A greedy set cover approach for the recomputation of the cluster centres in the FPAC Algorithm

Ivan Feliciano[1] and Edgar Hernandez-Gonzalez[2]

[1] National Institute of Astrophysics, Optics and Electronics
ivan.felavel@gmail.com
[2] National Institute of Astrophysics, Optics and Electronics
edgarmoy.28@gmail.com

Abstract. En este trabajo presentamos una modificación al algoritmo K-means usando una heurística que permite hacer el recalculo de centroides de una manera diferente al k-means tradicional. El procedimiento se basa en... primero, después, por ultimo.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

1.1 A Subsection Sample

Actualmente el numero de documentos en la web aumenta rápidamente, por tal motivo se necesitan algoritmos capaces de agrupar automáticamente grandes cantidades de datos. K-means es un algoritmo de agrupamiento, su objetivo es particionar un conjunto de datos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo. El algoritmo consta de tres pasos: 1. Inicialización: una vez escogido el número de grupos, k, se establecen k centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente. 2. Asignación objetos a los centroides: cada objeto de los datos es asignado a su centroide más cercano. 3. Actualización centroides: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo. Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso. A pesar de que el algoritmo K-means es muy popular no es escalable para datos de gran tamaño y dimensión. El principal cuello de botella de K-means es asignar cada vector no centroide a un grupo.

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Table 1. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 1 gives a summary of all heading levels.
Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

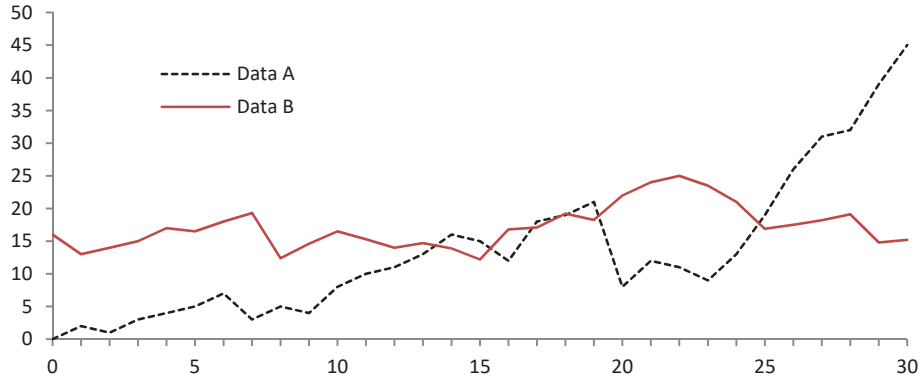

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], and a homepage [5]. Multiple citations are grouped [1–3], [1, 3–5].

2 Related Work

Se han hecho varias aportaciones para mejorar el algoritmo k-means. [la del articulo] desarrollo un algoritmo de partición rápida basado en una heurística de los vecinos mas cercanos. Dado un conjunto de centroides, evitar el calculo de distancia por pares entre vectores para obtener una partición de la colección, en su lugar se ocupó una asignación basada en el vecino mas cercano de cada centro, para esto se utilizo una lista invertida de vectores dispersos. También se evito el costoso calculo del centroide verdadero de cada grupo, se propuso una heurística para elegir el centroide de manera eficiente. En [9] los autores utilizan la heurística más lejana primero que consiste en seleccionar los centroides iniciales y evitar los cálculos de distancia redundante desde los no centroides a los centros. En [12, 19 y 20] se utilizaron estructuras de datos de partición de espacio como kd-trees. Esto aumenta la eficacia de k-means, pero solo para pocas dimensiones En [23] reasignaciones de clúster ocurren frecuentemente para puntos que no están cerca de los centroides. Identifica estos puntos al agrupar puntos vecinos usando múltiples arboles de partición. En [2 y 18] se utiliza una heurística la cual elije aleatoriamente el primer centroide y utiliza una distribución de probabilidad. En [4] se utilizo k-means escalable en el cual cada centroide se ve como una consulta para recuperar una lista de documentos que luego se asignan a ese grupo sin cálculo de distancia. [2] k-means ++.

3 Proposed Solution

4 Experiments

5 Results

6 Conclusions and future work

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017