Clasificación De Cerámicos Utilizando Árboles De Decisión: Una Comparativa Entre Árboles De Decisión, AdaBoost Y Random Forest. Y Un Análisis Del Mejoramiento De Rendimiento Utilizando PCA Y LDA

Brandon Marquez Salazar

***Resumen*—El reconocimiento de patrones es un área del procesamiento de señales que radica en la interpretación del comportamiento a fin de obtener un descriptor numérico. Diversas metodologías han sido implementadas para una correcta caracterización de las señales. Entre ellas, el uso de descriptores de comportamiento basado en la periodicidad.**

En el ámbito del procesamiento de señales, existen diversas formas de describir una imagen, estas se pueden ver descritas en libros como [1]–[3]. Estas caracterizaciones son, a la vista de un humano, poco intuitivas, ya que están, principalmente diseñadas para un reconocimiento relativo dentro de un conjunto de datos (clasificación) utilizando clasificadores [4].

Existen diferentes clasificadores como los árboles de decisión [5], Random Forest [6] y el AdaBoost [7].

Dentro del ámbito del aprendizaje de máquina, existen conceptos importante, como *the curse of dimensionality* [8] y *the non-free lunch theorem* [9]. Estos doc conceptos son esenciales para comprender que cada dataset y cada combinación de características puede comportarse mejor o peor según el algoritmo utilizado, ya que, numéricamente se tratan de diferentes ecuaciones resultantes al finalizar el aprendizaje.

I. Introducción

En este documento, se reporta el experimento de clasificación de cerámicos utilizando diferentes algoritmos de aprendizaje supervisado—árboles de decisión, AdaBoost y Random Forest—. Así mismo, se intentará mejorar la eficiencia de los algoritmos utilizando PCA y LDA, lo que puede equipararse a un proceso de *decoupled AutoML* [10], lo cual permite una selección de modelo y de características simultáneamente.

II. Metodología

La presente investigación se desarrolla mediante un experimento estructurado en cinco fases consecutivas, cada una diseñada para evaluar diferentes enfoques en la extracción de características y reducción de dimensionalidad para la clasificación de defectos en imágenes de cerámicos. Cada fase emplea un conjunto específico de parámetros y técnicas, con el objetivo fundamental de determinar la configuración óptima que maximice la precisión en la clasificación, basándose en principios establecidos de reconocimiento de patrones [4] y procesamiento de imágenes digitales [3].

A. Fase 1: Análisis de Características GLCM

En esta fase inicial se explora la utilización de características de matriz de co-ocurrencia de niveles de gris (GLCM) [3], tanto en su forma original como con aplicación de análisis de componentes principales. La extracción de características se realiza considerando distancias de 1, 3 y 7 píxeles, en cuatro direcciones angulares distintas de 0, 45, 90 y 135 grados. Las propiedades texturales evaluadas incluyen contraste, correlación, energía, homogeneidad y varianza. Los clasificadores empleados en esta fase son Random Forest [6], Máquinas de Vectores de Soporte y Vecinos Más Cercanos, con y sin la aplicación de reducción de dimensionalidad mediante PCA conservando el 95% de la varianza.

B. Fase 2: Evaluación de Características GLR

La segunda fase se centra en el análisis de características de longitud de ejecución de niveles de gris (GLR) [3]. Esta aproximación examina patrones en tres direcciones diferentes de 0, 45 y 90 grados, extrayendo cinco medidas características fundamentales: Énfasis de Ejecuciones Cortas, Énfasis de Ejecuciones Largas, No Uniformidad de Niveles de Gris, No Uniformidad de Longitud de Ejecución y Porcentaje de

Ejecuciones. Al igual que en la fase anterior, se emplean los tres clasificadores principales con configuración estándar y con aplicación de PCA manteniendo el 95% de la varianza explicada.

C. Fase 3: Implementación de Características SDH

La tercera fase incorpora el uso de histogramas de suma y diferencia (SDH) para la extracción de características [11]. Los parámetros de operación incluyen una distancia de 1 píxel y cuatro ángulos de análisis de 0, 45, 90 y 135 grados. Las siete características extraídas comprenden media, varianza, correlación, contraste, homogeneidad, sombreado y prominencia. La evaluación se realiza con los mismos tres clasificadores de las fases anteriores, tanto en configuración original como con reducción de dimensionalidad mediante PCA conservando el 95% de la varianza.

D. Fase 4: Integración de Características Combinadas

La cuarta fase representa un enfoque integrador que combina todas las características extraídas en las fases anteriores (GLCM, GLR y SDH) en un único conjunto de datos multivariado. Esta fase evalúa comparativamente la efectividad de dos técnicas de reducción de dimensionalidad: Análisis de Componentes Principales y Análisis Discriminante Lineal, ambas técnicas fundamentales en el procesamiento de características [4]. Los mismos tres clasificadores se aplican sobre el conjunto de características combinadas, buscando identificar la técnica de reducción dimensional que produzca los mejores resultados de clasificación.

E. Fase 5: Optimización con Análisis Discriminante Lineal

La fase final consiste en una etapa de optimización donde se aplica Análisis Discriminante Lineal al mejor conjunto de características identificado en las fases anteriores. Esta fase toma la combinación óptima de características y clasificador de las fases 1 a 3 y aplica una transformación lineal discriminante para maximizar la separabilidad entre clases, utilizando principios de aprendizaje supervisado [4]. El resultado de esta fase representa la configuración final recomendada para el sistema de clasificación de defectos.

III. Resultados

Los Resultados obtenidos fueron los siguientes

IV. Conclusiones

Referencias

[1] C. M., *Pattern Recognition and Machine Learning*, 1st ed., ser. Information Science and Statistics. New York, NY: Springer, aug 2006.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. John Wiley & Sons, 2012.

[3] A. K. Jain, *Fundamentals of digital image processing*. Upper Saddle River, NJ: Pearson, sep 1988.

[4] C. M. Bishop, "Pattern recognition and machine learning," 2006.

[5] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] Z.-H. Zhou, "Ensemble methods: foundations and algorithms," 2012.

[8] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[9] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, no. 7, pp. 1341–1390, 1996.

[10] A. Quemy, "Two-stage optimization for machine learning workflow," *Information Systems*, vol. 92, p. 101483, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S03064379193053

[11] R. E. Sánchez-Yañez, "Tarea(para el 0826): Caractersticas a partir de histogramas de sumas y de diferencias," 2025.

Cuadro I: Comparative Analysis of Classification Performance Across Multiple Feature Extraction and Dimensionality Reduction Techniques

Case	Features	Reduction	Classifier	Accuracy	Precision	Recall	F1-Score	AUC	CV Score	Prob.
Case 1	GLCM	No PCA	SVC	0.8125	0.8250	0.8125	0.8115	N/A	0.6718	No
			KNeighbors	0.8750	0.9167	0.8750	0.8667	0.9525	0.8449	Yes
			RandomForest	0.8125	0.8167	0.8125	0.8095	0.9674	0.8154	Yes
			LogisticRegression	0.7500	0.7875	0.7500	0.7431	0.9401	0.7038	Yes
	GLCM	PCA	SVC	0.6875	0.5750	0.6875	0.6032	N/A	0.7474	No
			KNeighbors	0.8125	0.8929	0.8125	0.8128	0.9408	0.8885	Yes
			RandomForest	0.8125	0.8929	0.8125	0.7818	0.9701	0.8590	Yes
			LogisticRegression	0.6875	0.5208	0.6875	0.5875	0.8555	0.7167	Yes
Case 2	GLR	No PCA	SVC	0.1875	0.1500	0.1875	0.1667	N/A	0.2192	No
			KNeighbors	0.6250	0.5595	0.6250	0.5747	0.8099	0.7821	Yes
			RandomForest	0.6250	0.5595	0.6250	0.5747	0.7897	0.7821	Yes
			LogisticRegression	0.4375	0.3833	0.4375	0.3984	0.6146	0.3410	Yes
	GLR	PCA	SVC	0.1875	0.0982	0.1875	0.1288	N/A	0.1731	No
			KNeighbors	0.6250	0.6250	0.6250	0.6111	0.7598	0.8436	Yes
			RandomForest	0.6250	0.6250	0.6250	0.6111	0.8711	0.8615	Yes
			LogisticRegression	0.4375	0.3500	0.4375	0.3571	0.6797	0.2333	Yes
Case 3	SDH	No PCA	SVC	0.8125	0.8250	0.8125	0.8115	N/A	0.9077	No
			KNeighbors	0.7500	0.8125	0.7500	0.7661	0.9154	0.8128	Yes
			RandomForest	0.8750	0.8875	0.8750	0.8740	0.9720	0.8295	Yes
			LogisticRegression	0.8125	0.8500	0.8125	0.8056	0.9310	0.6564	Yes
	SDH	PCA	SVC	0.8125	0.8167	0.8125	0.8095	N/A	0.8244	No
			KNeighbors	0.7500	0.7750	0.7500	0.7560	0.9160	0.7628	Yes
			RandomForest	0.7500	0.7500	0.7500	0.7500	0.9609	0.7628	Yes
			LogisticRegression	0.6875	0.7250	0.6875	0.7004	0.8607	0.6692	Yes
Case 4	Combined	LDA	SVC	0.8750	0.8875	0.8750	0.8740	N/A	0.8103	No
			KNeighbors	0.8125	0.8250	0.8125	0.8115	0.9492	0.7808	Yes
			RandomForest	0.8125	0.8375	0.8125	0.7986	0.9779	0.8282	Yes
			LogisticRegression	0.6875	0.7292	0.6875	0.6792	0.8932	0.7167	Yes
	Combined	PCA	SVC	0.7500	0.8125	0.7500	0.7542	N/A	0.7974	No
			KNeighbors	0.6250	0.7571	0.6250	0.6364	0.7656	0.7667	Yes
			RandomForest	0.7500	0.8167	0.7500	0.7476	0.9063	0.8756	Yes
			LogisticRegression	0.4375	0.5464	0.4375	0.4520	0.7852	0.7962	Yes
Case 5	Best	LDA	RandomForest	0.8125	0.8042	0.8125	0.8026	0.9642	0.8128	Yes