

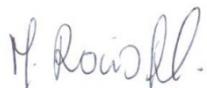
San José, 18 de noviembre del 2019.

A quien interese.

Reciban un cordial saludo de parte del programa TIC-as de la Cooperativa Sulá Batsú y Latinas in Computing como parte del comité organizador de la conferencia de Mujeres Latinoamericanas en Tecnologías: Latinity, y el comité organizador local conformado por organizaciones, instituciones y empresas comprometidas con la visibilización del aporte de las mujeres en las áreas STEAM.

Por medio de la presente hacemos constar que la ponencia titulada “Maquinas de Soporte Vectorial con Kernels Ortogonales para el diagnostico de la Diabetes Mellitus” fue presentada en formato de póster por la investigadora Karla Angélica Doctor Mauricio en la conferencia Latinity, llevada a cabo el pasado 6 y 7 de setiembre en el Instituto Interamericano de Cooperación para la Agricultura (IICA), en Coronado, San José, Costa Rica. Dicha ponencia fue resultado de la investigación que lleva el mismo nombre y tuvo como comité asesor a los señores Luis Carlos Padierna García y Carlos Villaseñor Mora.

Sin más por el momento, se despide.



Rocío Jiménez Calderón
Gestora de proyectos
Cooperativa Sulá Batsú



Latinity 2019

Tiene el honor de entregar el siguiente certificado a:

Karla Angelica Doctor Mauricio

Por su participación en la 4ta edición de la conferencia Latinity con el póster:

Máquinas de Soporte Vectorial con Kernels ortogonales para el diagnóstico de la Diabetes Mellitus

El 6 y 7 de setiembre del 2019 en el Instituto Interamericano de Cooperación para la Agricultura (IICA),
en Coronado, San José, Costa Rica.

KEMLY CAMACHO JIMÉNEZ
COOPERATIVA SULÁ BATSÚ

LUZA JARAMILLO
LATINAS IN COMPUTING



LATIN AMERICAN WOMEN IN TECHNOLOGY





Kernel Ortogonal s-Hermite para Máquinas de Soporte Vectorial aplicadas al diagnóstico de Diabetes Mellitus

Doctor Mauricio Karla , Padierna Luis Carlos , Villaseñor Mora Carlos
División de Ciencias e Ingenierías - Campus León

Las Máquinas de Soporte Vectorial (MSV) tienen sus raíces en la teoría del aprendizaje estadístico y se han convertido en herramientas poderosas para resolver problemas de aprendizaje automático: clasificación, regresión, detección de valores atípicos, entre otros. En esta teoría se explica el por qué es posible realizar un aprendizaje a partir de puntos de entrenamiento finitos y qué estrategia se emplea para superar algunas dificultades tradicionales, como la "maldición de la dimensionalidad", la "adaptación excesiva o sobreajuste", etc. En la presente investigación se diseñan y aplican MSV para determinar la presencia o ausencia de Diabetes Mellitus. Los datos sobre los cuales se aplicaron los modelos son los datasets Pima Indian y Diabetic Retinopathy, obtenidos del repositorio de la Universidad de California, Irvine (UCI). La hipótesis planteada es que el uso del kernel s-Hermite conllevará a la disminución de sobreajuste. Los experimentos se realizaron tomando al kernel RBF como referencia de comparación contra s-Hermite. El algoritmo empleado para resolver las MSV fue LIBSVM en C/C++. Los resultados preliminares obtenidos apoyan la hipótesis planteada.

1. INTRODUCCIÓN

En el área médica se requieren modelos confiables que auxilien a los profesionales de la salud a realizar mejores diagnósticos. Actualmente existen numerosas técnicas computacionales empleadas para diagnóstico médico, tales como las máquinas de soporte vectorial (MSV).

La Diabetes Mellitus es una enfermedad crónica y progresiva caracterizada por altos niveles de glucosa en la sangre que afectan al cuerpo en general con numerosas complicaciones incrementando el riesgo de morir prematuramente.

En esta investigación se diseñan y aplican MSV poniendo a prueba una función kernel ortogonal, recientemente desarrollada, para determinar la presencia o ausencia de Diabetes Mellitus y patologías asociadas

2. METODOLOGÍA

La selección de los parámetros de las MSV se realizó con la técnica Grid Search, necesaria para entrenar modelos con datos no reportados en la literatura, otras opciones son: búsqueda aleatoria, estimación de Algoritmos de distribución (EDA), metaheurísticas bioinspiradas, entre otros. Los parámetros fueron explorados en los siguientes valores $C = [2^{-5}, 2^2]$, $\gamma = [2^{-5}, 2^2]$ y $n = \{1, 2, \dots, 6\}$.

El objetivo de identificar el valor de los parámetros requeridos por el kernel es para que el clasificador pueda predecir con precisión las etiquetas de los datos; sin embargo, el buscar estos parámetros con los datos conocidos nos induce un posible sesgo al enfrentar al modelo a datos nuevos, por lo que es necesario realizar una validación cruzada. Esto consiste en separar el dataset en distintas partes, se empleó el método de validación cruzada de 10 pliegues.

Los conjuntos de datos fueron Pima Indian Diabetes y Diabetic Retinopathy[1]. Pima se conforma de 768 registros, con 8 medidas clínicas cada uno: concentración plasmática de glucosa, presión arterial diastólica, Índice de masa corporal, número de embarazos por paciente, edad, espesor de pliegues cutáneos en Tríceps, entre otros. Diabetic contiene 1151 registros y 19 medidas clínicas, las características de este conjunto de datos fueron extraídas de imágenes para predecir si presenta signos de retinopatía diabética o no.

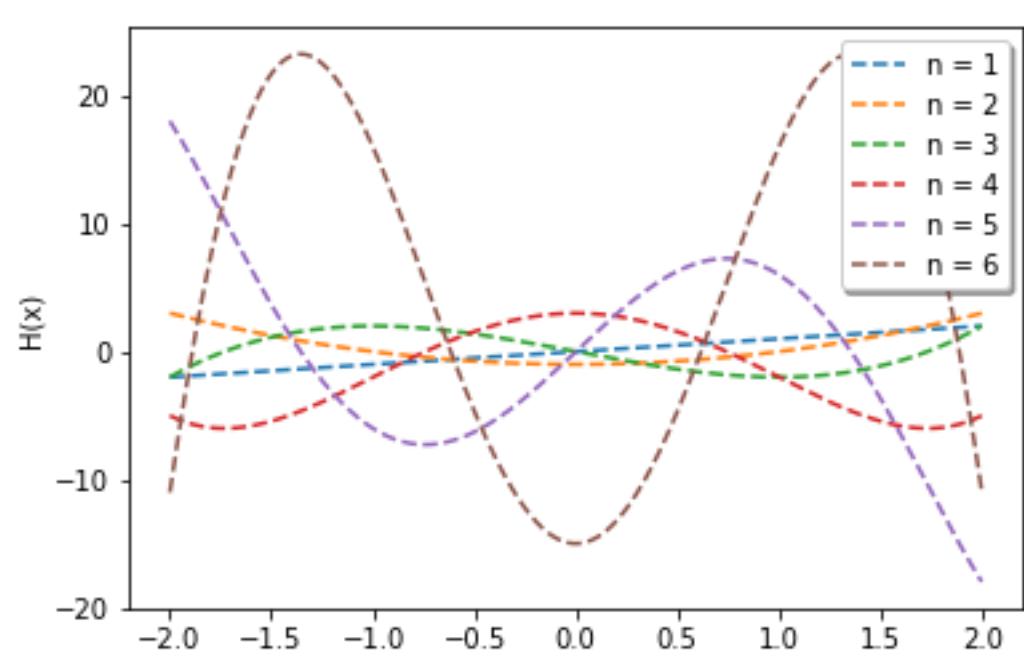


Fig. 1. Polinomio de s-Hermite

$$K_{s-Herm}(\mathbf{x}, \mathbf{z}) = \prod_{j=1}^d \sum_{i=0}^n H_{ei}(x_j) H_{ei}(z_j) (e^{-(x_j^2 + z_j^2)/2}) (2^{-2n}) \quad (\text{Eq.1})$$

$$K_{RBF}(\mathbf{x}, \mathbf{z}) = e^{(-\gamma \|\mathbf{x}_i - \mathbf{z}_i\|^2)} \quad (\text{Eq.2})$$

Mediante el uso de kernels, se construye un algoritmo no lineal a partir de uno lineal sin cambiar el diseño central del algoritmo. Esta observación conocida como el "truco del kernel". Por lo cual se utilizaron dos funciones kernel RBF(Eq.2) y s-Hermite(Eq.1)[2], además el entrenamiento de las MSV se realizó con el algoritmo LIBSVM en su versión para C/C++[3] y como índices de rendimiento se emplearon el Accuracy y la Proporción de Vectores Soporte (PVS).

$$\text{Accuracy} = \frac{\text{Predicciones correctas}}{\text{Predicciones totales}} \quad (\text{Eq.3})$$

IMPLEMENTACION KERNEL s-HERMITE

Pseudocódigo polinomio s-Hermite

INPUT: Parámetros dados por el usuario (número de ensayos y Grado del polinomio.)

OUTPUT : Polinomio de s-Hermite grado n

```

1 if grado del polinomio (n = 0)                                // n : 1 a 6 en este trabajo
2   Return 1
3 if grado del polinomio(n = 1)
4   Return x
5   Return x*H(x, n-1) - (n-1)*H(x, n-2)
```

INPUT : Parámetros definidos por el usuario (Grados de polinomio, número de registros, etc).

OUTPUT: Información de rendimiento (Accuracy y nVS) para realizar análisis estadísticos

```

1 for each d ∈ Datasets do
2   Partition Datasetd into DataFolds = {Trainingj,d, Testj,d}j=110    // 10-Validacion de pliegues
3   for each j ∈ Kernels do
4     for each i ∈ Kernels do
5       He(x, z) ← (Hei(xj)Hei(zj)(e-(xj2 + zj2)/2)(2-2n)
6       [Accuracyi,j, PVSi,j] ← SVM(Kerneli, Hi)
7     end
8   end
9 end
```

Fig 2. Metodología experimental para la evaluación del kernel de s-Hermite

3. RESULTADOS

La resolución de problemas de clasificación con máquinas de soporte vectorial mostró ser una herramienta potente para la detección y confirmación de personas con enfermedades crónicas como la Diabetes Mellitus. La ventaja de s-Hermite (Accuracy = 75.06%) de requerir menos vectores soporte en comparación con el kernel RBF (Accuracy = 69.89%) nos da la posibilidad de tener modelos con menor probabilidad de sobreajuste, lo cual resulta importante al momento de enfrentar la máquina a datos clínicos que no han sido valorados previamente.

Dataset	RBF				s-Hermite				
	Parámetros	C	Gamma	nVS	Accuracy	C	n	nVS	Accuracy
Diabetic(DR)	0.156225	1.6899	998	80.20	0.96875	2	903	83.20	
Diabetes Pima	0.156250	1.4899	511	69.89	6.468750	1	404	75.06	

Tabla 1. Resultados de clasificación con kernels RBF y s-Hermite

4. DISCUSIÓN

La figura 2 muestra la implementación del kernel s-Hermite basado en los polinomios con este mismo nombre, kernel con el cual obtenemos una ventaja en la clasificación del modelo al comparar con los resultados del kernel tradicional RBF, lo cual se reporta en la tabla 1. Esta ventaja está dada por el número de vectores soporte necesarios para construir la frontera de decisión, lo cual está asociado a una mayor capacidad de generalización o sobreajuste del modelo.

5. CONCLUSIÓN

Los resultados apoyan la hipótesis planteada, esto es, el kernel s-Hermite mostró disminuir el número de vectores soporte necesarios para construir la función de decisión de las MSV. Además de este resultado favorable, se observa que la disminución de vectores soporte no degrada el accuracy en clasificación, sino que lo eleva en los dos conjuntos de datos analizados. Estos resultados podrían derivarse de que (i) s-Hermite sea una mejor opción para clasificación de Diabetes Mellitus o que (ii) el rango de parámetros analizados esté sesgado a favor de s-Hermite. Para descartar esta última opción, como trabajo futuro, se realizarán pruebas exhaustivas en un rango más amplio de parámetros usando metaheurísticas para la exploración. También se tiene contemplado aplicar la metodología propuesta a otros datasets sobre Diabetes.

5. REFERENCIAS

- [1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
- [2] Padierna, Luis & Carpio, Martín & Rojas-Dominguez, Alfonso & Puga, Héctor & Fraire-Huacuja, Héctor. (2018). A Novel Formulation of Orthogonal Polynomial Kernel Functions for SVM Classifiers: the Gegenbauer Family. Pattern Recognition. 84. 10.1016/j.patcog.2018.07.010.
- [3] Bottou,L. and Lin, C. , Support Vector Machine Solvers , https://www.csie.ntu.edu.tw/~cjlin/papers/bottou_lin.pdf (Recuperado :Junio 2019).