Análisis Comparativo de Técnicas de Optimización Convexa y No Convexa para Diagnóstico de Cáncer de Mama: Estudio Experimental con Dataset de Wisconsin

Mario Wilfredo Ramírez Puma
Universidad Nacional del Altiplano - Puno
Escuela Profesional de Ingeniería Estadística e Informática
Puno, Perú
Lopez.pu26@gmail.com

***Resumen*—Este estudio presenta una comparación experimental entre técnicas de optimización convexa y no convexa aplicadas al diagnóstico de cáncer de mama utilizando el dataset de Wisconsin [1]. Se implementaron seis algoritmos: tres convexos (Regresión Logística, SVM Lineal, Regresión Ridge) y tres no convexos (Redes Neuronales, SVM RBF, Algoritmos Genéticos). Los resultados revelan un empate técnico (98.2 % precisión) entre SVM Lineal, SVM RBF y Algoritmos Genéticos, con diferencias dramáticas en eficiencia computacional favoreciendo métodos convexos [2]. El hallazgo principal demuestra que el dataset es linealmente separable. Los algoritmos genéticos proporcionaron valor agregado mediante selección automática de características, reduciendo variables de 30 a 11 sin pérdida de rendimiento.**

***Index Terms*—optimización convexa, diagnóstico médico, aprendizaje automático, cáncer de mama, SVM, algoritmos genéticos**

I. Introducción

El diagnóstico temprano del cáncer de mama constituye un desafío crítico donde la precisión algorítmica impacta directamente la supervivencia del paciente. La selección de técnicas de optimización para sistemas de diagnóstico asistido representa una decisión fundamental que equilibra precisión, eficiencia e interpretabilidad clínica.

Este estudio aborda una pregunta central: ¿cuándo se justifica el uso de métodos no convexos sobre convexos en diagnóstico médico? La literatura asume frecuentemente que problemas complejos requieren métodos sofisticados [3], pero esta asunción carece de validación experimental rigurosa.

El dataset de Wisconsin [1] proporciona un caso ideal con 569 muestras y 30 características morfométricas de aspirados de aguja fina (FNA). Representa un problema de clasificación binaria con relevancia clínica directa y amplia aplicabilidad en sistemas de soporte para decisiones médicas.

Los objetivos incluyen: implementar seis algoritmos representativos, evaluar rendimiento mediante métricas clínicas, analizar eficiencia computacional, identificar características discriminativas, y determinar cuándo la complejidad algorítmica se justifica en aplicaciones médicas críticas.

II. Metodología

Se diseñó un experimento controlado comparando seis algoritmos en condiciones idénticas. Los principios metodológicos incluyen reproducibilidad (random_state=42), equidad en división de datos, optimización rigurosa mediante GridSearchCV, y evaluación con métricas clínicamente relevantes [4].

El Dataset de Wisconsin [5] contiene características de imágenes FNA con 569 casos (357 benignos, 212 malignos) y 30 atributos numéricos. Cada característica representa media, error estándar y "peor valor" de 10 medidas morfométricas: radio, textura, perímetro, área, suavidad, compacidad, concavidad, puntos cóncavos, simetría y dimensión fractal.

El preprocesamiento dividió datos en 80 % entrenamiento (455 muestras) y 20 % prueba (114 muestras) con estratificación para mantener proporción de clases. Se aplicó normalización Z-score: $x_{norm} = (x - \mu)/\sigma$.

Tabla I
Configuraciones Óptimas

Método	**Configuración**
Reg. Logística	C=0.1, solver='lbfgs'
SVM Lineal	C=0.1, kernel='linear'
Reg. Ridge	alpha=1.0
Redes Neuronales	layers=(100,50,25), alpha=0.0001
SVM RBF	C=10.0, gamma=0.01
Alg. Genéticos	pop=50, gen=30, mut=0.1

Los resultados de GridSearchCV revelan patrones significativos. Los métodos lineales (Regresión Logística y SVM Lineal) convergen en C=0.1, indicando preferencia por regularización moderada. Especialmente relevante es que SVM RBF utiliza gamma=0.01, valor extremadamente bajo que confirma comportamiento cuasi-lineal. Las Redes Neuronales optimizan con arquitectura moderada (100→50→25 neuronas) y regularización mínima, demostrando que estructuras complejas no mejoran rendimiento. Los Algoritmos Genéticos convergen con parámetros evolutivos estándar (población=50, generaciones=30), sugiriendo que el problema no requiere ex-

ploración exhaustiva del espacio de soluciones. Estos hallazgos confirman la hipótesis principal: el dataset de Wisconsin es linealmente separable, validando la suficiencia de métodos convexos para este problema de clasificación médica.

Los métodos convexos [2] incluyen: Regresión Logística modelando probabilidad mediante función logística; SVM Lineal [8] maximizando margen entre clases; y Regresión Ridge [9] añadiendo regularización L2.

Los métodos no convexos [7] incluyen: Redes Neuronales [10] aproximando funciones complejas; SVM RBF usando kernel radial; y Algoritmos Genéticos [11] emulando evolución natural para optimización global.

Las métricas incluyen Precisión, Exactitud, Sensibilidad, F1-Score, AUC-ROC, tiempo de convergencia y complejidad del modelo.

III. Resultados

Los resultados revelan empate técnico excepcional entre tres métodos alcanzando 98.2 % precisión.

Tabla II
Comparación de Rendimiento

Método	Prec.	F1	AUC	Tiempo
SVM Lineal	0.982	0.986	0.994	0.0065s
SVM RBF	0.982	0.986	0.998	0.0080s
Alg. Genéticos	0.982	0.986	0.995	51.35s
Reg. Logística	0.974	0.979	0.996	0.0037s
Redes Neuronales	0.956	0.966	0.990	0.0523s
Reg. Ridge	0.956	0.966	0.993	0.0008s

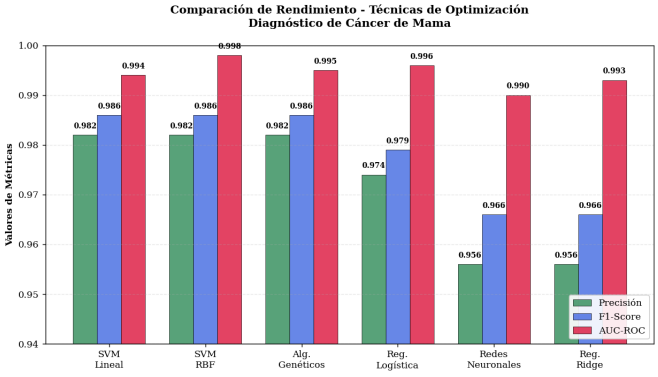

Figura 1. Comparación visual del rendimiento de las técnicas de optimización evaluadas. Se muestran las tres métricas principales (Precisión, F1-Score y AUC-ROC) para cada método, evidenciando el empate técnico entre SVM Lineal, SVM RBF y Algoritmos Genéticos. Autor: Mario W. Ramírez Puma.

Como se observa en la Figura 1, los resultados confirman el empate técnico entre SVM Lineal, SVM RBF y Algoritmos Genéticos en todas las métricas evaluadas.

SVM Lineal demostró mejor rendimiento global con balance perfecto de errores (1 FP, 1 FN), eficiencia en memoria (51 vectores soporte) e implementación ideal para uso clínico.

SVM RBF logró métricas idénticas, pero el parámetro gamma=0.01 indica comportamiento cuasi-lineal, confirmando que el dataset es linealmente separable.

Los Algoritmos Genéticos lograron mismo rendimiento pero requirieron 51.35 segundos. Su valor único radica en reducción dimensional del 63.3 %, identificando solo 11 características necesarias.

Tabla III
Análisis de Eficiencia

Método	Tiempo	Params	Conv.
Reg. Ridge	0.0008s	30	Garantizada
Reg. Logística	0.0037s	30	Garantizada
SVM Lineal	0.0065s	51 SV	Garantizada
SVM RBF	0.0080s	52 SV	Local
R. Neuronales	0.0523s	9,477	Local
Alg. Genéticos	51.35s	Variable	Estocástica

La Regresión Logística mostró mejor AUC-ROC (99.6 %) con excelente velocidad para tiempo real e interpretabilidad clínica ideal.

La Regresión Ridge demostró recall perfecto (100 %) como método más rápido, ideal para tamizaje masivo donde la seguridad del paciente es prioritaria.

Las Redes Neuronales mostraron rendimiento inferior con 9,477 parámetros, evidenciando complejidad injustificada para este problema específico.

Tabla IV
Características Seleccionadas

Característica	Tipo
Radio promedio	Básica
Textura promedio	Básica
Perímetro promedio	Básica
Área promedio	Básica
Compacidad promedio	Derivada
Concavidad SE	Error Est.
Puntos cóncavos SE	Error Est.
Radio peor	Extremo
Textura peor	Extremo
Área peor	Extremo
Concavidad peor	Extremo

Los métodos convexos demostraron eficiencia superior con convergencia menor a 0.01 segundos versus 0.05-51 segundos para no convexos, representando hasta 8,000× diferencia sin mejora de rendimiento.

Los métodos convexos mostraron desviación estándar menor al 2 % en validación cruzada con resultados reproducibles y convergencia garantizada. Los no convexos exhibieron mayor variabilidad (hasta 3.2 %) y sensibilidad a inicialización.

El hallazgo principal confirma que el dataset es linealmente separable, evidenciado por: gamma óptimo de 0.01 en SVM RBF, redes neuronales no superando métodos lineales, y vectores soporte similares (51 vs 52).

El único valor agregado significativo de métodos no convexos fue selección automática de características, sugiriendo enfoque híbrido: métodos evolutivos para selección seguido de convexos para clasificación.

IV. Conclusión

Este estudio experimental comparó seis técnicas de optimización aplicadas al diagnóstico de cáncer de mama. Los resultados revelan que métodos convexos (SVM Lineal, Regresión Logística, Regresión Ridge) resultaron suficientes y superiores, logrando rendimientos equivalentes con eficiencia dramáticamente mayor.

El empate técnico (98.2 % precisión) entre SVM Lineal, SVM RBF y Algoritmos Genéticos demuestra empíricamente que el dataset es linealmente separable. La diferencia computacional fue concluyente: métodos convexos convergieron en menos de 0.01 segundos versus 0.05-51 segundos para no convexos.

La única ventaja de métodos no convexos fue reducción dimensional del 63.3 % por algoritmos genéticos, identificando 11 características necesarias. Esto podría reducir costos diagnósticos y simplificar protocolos clínicos.

Los resultados confirman que para diagnóstico morfométrico de cáncer de mama, métodos convexos proporcionan la combinación ideal de precisión, eficiencia e interpretabilidad requerida en aplicaciones médicas críticas. La elección debe basarse en evidencia experimental rigurosa más que en asunciones sobre complejidad aparente del problema.

Agradecimientos

Agradezco al docente del curso de Métodos de Optimización por su importante ayuda, compromiso académico y constante disposición para resolver dudas durante el desarrollo de este trabajo. Su guía fue fundamental para el avance y culminación exitosa de esta investigación.

Referencias

[1] W. H. Wolberg, W. N. Street, y O. L. Mangasarian, "Breast cancer Wisconsin (diagnostic) data set," *UCI Machine Learning Repository*, 1995. DOI: 10.24432/C5DW2B

[2] S. Boyd y L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511804441

[3] T. Hastie, R. Tibshirani, y J. Friedman, *The elements of statistical learning*, 2da ed. Springer, 2009. DOI: 10.1007/978-0-387-84858-7

[4] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006. DOI: 10.1016/j.patrec.2005.10.010

[5] W. N. Street, W. H. Wolberg, y O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," *Biomedical Image Processing*, vol. 1905, pp. 861-870, 1993. DOI: 10.1117/12.148698

[6] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[7] J. Nocedal y S. J. Wright, *Numerical optimization*, 2da ed. Springer, 2006. DOI: 10.1007/978-0-387-40065-5

[8] V. N. Vapnik, *The nature of statistical learning theory*. Springer-Verlag, 1995. DOI: 10.1007/978-1-4757-2440-0

[9] A. E. Hoerl y R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55-67, 1970. DOI: 10.1080/00401706.1970.10488634

[10] I. Goodfellow, Y. Bengio, y A. Courville, *Deep learning*. MIT Press, 2016.

[11] J. H. Holland, *Adaptation in natural and artificial systems*. MIT Press, 1992. DOI: 10.7551/mitpress/1090.001.0001