UNIVERSIDADE DA CORUÑA

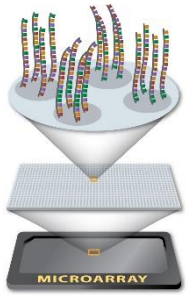
Aprendizaje en Grandes Volúmenes de Datos Mediante un Nuevo Método Distribuido y No Iterativo para Redes de Neuronas de Una Capa

Óscar Fontenla Romero
Marcelo Gómez Casal
Bertha Guijarro Berdiñas
Beatriz Pérez Sánchez

Aplicaciones Prácticas del Aprendizaje Automático

- Detección de ataques informáticos y/o spam
- Identificación de fraudes
- Sistemas de recomendación
- Clasificación de clientes
- Diagnóstico médico o de sistemas
- Análisis de estructuras complejas
- Estudio del estado de la bolsa

amazon



Problema: Tamaño de los Conjuntos de Datos

- Incremento del tamaño en dos direcciones:

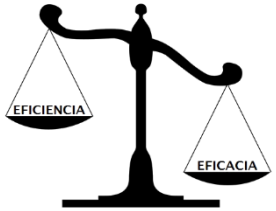
  1. Crecimiento constante de las bases de datos
     - Aumento continuo del número de muestras
     - Ejemplo: Bolsa, Industria, Sistemas de Recomendación

  2. Estudio de campos con datos complejos
     - Manejo de muestras con muchas características
     - Ejemplo: Análisis de Estructuras de ADN

- Conjuntos más grandes = Proceso de aprendizaje más lento
  - Minimizar el impacto del tamaño sobre los tiempos de entrenamiento
  - Importancia de la escalabilidad en los métodos de Aprendizaje Automático

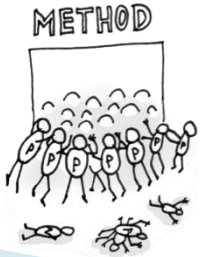
Problema: La Escalabilidad de los Métodos

- Uso de algoritmos iterativos
- Se centran en la eficacia, escalabilidad relegada a segundo plano



- No se explota la posibilidad de distribución

- Ajuste exhaustivo parámetros para maximizar eficacia



Propuesta de un Nuevo Método: El LANN-DSVD

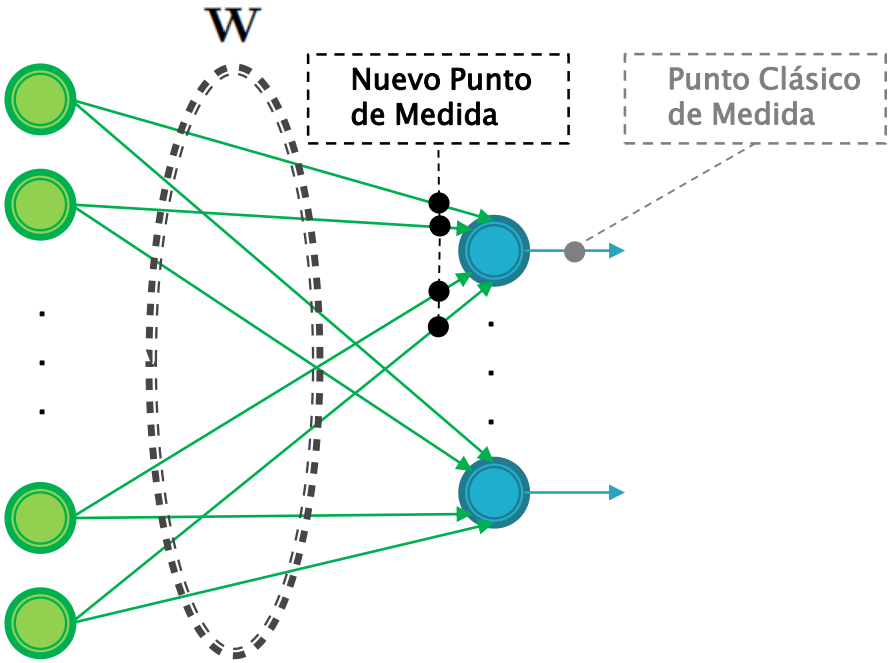
- Red de neuronas sin capa oculta: Tantas entradas como atributos

- Aprendizaje no iterativo gracias a:
  - Nuevo punto de medida de error (antes de función de activación)
  - Factorización de matrices SVD (Singular Value Decomposition)

SVD

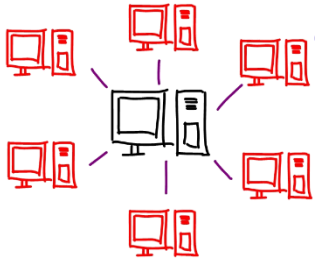
$$\mathbf{W} \approx \mathbf{U}(\mathbf{S}\mathbf{S}^T)^{\dagger}\mathbf{U}^T\mathbf{X}\mathbf{F}\mathbf{F}\bar{\mathbf{d}}$$

Medida del Error

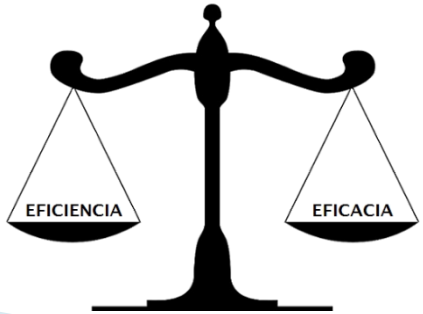


Propuesta de un Nuevo Método: El LANN-DSVD

- Aprendizaje distribuido e incremental



- Buena escalabilidad
  - No iterativo
  - No precisa ajuste de parámetros por parte del usuario
  - Equilibrio entre eficacia y eficiencia



Validación: SVM vs. LANN-DSVD

- Se plantean cinco escenarios de clasificación

Conjunto	Nº de Muestras	Nº de Atributos
Breast	683	10
MiniBooNE	130.064	50
Susy	5.000.000	18
HearthStone	2.000.000	44
Higgs	11.000.000	28

  - Comparación de eficacia y eficiencia
    - ¿Qué método resuelve mejor el problema?
    - ¿Qué método termina antes el entrenamiento?

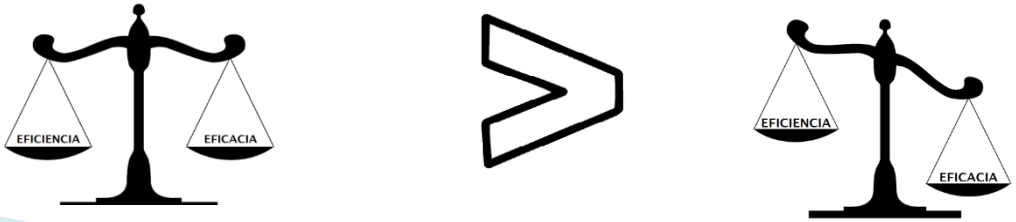
Validación: SVM vs. LANN-DSVD

	Eficacia (% AUC)		Eficiencia (s)	
Conjunto	SVM	LANN-DSVD	SVM	LANN-DSVD
Breast	96,66 ± 2,37	99,5 ± 0,50	0,034 ± 0,04	0,18 ± 0,01
MiniBooNE	89,54 ± 0,34	95,50 ± 0,20	383,35 ± 35,52	0,37 ± 0,01
Susy	Más de 1 semana sin terminar	83,60 ± 0,00	Más de 1 semana sin terminar	3,66 ± 0,21
HearthStone	Más de 1 semana sin terminar	78,60 ± 0,70	Más de 1 semana sin terminar	4,51 ± 0,11
Higgs	Más de 1 semana sin terminar	68,30 ± 0,00	Más de 1 semana sin terminar	12,05 ± 0,42

Conjunto	Método de Referencia	Eficacia de Referencia
Susy	Deep Neural Network	87,9 ± 0,1
HearthStone	(No Revelado)	80,18 ± ¿?
Higgs	Deep Neural Network	88,50 ± 0,2

Conclusiones

- En general la precisión es comparable a la de los métodos más eficaces
- Para conjuntos grandes, el LANN-DSVD es más rápido
  - Idóneo para conjuntos con muchas muestras y/o atributos
- Importancia del equilibrio entre eficacia y eficiencia
  - Las SVM no terminan a tiempo en el caso de los conjuntos más grandes
- Aprendizaje absolutamente automático
  - LANN-DSVD no precisa de un ajuste de parámetros por parte del usuario
- Permite aprendizaje compartido preservando la privacidad
  - Entre varias entidades, transmitiendo solo el modelo (vector de pesos)

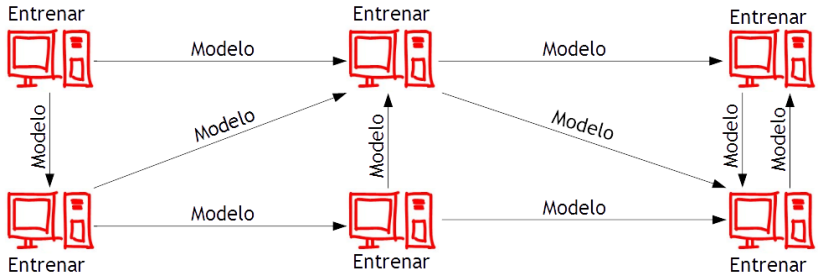


Aplicaciones del LANN-DSVD

- Posible entrenar muchos modelos sin necesidad de un supercomputador
  - Cualquier máquina de hoy en día soporta 4 u 8 hilos de ejecución

- Escenarios que precisen de aprendizaje en tiempo real
  - No entrenar el modelo con todas las muestras al querer añadir nuevas
  - Ejemplo: Sistemas de Recomendación

- Aprendizaje compartido entre diferentes entidades
  - Distribuir el entrenamiento para aprender con datos de diferentes fuentes
  - Al contar con más datos, todas las entidades implicadas se benefician
  - Se preserva la privacidad de los datos originales
  - Ejemplo: Bolsa, Medicina



UNIVERSIDADE DA CORUÑA

Aprendizaje en Grandes Volúmenes de Datos Mediante un Nuevo Método Distribuido y No Iterativo para Redes de Neuronas de Una Capa

Óscar Fontenla Romero
Marcelo Gómez Casal
Bertha Guijarro Berdiñas
Beatriz Pérez Sánchez