Optimización de ReliefF para entornos IoT con recursos limitados *

Samuel Suárez-Marcote, Laura Morán-Fernández y Verónica Bolón-Canedo

Universidade da Coruña, CITIC, A Coruña, España
{s.suarez.marcote, laura.moranf, veronica.bolon}@udc.es

Resumen

Ante el aumento actual del volumen de datos, en este trabajo se proponen dos mejoras de ReliefF diseñadas para la selección de características en entornos con recursos limitados. La primera reemplaza la aritmética en punto flotante por punto fijo, reduciendo tanto el consumo de memoria como energético, disminuyendo hasta una cuarta parte los bits necesarios mientras mantiene resultados comparables al uso de 64 bits. La segunda añade un paso de *clustering* para reducir el número de muestras y acelerar la ejecución sin perder información. Evaluadas en nueve conjuntos de datos con diversas propiedades, las optimizaciones ofrecen un equilibrio óptimo entre eficiencia y rendimiento, haciendo de ReliefF una solución útil para la implementación en sistemas IoT con restricciones energéticas.

1. Introducción

Ante el aumento del número de dispositivos de Internet de las Cosas (*IoT*, por sus siglas en inglés), el volumen de datos generado a diario ha aumentado enormemente. Por ello, surge la necesidad de eliminar aquellos datos irrelevantes para poder extraer información de calidad. Una buena forma de abordar dicho problema es recurrir a los métodos de selección de características. Sin embargo, ante las limitaciones de memoria, energía y capacidad de cómputo de dichos dispositivos, es necesario adaptar estos métodos a entornos restringidos, resultando también en un mejor uso del ancho de banda disponible. ReliefF [1] es un método de selección de características de tipo filtro capaz de detectar dependencias de alto orden entre características, aunque su coste computacional escala de manera cuadrática con el número de muestras. Por estas razones, se proponen dos mejoras para la adaptación de ReliefF a entornos embebidos.

2. Método propuesto

La primera de las mejoras busca adaptar ReliefF a los dispositivos *IoT* mediante la reducción de la precisión numérica. Para ello, se convirtieron sus operaciones internas a punto fijo. De esta manera, operaciones como el cálculo de distancias fueron reformuladas usando aritmética de enteros, permitiendo operaciones más eficientes al no recurrir a una unidad de punto flotante.

En la segunda, se recurre a un paso de *clustering* acumulativo para la reducción del número de muestras, generando prototipos representativos que reemplazan a grupos de instancias similares. De esta forma, se reduce el número de cálculos de distancia realizados por ReliefF, mitigando la complejidad a la vez que permite balancear el conjunto de datos.

*Este trabajo ha sido financiado por la Xunta de Galicia/FEDER (ED431C 2022/44); Ministerio de Ciencia e Innovación MCIN/AEI/10.13039/501100011033 y "Next-GenerationEU"/PRTR [PID2023-147404OB-I00, TED2021-130599A-I00]; Ministerio para la Transformación Digital y de la Función Pública (TSI-100925-2023-1) y Programa de axudas á etapa predoutoral, Xunta de Galicia (ED481A 2023/034). CITIC, como parte de la Red de Centros de Investigación del Sistema Universitario de Galicia, está financiado por fondos ERDF Funds y por la "Secretaría Xeral de Universidades" (Grant ED431G 2023/01).

3. Resultados

Se emplearon un total de nueve conjuntos de datos, cinco de ellos microarrays con un alto número de características. Para la evaluación de cada una de las mejoras propuestas se definieron dos pasos: estudio de la similaridad de los *rankings* frente a la versión de ReliefF en doble precisión mediante la Tasa de Verdaderos Positivos (TVP) y evaluación de la información retenida por las características seleccionadas en un paso de clasificación con una SVM Lineal.

Como muestra la Tabla 1, las configuraciones de 32 bits obtuvieron *rankings* casi prácticamente idénticos a la versión de 64 bits. En el caso de la representación Q12.4, esta ofreció un balance favorable entre preservación de los *rankings* y el ahorro de recursos. Las configuraciones de 8 bits mostraron mayores diferencias en TVP. Aún así, en muchos casos, no se tradujeron en pérdidas de información en el paso posterior de clasificación.

Aunque mostró mayores diferencias con respecto a los *rankings* obtenidos, como se puede ver en la Figura 1, el paso de *clustering* redujo los tiempos de ejecución y el consumo energético drásticamente, destacando los conjuntos de datos con gran número de instancias, mientras que la exactitud de clasificación permaneció sin diferencias significativas en la mayoría de los casos.

Conjunto de datos	$Q16.16$	$Q24.8$	$Q8.8$	$Q12.4$	$Q4.4$	$Q6.2$
congress	0.975 ± 0.06	0.975 ± 0.06	0.875 ± 0.00	0.975 ± 0.06	0.925 ± 0.07	0.875 ± 0.00
parkinsons	0.975 ± 0.06	0.975 ± 0.06	0.416 ± 0.03	0.975 ± 0.06	0.657 ± 0.14	0.457 ± 0.04
waveform	1.000 ± 0.00	1.000 ± 0.00	0.182 ± 0.03	0.875 ± 0.00	0.262 ± 0.04	0.274 ± 0.06
connect4	1.000 ± 0.00	1.000 ± 0.00	0.233 ± 0.07	0.273 ± 0.05	0.242 ± 0.06	0.215 ± 0.09
colon	1.000 ± 0.00	0.964 ± 0.02	0.367 ± 0.08	0.849 ± 0.06	0.023 ± 0.01	0.000 ± 0.00
SRBCT	1.000 ± 0.00	0.989 ± 0.01	0.000 ± 0.00	0.890 ± 0.04	0.015 ± 0.01	0.008 ± 0.01
DLBCL	1.000 ± 0.00	0.984 ± 0.01	0.538 ± 0.05	0.839 ± 0.04	0.008 ± 0.01	0.004 ± 0.00

Tabla 1: Media y desviación típica de TVP para las variantes de ReliefF en punto fijo frente a la versión en doble precisión. $Qm.n$ muestra la representación con m bits enteros y n fraccionarios.

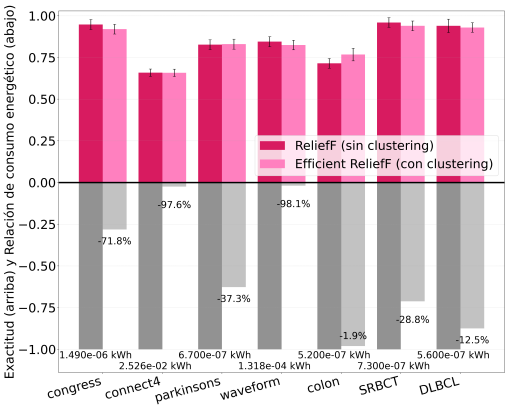

Figura 1: Exactitud de clasificación y relación del consumo energético de las versiones optimizadas de ReliefF con *clustering* frente a la línea base utilizando un clasificador SVM Lineal.

4. Conclusiones

En este trabajo se ha propuesto la adaptación de ReliefF para su ejecución en dispositivos IoT mediante la combinación de la aritmética de punto fijo y *clustering*, reduciéndose el consumo energético y tiempo de ejecución a la vez que se mantiene la efectividad del método.

Referencias

[1] Igor Kononenko. Estimating attributes: Analysis and extensions of relief. In *European conference on machine learning*, pages 171–182. Springer, 1994.