados obtenido en esa primera instancia y se realizan 15 lanzamientos para cada uno, con el fin de obtener el mejor modelo a mostrar.

Se realizaron 3 experimentos para el análisis de este tipo de clasificación en el *dataset* de semillas de trigo:

- Experimento 1: Se utilizan todas las variables del *dataset*.

- Experimento 2: Se utilizan todas las variables del *dataset* menos "compactness".
- Experimento 3: Se utilizan las variables "l_groove", "perimeter" y "area".
- Experimento 4: Se utilizan las variables "l_groove", "area" y "asymmetry".
- Experimento 5: Se utilizan las variables "l_groove" y "area".

Para mostrar los resultados, se consideran 2 experimentos, aquel donde se utiliza el *dataset* completo (Experimento 1) y otro quitando varias variables que estaban demasiado correlacinadas entre sí (Experimento 4), principalmente las referentes a la geometría de la semilla como "perimeter", "area", "l_kernel" y "w_kernel".

5 Resultados

En esta sección se presentan los resultados obtenidos de los diferentes experimentos planteados y explicados en la sección 4. Estos resultados serán discutidos en la siguiente sección.

	ntrees			
mtry	**1000**	**1500**	**2000**	**3000**
1	6,67	7,14	7,14	8,1
2	6,19	6,67	5,71	6,19
3	5,71	6,19	6,19	6,19
4	5,71	5,71	5,71	5,24
5	5,24	5,24	5,24	5,24
6	6,19	5,71	5,71	5,71
7	6,19	6,19	6,19	6,19

Table 1: Parámetros y error *Out of Bag* - Experimento 1

	ntrees			
mtry	**1000**	**1500**	**2000**	**3000**
1	6,67	6,67	7,14	7,14
2	6,67	6,19	5,71	5,71
3	5,24	5,24	5,24	5,24
4	5,24	5,71	5,71	5,24
5	5,71	6,19	6,19	5,71
6	5,24	5,24	5,71	5,71

Table 2: Parámetros y error *Out of Bag* - Experimento 2

	ntrees			
mtry	**1000**	**1500**	**2000**	**3000**
1	5,71	5,71	5,71	5,71
2	6,19	6,19	5,71	6,19
3	6,19	6,19	5,71	6,19

Table 3: Parámetros y error *Out of Bag* - Experimento 3

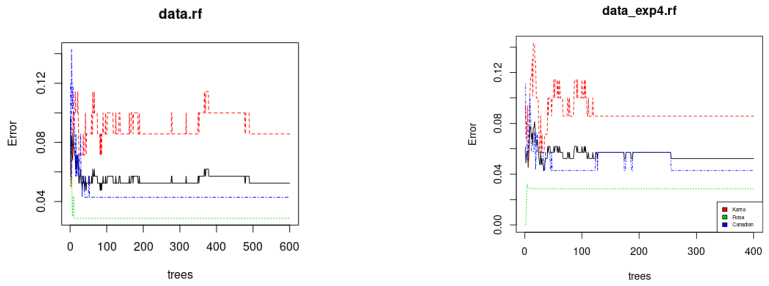
	ntrees			
mtry	**1000**	**1500**	**2000**	**3000**
1	5,24	5,24	5,71	5,71
2	5,24	5,24	5,24	5,24
3	5,24	5,71	5,71	5,71

Table 4: Parámetros y error *Out of Bag* - Experimento 4

	ntrees			
mtry	**1000**	**1500**	**2000**	**3000**
1	6,67	6,67	6,19	6,19
2	7,62	7,62	7,62	7,62

Table 5: Parámetros y error *Out of Bag* - Experimento 5

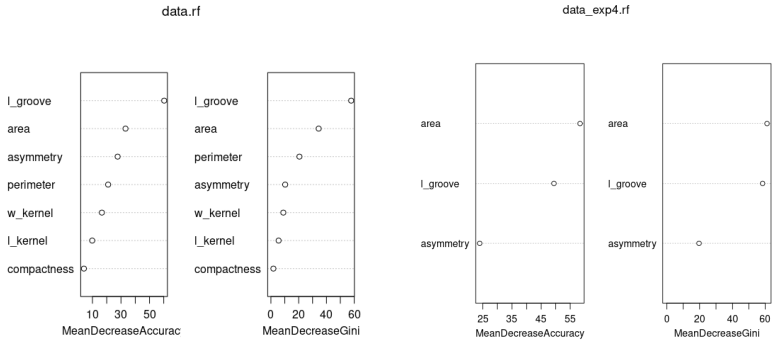
A continuación se muestran los resultados obtenidos para el experimento 1 y el experimento 4 a manera de facilitar la discusión de los resultados. Los cuales fueron elegidos para ser discutidos, ya que muchos de los modelos obtuvieron un alto rendimiento con ciertas combinaciones de los parámetros número de árboles y *mtry*. En primera instancia se muestran los gráficos de Error Out of Bag versus la cantidad de árboles, de manera de mostrar cuando el algoritmo comienza a converger a un error menor, como se puede apreciar en la Fig. 1. Otro output importante del modelo *Random Forest* es que entrega la importancia de las variables (Fig. 2) para el modelo de clasificación en base a dos métricas: *Mean Decrease Accurracy* y *Mean Decrease Gini*.



(a) Error v/s número de árboles para experimento 1

(b) Error v/s número de árboles para experimento 4

Fig. 1: Error v/s número de árboles

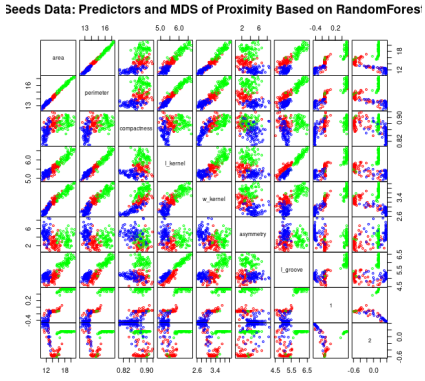


(a) Importancia de variables para experimento 1

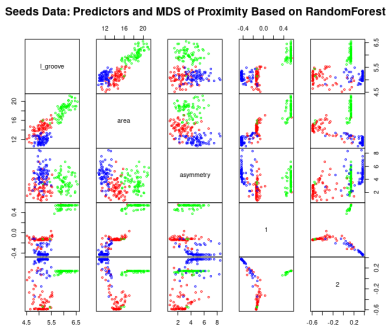
(b) Importancia de variables para experimento 4

Fig. 2: Importancia de variables

Además es posible obtener las matrices de proximidad (Fig.3) para analizar la relación de las variables de la clasificación en base a dos dimensiones, agregando además el escalamiento multidimencional en el sector inferior derecho de la matriz (Como se muestra con más detalle en la Fig. 4).

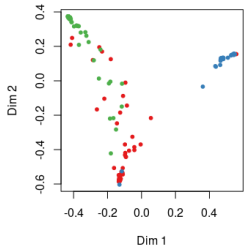


(a) Matriz de proximidad para experimento 1

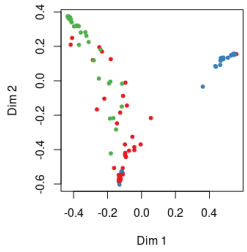


(b) Matriz de proximidad para experimento 4

Fig. 3: Matriz de proximidad



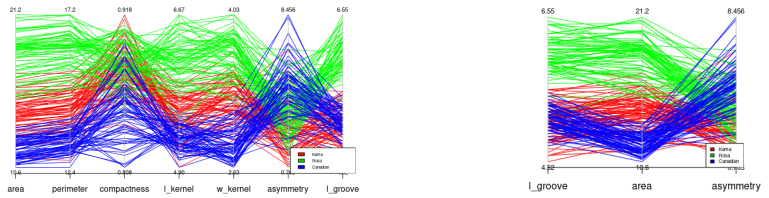
(a) Escalamiento Multidimencional para experimento 1



(b) Escalamiento Multidimencional para experimento 4

Fig. 4: Escalamiento Multidimencional

Finalmente, se presentan las coordenadas paralelas para las variables de cada uno de los experimentos.



(a) Coordenadas paralelas para experimento 1

(b) Coordenadas paralelas para experimento 4

Fig. 5: Coordenadas paralelas

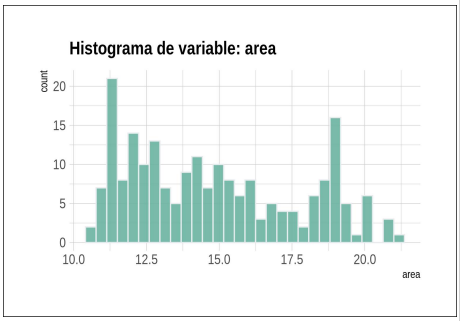
6 Discusión y conclusiones

En base a las métricas *Mean Decrease Accurac* y *Mean Decrease Gini*, se definen como variables más importantes "área","l_groove" , "perimeter" y "asymmetry" (en general, por lo visto en todos los lanzamientos realizados). Para los experimentos 1, 2 y 4, se encontró un valor mtry que obtenía los menores valores del error *Out of Bag* (OOB), para todas las combinaciones de árboles. Aún así, para el experimento 1, perdía sentido al aumentar la cantidad de árboles. Para el experimento 2 y 4, el error OOB se mantenía igual para mtry = 2 y mtry = 3 respectivamente, independiente de la cantidad de árboles. En los 3 experimentos nombrados anteriormente, los mejores valores para mtry (menor OOB error rate), se ajustaban a la teoría, donde se define *mtry* como la raíz cuadrada de n con n como la cantidad de variables.

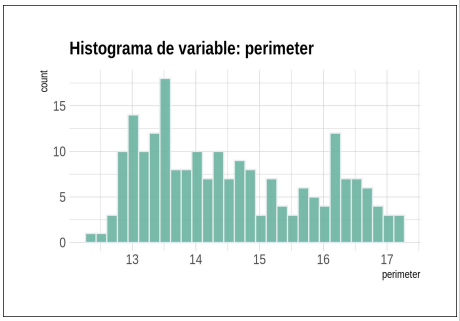
Si analizamos los dos experimento seleccionados (1 y 4), podemos observar que el modelo con menos variables converge mucho más rápido que aquel que tiene todas las variables. También es posible observar que siempre la variables más importantes son el largo de la ranura y el área de la semilla. La matriz de proximidad en ambos experimentos es exactamente igual, lo que sugiere que la clasificación se realiza de la misma manera, no fue posible mejorar la clasificación y bajar su error OOB mínimo de 5.24%. Por último y en base a las coordenadas paralelas, se puede observar que son principalmente las variables geométricas las que permiten diferenciar de mejor manera las variedades, sobre todo para la variedad Rosa.

Algo que llama la atención es la capacidad del *Random Forest* para clasificar el *dataset* utilizado, ya que pudo obtener el mismo error mínimo en varias combinaciones de variables y parámetros como se muestra en las tablas 1 a la 5. Cabe mencionar que el artículo donde se muestra la construcción de este dataset, logra resultados con más error que los obtenidos en este análisis. También que se obtiene el mismo resultado en el artículo donde se dicuten distintos métodos de clasificación y efectivamente *Random Forest* es el de mejor rendimiento.

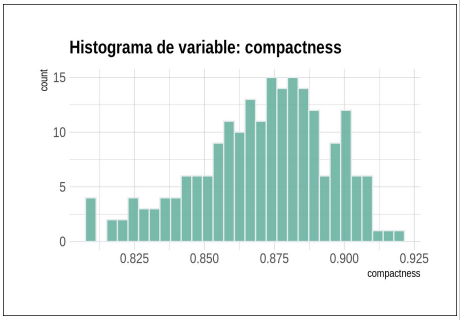
7 Anexos



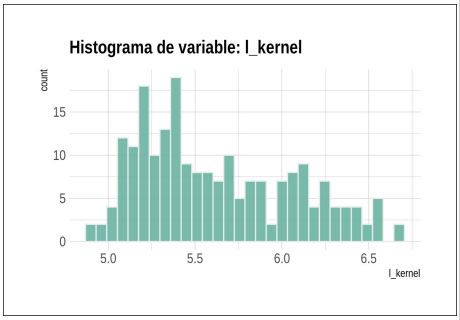
Histograma variable area



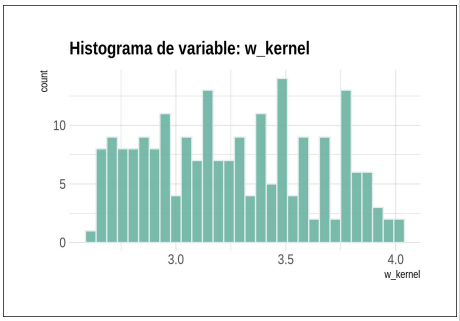
Histograma variable perimeter



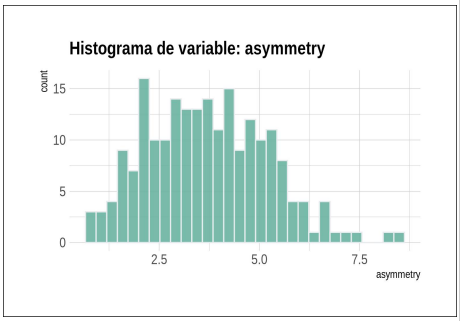
Histograma variable compactness



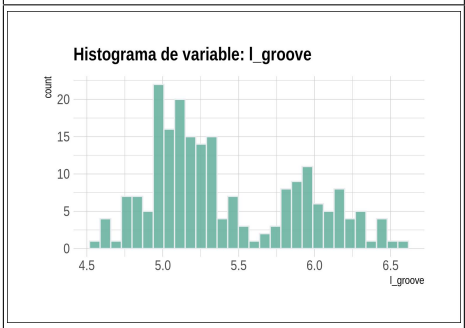
Histograma variable large kernel



Histograma variable width



Histograma variable assymetry



Histograma variable large kernel groove

References

1. Karim Laabassi: Wheat varieties identification based on a deep learning approach. Journal of the Saudi Society of Agricultural Sciences (2021)
2. Taheri-Garavand, A., Nasiri, A., Fanourakis, D., Fatahi, S., Omid, M., & Nikoloudakis, N.: Automated In Situ Seed Variety Identification via Deep Learning: A Case Study in Chickpea. Journal of the Saudi Society of Agricultural Sciences Plants (Basel, Switzerland), 10(7), 1406 (2021). https://doi.org/10.3390/plants10071406
3. UCI Repository, https://archive.ics.uci.edu/. Last accessed 12 Abr 2022
4. Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.
5. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
6. Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.
7. Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222.
8. Singh, P., Nayyar, A., Singh, S., & Kaur, A. (2020). Classification of wheat seeds using image processing and fuzzy clustered random forest. International Journal of Agricultural Resources, Governance and Ecology, 16(2), 123-156.
9. Priya, B. G. (2019). A comparison of various machine learning algorithms for wheat seed data set classification. An international journal of advanced computer technology, 7(5), 10-9756.