UNIVERSIDAD DE CUENCA

Text Mining

An improvised sine cosine algorithm to select features for text categorization

Journal of King Saud University

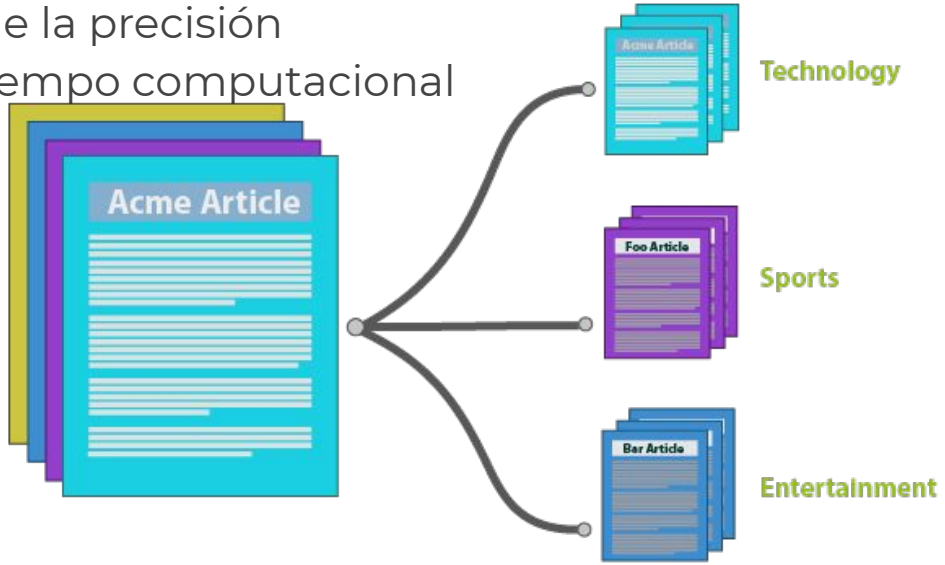
Miguel Á. Macías & **Jonnathan** Campoberde & **Moises** Arévalo

Contenido

1. Contexto del problema
2. Estado del Arte
3. Problemas que ataca la investigación
4. Metodología utilizada
5. Resultados y Discusiones
6. Conclusiones

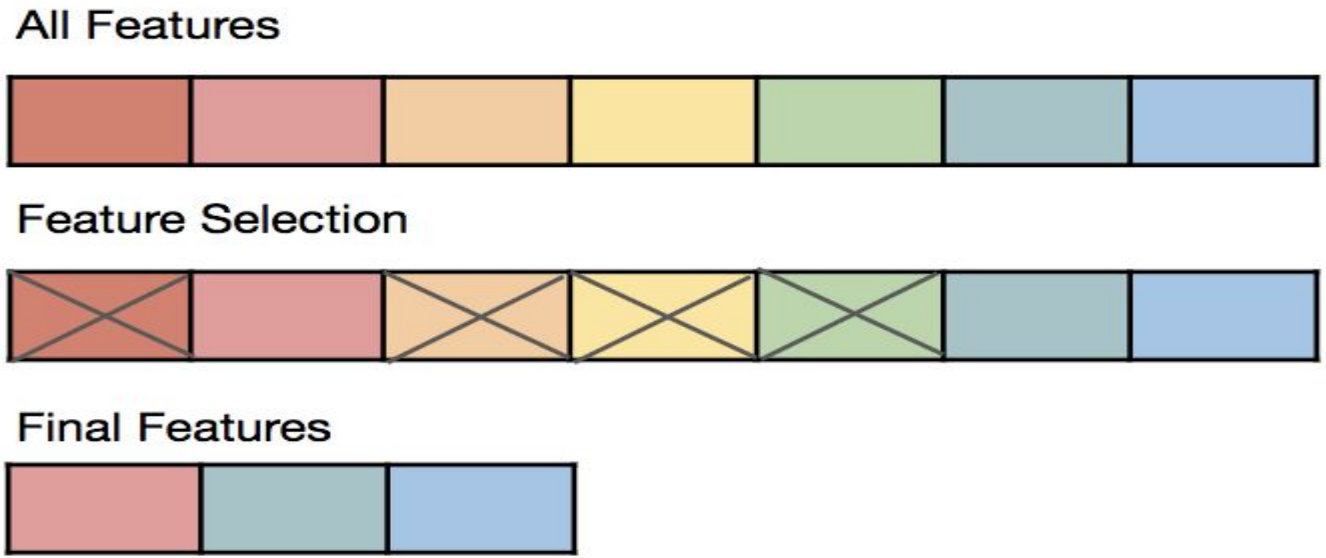
1 Contexto del problema

- **Categorización de texto (TC)**
  - La creciente cantidad de documentos electrónicos disponibles en la actualidad hace necesario métodos de organización automáticos.
    - Bag of Words(BoW): Cada documento está representado por un vector de términos.
      - Alta dimensionalidad
      - Degradación de la precisión
      - Aumento de tiempo computacional



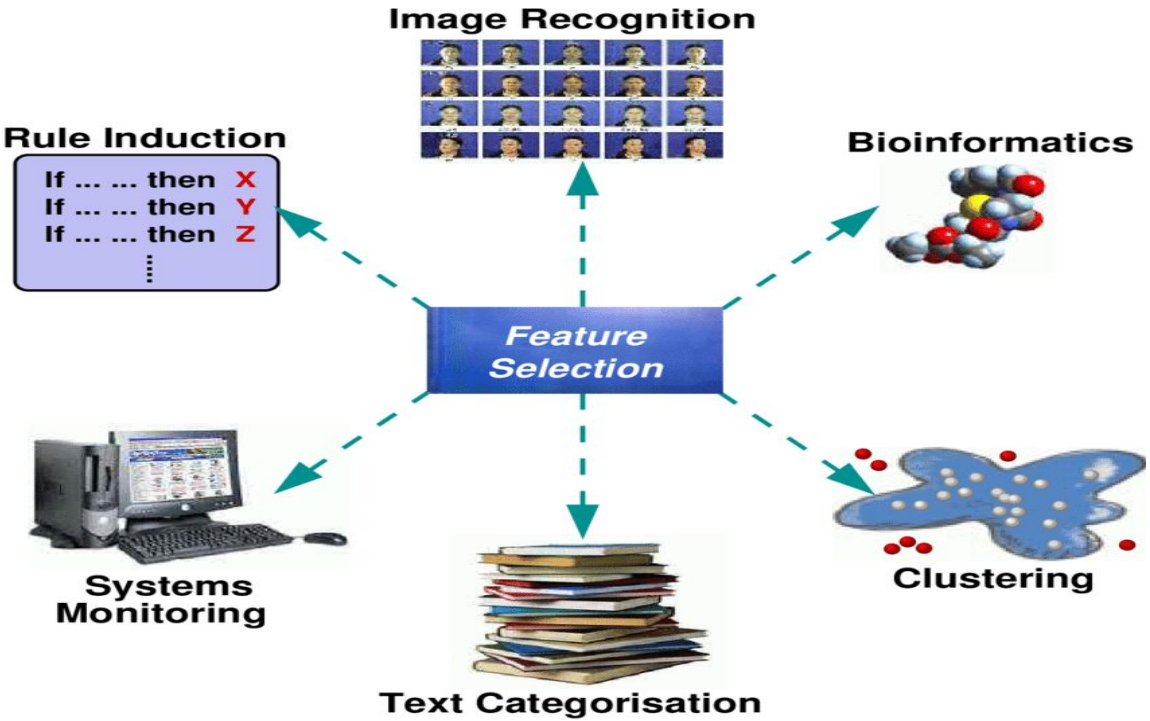
1 Contexto del problema

- **Selección de características(Feature Selection)**
  - Paso importante en el preprocesamiento de TC para superar este problema.
  - Determinar un subconjunto que contiene un número limitado de características relevantes.



1 Contexto del problema

- **Selección de características(Feature Selection)**
  - Algunas de las aplicaciones que se tiene con la selección de características son:



1 Contexto del problema

- **Selección de características(Feature Selection)**
  - Los métodos de FS se pueden categorizar en tres enfoques basados en el procedimiento de evaluación:
    - ***Filtros***: Cada característica se evalúa individualmente utilizando una medida estadística y se seleccionan las características con las puntuaciones más altas.
    - ***Envoltura:*** Se basa en un algoritmo de búsqueda y en un criterio de medida. En este caso la función objetivo es el desempeño del algoritmo de clasificación.
    - ***Enfoques integrados:*** La selección de características y los algoritmos de aprendizaje están intercalados. Los filtros son incluidos en la máquina de aprendizaje del algoritmo de envoltura.

2 Estado del arte

Año	Breve Descripción
1997	Los autores (Yang y Pedersen) comparó varios métodos como Frecuencia de documentos (DF), Ganancia de información (IG), Información mutua (MI), Chi-cuadrado (X^2). Esta comparación ha demostrado que IG y X^2 son los más eficaces.
2009	Aghdam aplicó un método de selección de características basado en el algoritmo genético, colonias de hormigas para la categorización de texto.
2016	Algoritmo seno-coseno (SCA) es un enfoque de optimización global que utiliza un modelo matemático basado en funciones seno y coseno para actualizar iterativamente un conjunto de soluciones candidatas.
2017	Implementación de SCA para la detección de galaxias mediante recuperación de imágenes. Además de implementar un SCA mejorado para problemas de optimización en general.
2017	Sindhu ha aplicado una versión mejorada de SCA con una estrategia de elitismo para la selección de características en el aprendizaje automático. En esta variante, muestran un avance de SCA sobre los otros métodos de búsqueda como PSO (Optimización de Enjambre de partículas) y GA (Algoritmos Genéticos).

3 Problemas que ataca la investigación

Selección de características

- Para la categorización de texto.
- Eficiencia del método de envoltura para la selección de características. En este caso ISCA(Algoritmo de seno coseno mejorado) se combina con el filtro de ganancia de información para superar el problema de la enorme dimensión.

4 Metodología

Preliminares

SCA algorithm

El algoritmo seno-coseno es un algoritmo estocástico iterativo reciente donde, en cada iteración, las soluciones se actualizan en función de las funciones seno o coseno.

$$X(i,j)_{t+1} = X(i,j)_t + r1 * \sin(r2) * \left| r3P(j)_t - X(i,j)_t \right| \quad (1)$$

$$X(i,j)_{t+1} = X(i,j)_t + r1 * \cos(r2) * \left| r3P(j)_t - X(i,j)_t \right| \quad (2)$$

dónde X(i,j) es la posición de la solución i en la j-ésima dimensión de la t-ésima iteración, P(j) es la posición del mejor individuo en la jth dimensión en la t-ésima interacción

4 Metodología

Preliminares

SCA algorithm

Las dos ecuaciones anteriores son combinadas dentro de la siguiente ecuación:

$$X(i,j)_{t+1} = \begin{cases} X(i,j)_t + r1 * \sin(r2) * \left| r3P(j)_t - X(i,j)_t \right| if r4 < 0.5 \\ X(i,j)_t + r1 * \cos(r2) * \left| r3P(j)_t - X(i,j)_t \right| if r4 \geq 0.5 \end{cases} \quad (3)$$

r4: Cambia igualmente entre componentes seno y coseno en la ecuación.

4 Metodología

I) Objetivo

Realizar una mejora al algoritmo de seno coseno(SCA) para feature selection

II) Problemática del SCA:

- Tiende a quedarse atascado en regiones subóptimas y esto se refleja en el esfuerzo computacional requerido para obtener el mejor rendimiento.
- La razón depende de un límite en la exploración del espacio de búsqueda, ya que existe una limitación en la actualización de la solución hacia afuera con respecto a la solución encontrada.

4 Metodología

III) Forma de abordar el problema:

- Lo que tratan de hacer es que el algoritmo pueda referirse a otras soluciones en otras regiones del espacio de investigación, obtiene más probabilidades de descubrir regiones destacadas, y evitar la convergencia hacia óptimos locales.

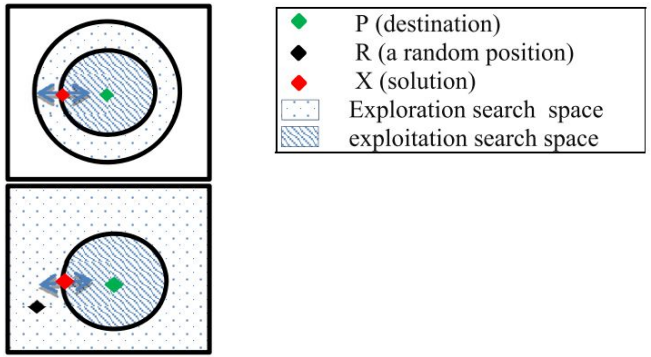


Fig. 2. Exploration and exploitation of algorithms SCA and ISCA, Sine Cosine Algorithm on the top; Improved Sine Cosine Algorithm on the bottom.

4 Metodología

IV) Ecuación:

$$X(i,j)_{t+1} = (1 - r1) * X(i,j)_t + r1 * R$$

- Donde
  - X(i,j) es la posición de la solución actual i, en la j-esima dimensión en la t-esima iteración.
  - R es una posición aleatoria incluida en el espacio de búsqueda
  - Los pesos (1-r1) y (r1) representa la importancia relativa de la posición actual y la posición aleatoria R respectivamente

4 Metodología

V) Realiza modificaciones en el sistema de ecuaciones (3):

$$X(i,j)_{t+1} = \begin{cases} X(i,j)_t + r1 * \sin(r2) * \left| r3P(j)_t - X(i,j)_t \right| if r4 < 0.5 \\ X(i,j)_t + r1 * \cos(r2) * \left| r3P(j)_t - X(i,j)_t \right| if r4 \geq 0.5 \end{cases} \quad (3)$$

- Proponen reducir las dos fórmulas de seno y coseno en una sola

$$X(i,j)_{t+1} = \begin{cases} (1 - r1) * X(i,j)_t + r1 * R & if\ c < r1 \\ P(j)_t + r1 * \sin(r2) * \left| r3 * P(j)_t - X(i,j)_t \right| & if\ c \geq r1 \end{cases} \quad (6)$$

4 Metodología

- Para alcanzar un buen equilibrio entre intensificación y diversificación, introducen una nueva ecuación que depende de las variables aleatorias c y r1
- Donde c = b*r4, b es entero constante [1,5]

$$X(i,j)_{t+1} = \begin{cases} (1-r1) * X(i,j)_t + r1 * R & if\ c < r1 \\ P(j)_t + r1 * \sin(r2) * \left| r3 * P(j)_t - X(i,j)_t \right| & if\ c \geq r1 \end{cases} \quad (6)$$

4 Metodología

VI) Pseudocódigo

Algorithm 2: ISCA (nb_agent,dimension_size, max_iteration, a,b)

1. Initialize positions of all agents (X).
2. Calculate the cost function of each agent (Fit), and select the best position's agent (Best_pos).
3. T ← 2;**WHILE** *(t <= max_iteration)* **DO** $r_1 \leftarrow a - t^*(a/max\_iteration)$; % decreasing r1 iteratively **FOR** each agent *(X)*, from the dimension *(j)* **DO** $r_2 \leftarrow (2^*pi)^*rand()$; $r_3 \leftarrow rand()$; $r_4 \leftarrow b^*rand()$; % updating of the position of the solution **IF** $(r_4 < r_1)$ *m = Get_rand_position();* %get a random position from the search space *[lb,ub]* $X(j)=(1-r_1)^* X(j)+(r_1)^*m$; **ELSE** $X(j) = Best\_pos\ (j) + r_1^*(\sin(r_2))^*abs(r_3^* Best\_pos\ (j)-X(j))$; **ENDIF** % end of the updating solution's position **END FOR**
1. Calculate the new cost function (*Fit*) of each agent (*X*), and get the best position's agent (Best_pos).
2. Increment (*t*);**END WILEEND ALGORITHM**

4 Metodología

VII) Representación del subconjunto de características

- Una solución potencial(feature subset) está representada por un feature vector donde cada característica corresponde a una dimensión y cada variable fija a un rango dentro de [0,1].
- Para saber si se selecciona o se rechaza debe cumplir la condición:
  - si el valor de la posición es mayor o igual a .5, se selecciona, caso contrario se descarta.

$$f_{i,j} = \begin{cases} 1 & if\ Xi,j \geq 0.5 \\ 0 & otherwise \end{cases}$$

4 Metodología

VIII) Evaluación de las características de un subconjunto.

- En cada generación, la prioridad de un subconjunto de características se clasifica de acuerdo a sus valores de aptitud.
- Donde fit es la función de aptitud de ISCA maximiza la precisión e intentar alcanzar un subconjunto corto.
- Para esto combinan el número de características y la tasa de error en una fórmula.

$$fit = \alpha * Error\_Rate + (1 - \alpha) * \left(\frac{number\ of\ features\ selected}{total\ features}\right) \quad (8)$$

4 Metodología

VIII) Evaluación de las características de un subconjunto.

- Donde
  - α es una constante entre 0 y 1.
  - (1 - α) la importancia relativa del número de características seleccionadas.

$$fit = \alpha * Error\_Rate + (1 - \alpha) * \left(\frac{number\ of\ features\ selected}{total\ features} \right) \quad (8)$$

4 Metodología

VIII) Evaluación de las características de un subconjunto.

- La medida F1 del clasificador Naive Bayes, se usa para calcular la tasa de error involucrada en (8)
- Aplican un método de prueba de validación cruzada 3fols del conjunto de datos de entrenamiento.
- Con métodos empíricos, se establece en .9, ya que su rango de valor es bueno más allá de ese valor.

4 Metodología

IX) Estudio Experimental

- Este método a sido evaluado con una gran cantidad de documentos. Y para evaluar el desempeño lo comparan con los siguientes algoritmos:
  - ISCA, GA, ACO, SCA y OBL-SCA.
- Utiliza Naive bayes, para la tarea de categorización.

4 Metodología

IX) Estudio Experimental

- Para el estudio hacen uso de 9 colecciones de texto:

Re0	**La1s**
La2s	**Oh0**
Oh5	**Oh10**
Oh15	**FBIS**
Tr41	

4 Metodología

IX) Estudio Experimental

- Para la evaluación de los métodos hacen usos de métricas como:
  - Precisión
  - Recall
  - F-measure
  - Promedio Macro
  - Promedio Micro
  - Micro F1
  - Macro F1

5 Resultados y discusiones

Parámetros comunes: Population size = 30 Max. iteration = 50

SCA, OBL_SCA, Weighted_SCA

Coeficiente **a = 2**

ACO

Initial pheromone = 1
α = 1.5
β = 0.1
σ = 0.2

Levy_SCA

Coeficiente **a = 2**
α = 1.5

GA

Crossover probability = 0.7
Mutation probability = 0.05

MFO

Coeficiente **b = 1**

ISCA

Coeficiente **a = 1**
Coeficiente **b = 8**

5 Resultados y discusiones

Promedios de Precision (P), recall (R) y F-score (F1)

Collection	Wgh-ISCA			Levy-SCA			ISCA			Obl_SCA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Re0	65,09	63,28	62,66	72,49	72,04	70,78	68,67	71	89	65,89	67,37	65,02
La1s	70,29	72,44	70,64	73,35	75,84	73,91	76,13	76,88	76,1	72,14	74,87	72,6
La2s	72,98	74,43	72,95	76,59	77,03	76,4	78,23	79,22	78,18	72,98	75,87	73,33
Oh0	70,21	79,39	72,45	77,07	85,36	79,37	78,85	87,19	81,52	72,18	81,09	74,2
Oh5	71,05	75,82	72,08	80,09	86,16	81,88	81,59	86,82	83,42	72,38	78,8	73,83
Oh10	65,36	70,82	66,6	70,95	76,98	72,42	73,05	78,41	74,48	66,02	72,29	67,34
Oh15	66,89	74,85	68,87	73,76	80,45	75,62	76,67	83,92	79,02	68,56	75,57	70,2
Fbis	74,9	73,05	73,2	80,58	79,24	79,3	81,39	80,01	80,2	75,26	73,97	73,9
Trec41	83,34	80,83	81,08	86,79	84,67	85,17	88,23	86,49	86,9	83,35	82,03	81,95
AVG	71,12	73,88	71,17	76,85	79,75	77,20	78,09	81,10	80,98	72,08	75,76	72,49

5 Resultados y discusiones

Promedios de Precision (P), recall (R) y F-score (F1)

Collection	MFO			SCA			ACO			GA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Re0	71,09	68,5	68,79	68,61	68,09	67,01	66,83	66,23	65,33	71,09	69,37	68,79
La1s	73,72	76,02	74,33	71,32	75,2	72,08	71,06	72,9	71,42	74,69	77,16	75,4
La2s	76,77	78,21	76,89	73,54	74,92	73,52	72,62	75,37	72,99	76,9	78,52	77,11
Oh0	76,91	86,33	79,67	72,79	80,97	74,44	71,05	79,23	73,06	76,95	84,98	79,11
Oh5	78,89	86,46	80,51	74,41	81,59	75,83	71,4	78,13	72,81	81,17	85,76	82,58
Oh10	71,14	78,35	72,87	68,56	74,07	69,69	64,68	70,34	65,75	72,23	78,14	73,63
Oh15	73,85	82,25	76,31	69,34	77,92	71,74	67,44	74,35	68,91	74,55	82,2	76,99
Fbis	80,96	79,39	79,58	76,39	75,43	75,2	74,25	73,1	72,73	81,58	79,93	80,24
Trec41	87,17	85,88	86,02	85,65	84,13	83,8	84,56	81,66	82,47	87,06	86,04	85,93
AVG	76,72	80,15	77,22	73,40	76,92	73,70	71,54	74,59	71,72	77,36	80,23	**77,75**

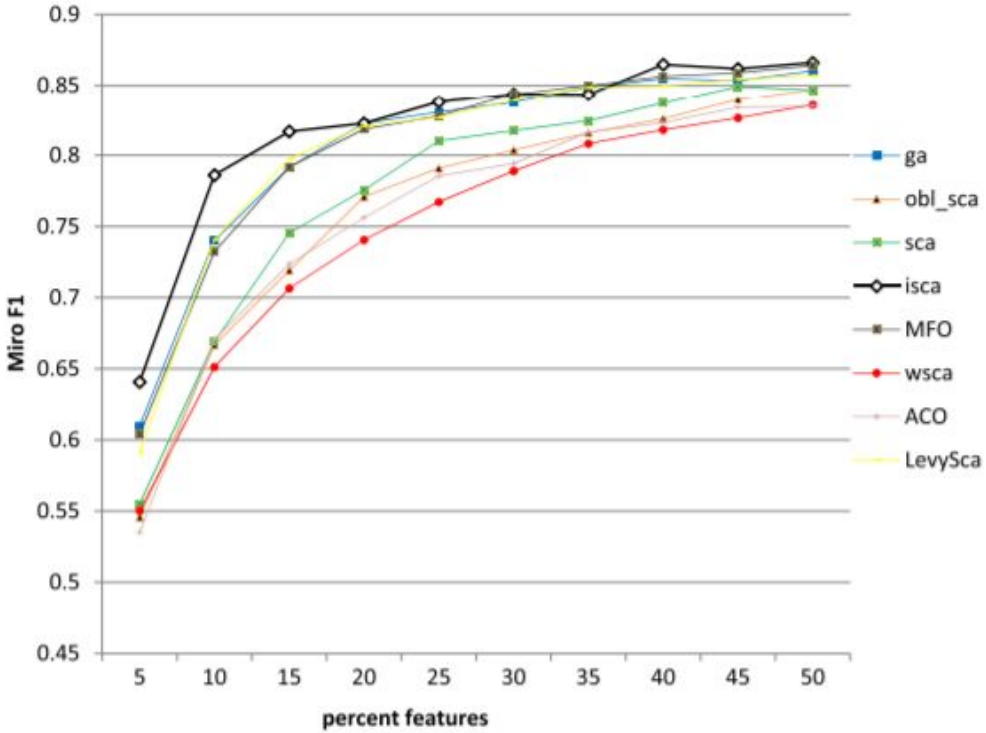
5 Resultados y discusiones

Promedios de Precision (P), recall (R) y F-score (F1)

	P	R	F1
Weighted_SCA	71.12	73.88	71.17
Levy_SCA	76.85	79.75	77.20
ISCA	**78.09**	**81.10**	**80.98**
OBL_SCA	72.08	75.76	72.49
MFO	76.72	80.15	77.22
SCA	73.40	76.92	73.70
ACO	71.54	74.59	71.72
GA	77.36	80.23	77.75

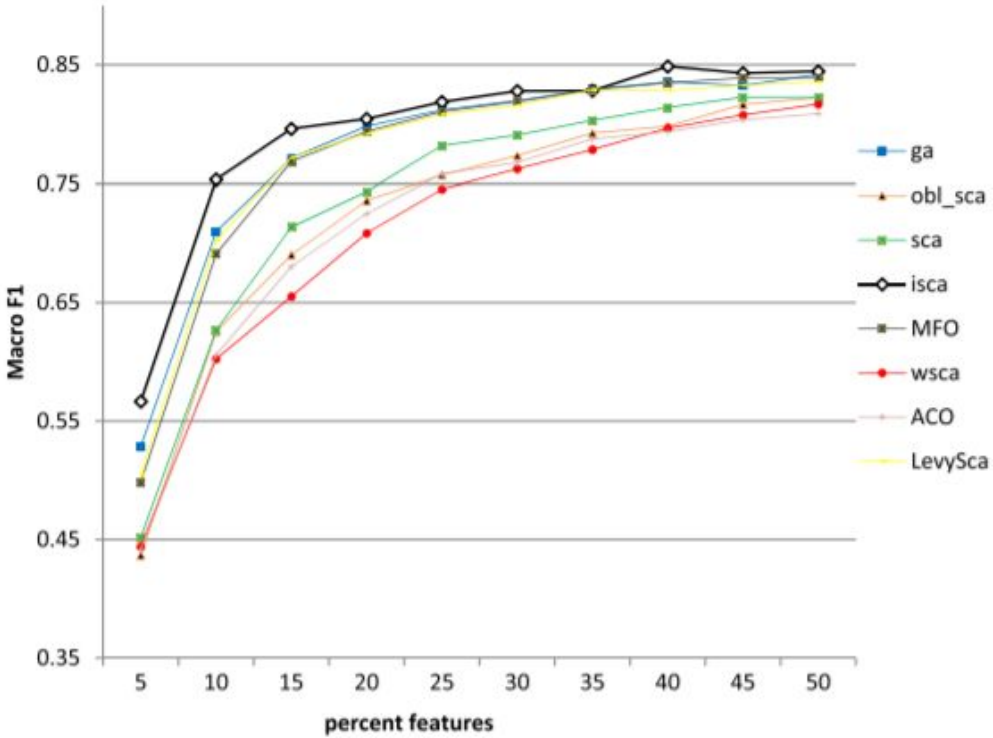
5 Resultados y discusiones

Average microF1-macroF1 scores



5 Resultados y discusiones

Average microF1-macroF1 scores



5 Resultados y discusiones

Rendimiento de ISCA vs filter methods

Collection	ISCA Search Algorithm			IG Attr. Eval.			Correlation Attr. Eval.			Significance Attr. Eval.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Re0	82,52	80,28	81,08	70,95	63,10	65,36	50,91	48,98	48,98	48,55	52,54	48,75
la1s	81.40	81.37	81.38	69.04	68.87	68.36	62.07	65.63	62.53	40.56	67.33	40.24
La2s	78,56	77,63	77,81	75,15	74,21	74,27	68,59	67,23	66,63	43,90	67,26	45,34
oh0	82,82	86,52	84,11	80,27	83,25	81,12	64,42	65,15	63,55	63,46	82,02	66,74
Oh5	84,06	85,93	84,71	79,37	80,71	79,75	76,93	80,99	78,09	65,25	90,49	69,83
oh10	74,16	77,69	75,56	70,57	73,50	71,27	69,58	72,15	70,34	55,70	71,60	56,71
Oh15	79,16	84,39	81,12	76,16	79,66	77,28	73,38	76,33	74,46	64,72	80,93	66,14
Fbis	85,86	82,80	83,66	74,40	65,25	65,04	12,86	9,92	9,00	8,21	7,92	7,39
Trec41	92,34	89,39	90,56	74,40	65,25	65,04	64,42	57,16	53,73	59,33	79,84	63,47
Avg	82,32	82,89	82,22	74,48	72,64	71,94	60,35	60,39	58,59	49,96	66,66	51,62

	P	R	F1
ISCA avg	82.32	82.89	82.22

5 Resultados y discusiones

Wilcoxon's Nonparametric statistical test

Collection	ISCA Algorithm Vs						
	GA	levySca	OblSca	Sca	WgSca	MFO	ACO
Re0	0.0340	0.0901	0.0000	0.0001	0.0000	0.0266	0.0000
la1s	0.7369	0.5673	0.2145	0.3556	0.0810	0.9707	0.1273
La2s	0.6746	0.5392	0.0876	0.1112	0.0956	0.5426	0.0447
Oh0	0.8719	0.1400	0.0001	0.0002	0.0000	0.3864	0.0000
oh5	0.8786	0.9698	0.0000	0.0000	0.0000	0.3381	0.0000
Oh10	0.8556	0.5609	0.0046	0.0704	0.0012	0.6113	0.0016
oh15	0.2678	0.3481	0.0000	0.0005	0.0000	0.1862	0.0000
FBIS	0.9270	0.4643	0.0050	0.0743	0.0023	0.7085	0.0018
Trec41	0.9348	0.8603	0.0904	0.2008	0.0076	0.6761	0.1234

6 Conclusiones

ISCA superó ampliamente al SCA original, incluidos sus sucesores

- ISCA no solo tiene un buen desempeño, posee una fórmula matemática simple para actualizar las posiciones de la solución
- La configuración de pocos parámetros de ISCA lo hacen flexible y fácil de adaptar a la mayoría de problemas de búsqueda
- ISCA se puede combinar con otros algoritmos de búsqueda para obtener un mejor rendimiento

Referencias

Belazzoug, M., Touahria, M., Nouioua, F., & Brahimi, M. (2020). An improved sine cosine algorithm to select features for text categorization. Journal of King Saud University - Computer and Information Sciences, 32(4), 454–464. https://doi.org/10.1016/j.jksuci.2019.07.003

¡Gracias!