UNIVERSIDAD LAICA "ELOY ALFARO" DE MANABÍ

TIC´S

TECNOLOGIA DE LA INFORMACION Y COMUNICACION

Machine learning in the identification of native species from seed image analysis

---

Autores:
CHÁVEZ SIMALEZA JOSÉ RAÚL Y VELIZ FALLU DILAN STALIN

Profesor:
Ing. CESAR AUGUSTO SINCHIGUANO CHIRIBOGA

25 de mayo de 2024

Índice

1. Resumen

La identificación de semillas a nivel global es una tarea extremadamente difícil para los investigadores, incluso para los botánicos experimentados. Incluso hablando de especies nativas, sigue representando un desafío, ya que solo en Brasil existe una amplia gama de especies, todas ellas con un gran parecido biológico entre sí. Es por eso que el objetivo fue aplicar diferentes clasificadores de aprendizaje automático asociados al análisis de imágenes para identificar semillas de especies forestales. En total, se analizaron 155 especies nativas pertenecientes a 42 familias botánicas y se utilizaron diferentes tipos de algoritmos para entrenar a los clasificadores. El clasificador de árboles de decisión (DT) presentó una mayor precisión en la identificación correcta de las especies (82,8), seguido de los clasificadores ANN (81,7), k-NN (81,7), NBC (81,1) y SVM (78,7).

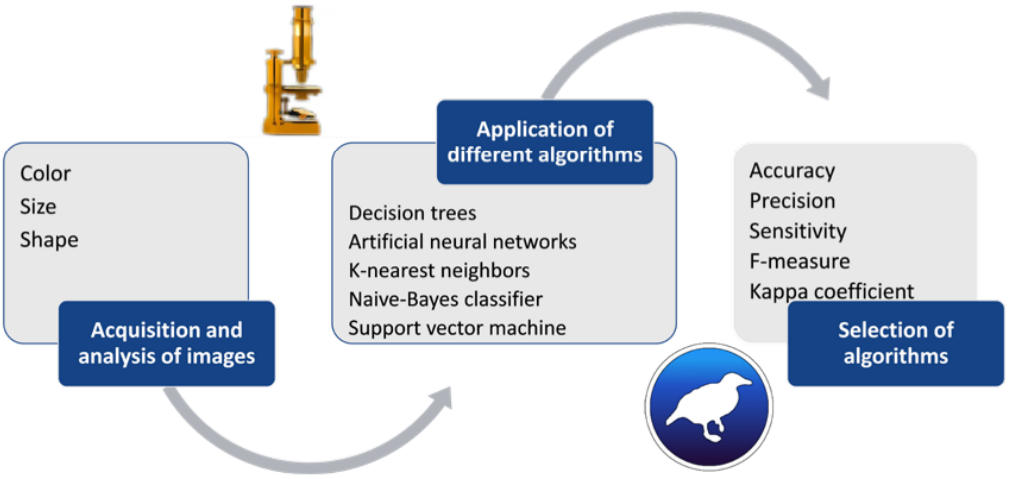
2. Abstract

The identification of seeds on a global scale is an extremely difficult task for researchers, even for experienced botanists. Even when dealing with native species, it remains a challenge, as Brazil alone has a wide range of species, all of which have a great biological resemblance to each other. That is why the objective was to apply different machine learning classifiers associated with image analysis to identify seeds of forest species. In total, 155 native species belonging to 42 botanical families were analyzed, and different types of algorithms were used to train the classifiers. The decision tree (DT) classifier showed the highest accuracy in correctly identifying the species (82.8), followed by the ANN (81.7), k-NN (81.7), NBC (81.1), and SVM (78.7) classifiers.

3. Introduccion

La identificación de semillas representa una tarea extremadamente difícil, ya que tan solo en la identificación de semillas a nivel de las especies nativas de Brasil existen un total de 35,653 especies de plantas diferentes, de las cuales 8,320 son especies arbóreas distribuidas en 138 familias y 938 géneros botánicos. Las semillas de todas estas tienen un gran parecido, ya que muchas de ellas comparten la misma forma, tamaño o color, lo que complica más la tarea de identificarlas. Es por eso que la identificación de semillas mediante el reconocimiento de inteligencia artificial es la opción más práctica para esta tarea. Aunque en los últimos años ha habido diversos avances en el uso de inteligencia artificial para la identificación de especies de semillas a partir de imágenes de semillas y frutos, esta sigue siendo un área casi inexplorada.

4. Materiales y metodos



4.1. Adquisición y procesamiento de imágenes de semillas

Se fotografió un total de 155 especies nativas, las cuales fueron capturadas con una cámara Canon PowerShot SX500 IS (f/4) con lente de 12 MP a una distancia de 50 cm de las semillas. Para poder ser fotografiada, la semilla se colocó en un escenario de fondo blanco con dimensiones de 50 cm x 50 cm x 50 cm, el cual permite que la luz no genere sombras ni interfiera con las imágenes.

Las imágenes por defecto aparecen en formato JPEG, pero se necesitan en una escala milimétrica, por lo que fueron subidas al software ImageJ.js versión 1.53, el cual permitió cambiar la escala de las imágenes a milimétrica.

El procesamiento de las imágenes resultó en un total de 1.827 millones de datos morfobiométricos obtenidos del procesamiento de imágenes de 101,521 semillas de 155 especies. Los datos fueron convertidos en un archivo (.CSV) que contiene los datos de las semillas, los cuales se utilizaron para el procesamiento de los modelos de aprendizaje automático en el software Weka, versión 3.8.3, el cual permite hacer minería de datos y ayuda al entrenamiento de la inteligencia artificial.

4.2. Aprendizaje Automático y Clasificación

Para el aprendizaje de los clasificadores que se utilizaran para el reconocimiento de las images se utilzaron los siguientes cinco técnicas de clasificación de aprendizaje super-

visado en este estudio: árboles de decisión (DT), redes neuronales artificiales (ANN), vecinos más cercanos (k-NN), clasificador Naive-Bayes (NBC) y máquina de vectores de soporte (SVM). El clasificador se seleccionó en función del rendimiento superior en precisión, sensibilidad y F-medida

4.3. Integración con Herramientas de Versionado

LaTeX es compatible con sistemas de control de versiones como Git, lo que facilita la colaboración entre múltiples autores. Los documentos de LaTeX son archivos de texto plano, lo que permite realizar un seguimiento eficiente de los cambios y gestionar diferentes versiones del documento. Esta compatibilidad es especialmente útil en proyectos de investigación colaborativa, donde es crucial mantener un registro claro de las contribuciones y modificaciones de cada autor.

- Árboles de decisión: organizan el conocimiento extraído del conjunto de datos en una estructura jerárquica similar a un árbol, compuesto por nodos y ramas; cada nodo interno representa un atributo y está asociado con una prueba para la clasificación de datos, mientras que los nodos y hojas del árbol corresponden a las clases y las ramas representan cada uno de los posibles resultados de las pruebas aplicadas (Quinlan, 1996).
- Redes neuronales artificiales: simulan el comportamiento del cerebro humano, compuesto por un gran número de elementos de procesamiento altamente interconectados, similares al funcionamiento de las neuronas biológicas, vinculados con conexiones ponderadas correspondientes a las sinapsis cerebrales (McCulloch y Walter, 1943).
- Vecinos más cercanos (k-NN): aprende en función de instancias, analizando las instancias o ejemplos alrededor de un caso específico. Este modelo calcula la distancia entre cada muestra de entrenamiento y el caso de prueba en función de la distancia euclidiana. Después de clasificar todas las distancias, el modelo selecciona los k más cercanos de aquellos que se consideran los k vecinos más cercanos (Aha et al., 1991).
- Clasificador Naive-Bayes: predice la clase para la cual la probabilidad a posteriori es mayor, dadas las variables predictoras del caso a clasificar, basado en la teoría de la probabilidad utilizando el teorema de Thomas Bayes (Shannon, 1948).
- Máquina de vectores de soporte (SVM): construye un hiperplano con una línea de decisión para la clasificación de instancias ampliamente utilizada en varias aplicaciones (Cortes y Vapnik, 1995; Vapnik, 1995).

4.4. Evaluación del rendimiento

4.4.1. Precisión General

$$\text{Precisión} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Descripción: Representa el porcentaje de predicciones correctas (tanto positivas como negativas) realizadas por el modelo en comparación con la clasificación real del conjunto de datos de validación. Es la relación entre las semillas correctamente clasificadas y el número total de semillas.

4.4.2. Precisión

$$\text{Precisión} = \frac{TP}{TP + FP} \tag{2}$$

Descripción: Indica la proporción de predicciones positivas que son realmente correctas. Un falso positivo ocurre cuando una semilla es predicha incorrectamente como positiva cuando en realidad es negativa.

4.4.3. Sensibilidad

$$\text{Sensibilidad} = \frac{TP}{TP + FN} \tag{3}$$

Descripción: Mide la capacidad del modelo para identificar correctamente las semillas positivas. Un falso negativo ocurre cuando una semilla predicha como negativa es en realidad positiva.

4.4.4. F-medida

$$\text{F-medida} = 2 \times \frac{\text{precisión} \times \text{sensibilidad}}{\text{precisión} + \text{sensibilidad}} \tag{4}$$

Descripción: Es la media armónica de la precisión y la sensibilidad, proporcionando una métrica única que equilibra ambas métricas.

4.4.5. Coeficiente de Kappa

$$\kappa = \frac{2 \times TP \times TN - FP \times FN}{TP \times FP + TN \times FN + 2 \times TP \times TN + FN^2 + FP \times TP + FP^2 + FN \times TP} \tag{5}$$

Descripción: Evalúa el nivel de acuerdo entre las clasificaciones realizadas por el modelo y las clasificaciones reales, ajustando por la posibilidad de que el acuerdo ocurra por azar.

5. Resultado

Los clasificadores de aprendizaje automático probados demostraron ser prometedores para la identificación de especies nativas a partir del procesamiento de imágenes y la obtención de atributos morfobiométricos de semillas forestales. El modelo de árbol de decisión mostró una mayor precisión para la identificación correcta de semillas del conjunto de datos de validación (82.8), cuyo coeficiente Kappa fue 0.780 (Tabla 2), seguido por los clasificadores que utilizan redes neuronales artificiales (81.7; 0.763), vecinos más cercanos (81.7; 0.754), clasificador Naive-Bayes (72.4; 0.683) y máquina de vectores de soporte (61.6; 0.584).

El modelo J48 (C4.5) alcanzó la mayor precisión, sensibilidad y F-medida (85.5, 83.8 y 84.3, respectivamente) para la identificación de especies nativas brasileñas a partir de semillas (Figura 2), superando a los otros clasificadores de aprendizaje automático. Esto muestra que las técnicas de procesamiento de imágenes y la identificación de atributos específicos de tamaño, forma y color, utilizados en este estudio, son adecuados para el análisis e identificación de semillas de especies nativas.

La identificación precisa de especies forestales nativas es esencial para el desarrollo de programas de conservación, restauración ecológica y manejo sostenible de los recursos forestales, lo que destaca la importancia de este estudio para la clasificación y preservación de la biodiversidad.

6. Conclusión

El procesamiento de imágenes y el uso de técnicas de aprendizaje automático hacen posible identificar semillas forestales nativas con una tasa de precisión satisfactoria. Se recomiendan los clasificadores basados en árboles de decisión.

Referencias

[1] BAO, F.; BAMBIL, D. Applicability of computer vision in seed identification: deep learning, random forest, and support vector machine classification algorithms. *Acta Botanica Brasilica*, v.35, n.1, p.17–21, 2021. https://doi.org/10.1590/0102-33062020abb0361

[2] BAO, Y.; MI, C.; WU, N.; LIU, F.; HE, Y. "Rapid classification of wheat grain varieties using hyperspectral imaging and chemometrics." *Applied Sciences*, v.9, n.19, e4119, 2019. https://doi.org/10.3390/app9194119

[3] CAO, J.; LIU, K.; LIU, L.; ZHU, Y.; LI, J.; HE, Z. Identifying mangrove species using field close-range snapshot hyperspectral imaging and machine-learning techniques. *Remote Sensing*, v.10, n.12, e2047, 2018. https://doi.org/10.3390/rs10122047

[4] CORTES, C.; VAPNIK, V. Support-vector network. *Machine Learning*, v.20, n.3, p.273–297, 1995. http://dx.doi.org/10.1007/BF00994018

[5] COSTA, M.F.; LOPES, A.C.A.; GOMES, R.L.F.; ARAÚJO, A.S.F.; ZUCCHI, M.I.; PINHEIRO, J.B.; VALENTE, S.E.S. Characterization and genetic divergence of Casearia grandiflora populations in the Cerrado of Piaui State, Brazil. *Floresta e Ambiente*, v.23, n.3, p.387-396, 2016. https://doi.org/10.1590/2179-8087.007115

[6] DUAN, Z.; MIN, Z.; ZHIFANG, Z.; SHAN, L.; LEI, F.; XIA, Y.; YAQIN, Y.; YI, P.; GUOAN, Z.; SHULIN, L.; ZHIXI, T. Natural allelic variation of GmST05 controlling seed size and quality in soybean. *Plant Biotechnology Journal*, v.20, n.9, p.1807-1818, 2022. http://dx.doi.org/10.1111/pbi.13865

[7] FARRIS, E.; ORRÙ, M.; UCCHESU, M.; AMADORI, A.; PORCEDDU, M.; BACCHETTA, G. Morpho-colorimetric characterization of the Sardinian endemic taxa of the genus Anchusa L. by seed image analysis. *Plants*, v.9, n.10, p.1–14, 2020. https://doi.org/10.3390/plants9101321

[8] FELIX, F.C.; MEDEIROS, J.A.D.; FERRARI, C.S.; VIEIRA, F.A.; PACHECO, M.V. Biometry of Pityrocarpa moniliformis seeds using digital imaging: implications for studies of genetic divergence. *Brazilian Journal of Agricultural Sciences*, v.15, n.1, e6128, 2020. https://doi.org/10.5039/agraria.v15i1a6128

[9] FELIX, F.C.; KRATZ, D.; RIBEIRO, R.; NOGUEIRA, A.C. Characterization and differentiation of forest species by seed image analysis: a new methodological approach. *Ciência Florestal*, v.33, n.3, e73427, 2023. https://doi.org/10.5902/1980509873427

[10] FERREIRA, R.L.A.; CERQUEIRA, R.M.; CARDOSO-JUNIOR, R.C. Analysis of botanical identification in forest inventories of sustainable management plans on wester Pará state, Brazil. *Nature and Conservation*, v.13, n.3, p.136-145, 2020. https://doi.org/10.6008/CBPC2318-2881.2020.003.0014

[11] FERREIRA, T.; RASBAND, W. ImageJ: user guide (IJ 1.46r), 2012. 198p.

[12] FRANKLIN, S.E.; AHMED, O.S. Deciduous tree species classification using object-based analysis and machine learning with unmanned aerial vehicle multispectral data. *International Journal of Remote Sensing*, v.39, p.5236-5245, 2017. https://doi.org/10.1080/01431161.2017.1363442

[13] McCULLOCH, W.S.; WALTER, P. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, v.5, n.4, p.115-133, 1943.

[14] MEDEIROS, A.D.; PINHEIRO, D.T.; XAVIER, W.A.; SILVA, L.J.; DIAS, D.C.F.S. Quality classification of Jatropha curcas seeds using radiographic images and machine learning. *Industrial Crops and Products*, v.146, p.112-162, 2020. https://doi.org/10.1016/j.indcrop.2020.112162

[15] MITCHELL, T. M. *Machine Learning*. McGraw–Hill Science/Engineering/Math, 1997. 421p.

[16] MUKASA, P.; WAKHOLI, C.; FAQEERZADA, M.A.; AMANAH, H.Z.; KIM, H.; JOSHI, R.; SUH, H.K.; KIM, G.; LEE, H.; K