ISSN 1870-4069

RENTOL: Un algoritmo de agrupamiento basado en K-means

Erendira Rendon Lara, Itzel María Abundez Barrera

Instituto Tecnológico de Toluca, División de Estudios de Posgrado e Investigación, México

erendir@prodigy.net.mx erendonl@toluca.tecnm.mx, iabundezb@toluca.tecnm.mx

Resumen. Sin lugar a duda el algoritmo K-means es el más utilizado en la comunidad de aprendizaje no supervisado. Desafortunadamente es muy sensible a la selección de los centroides iniciales. Debido a ello, se han propuesto un gran número de métodos para la selección de los centros iniciales. En este artículo se presenta un algoritmo de agrupamiento que tiene como base al algoritmo K-means, en el cual se implementa un nuevo método para la selección de centros iniciales. Para evaluar los resultados se utilizaron índices internos de validación de agrupamiento. Los resultados del algoritmo propuesto fueron comparados con los algoritmos K-means y el Algoritmo K-means++. De acuerdo a las pruebas realizadas, RENTOL mejoró los resultados de K-means y K-means++.

Palabras clave: Agrupamiento K-means, indices de validación, algoritmo K++ means.

RENTOL: A Clustering Algorithm Based on K-means

Abtract. K-means algorithm is the most widely used in the community unsupervised learning. Unfortunately it is very sensitive to the selection of initial centroids. For this reason, there are proposed many methods for selection of initial centers. This article presents a clustering algorithm, which is based on K-means algorithm, which a new method for selecting implemented centers initial. To evaluate the results internal clustering validation indices were. The results of the proposed algorithm were compared with the K-means algorithm and K-means ++ algorithms. According to tests, the results of RENTOL improved K-means and K-means ++.

Keywords: K-means clustering, validation indexes, K++ means, algorithm.

1. Introducción

Agrupamiento es una técnica de clasificación no supervisada, cuyo objetivo es encontrar o descubrir grupos en un conjunto de patrones u objetos, además de ser una de las tareas más importantes en el análisis de datos y en la minería de datos [8]. La técnica de agrupamiento ha sido ampliamente utilizada en la detección de anomalías, identificación de características sobresalientes de conjuntos de datos, etc., en diferentes áreas del conocimiento como: biología, antropología, medicina, estadística, matemáticas entre otras. Por lo cual se han desarrollado una gran diversidad de técnicas desde sus inicios en los años 50´s.

Existe en la literatura diferentes clasificaciones de los algoritmos de agrupamiento [10], aunque la más general es proporcionada por Jain [5], donde los algoritmos son clasificados en dos categorías: jerárquicos y de partición. Los algoritmos jerárquicos trabajan interactivamente encontrando una secuencia anidada de particiones tomando en cuenta un criterio para unir o dividir grupos en función de una medida de similitud. Por otra parte los algoritmos de partición, encuentran particiones del conjunto de datos sin ninguna jerarquía. La mayoría de los algoritmos jerárquicos tienen un orden (complejidad) cuadrático [6], por lo que tienen problemas cuando trabajan con grandes volúmenes de datos, mientras que los algoritmos de partición tienen una complejidad menor.

Podemos encontrar en la literatura una gran cantidad de algoritmos de partición [10]. Pero uno de los más utilizados y referenciados es el algoritmo de agrupamiento K-Means. Esto se atribuye a varias razones que a continuación se enlistan:

- La facilidad para comprender su funcionamiento, así como su implementación.
- El algoritmo puede tomar diferentes caminos si se utilizan diferentes criterios como: distancia, método de selección de centroides iniciales o el criterio de terminación del algoritmo.
- El tiempo de complejidad es lineal

A pesar de las ventajas mencionadas anteriormente, no está exento de algunas desventajas que han sido reportadas ampliamente en la literatura [5], las más importantes se enlistan a continuación:

- Tiene dificultades para detectar grupos con formas no esféricas y de tamaño diferente. Cuando los grupos naturales son muy grandes existe una alta probabilidad que seleccionen, como centroides iniciales a puntos del mismo grupo natural, esto ocasionará que el algoritmo divida al grupo natural.
- Es muy sensible al ruido y a puntos atípicos. Si se llegará a elegir como centroide uno de estos puntos, los centroides resultantes pudieran no ser tan representativos como debieran serlo.
- Converge a un mínimo local de la función criterio, obteniendo una solución pobre.
- Es altamente sensible al método para seleccionar los centroides iniciales.

Una forma de contribuir a la eficiencia del algoritmo, es mejorar el método de selección de centros iniciales. Así en este trabajo se presenta un algoritmo de

agrupamiento basado en el algoritmo K- Means, que incluye un método diferente de seleccionar los centros iniciales.

El resto del artículo está organizado de la siguiente manera: en la sección 2 se describe el algoritmo K-Means mencionando algunas de sus ventajas y desventajas, en la sección 3 se describen brevemente los índices de validación internos utilizados en este trabajo. En la sección 4 se presentan algunos trabajos sobre métodos de inicialización de centros, en la sección 5 se describe el algoritmo propuesto, denominado RENTOL. En las secciones 6 y 7 se presentan los resultados obtenidos y la discusión de ellos respectivamente.

2. El algoritmo K-means

La mayoría de los algoritmos de partición tienen como base la optimización de una función criterio que denominaremos F , el valor de esta función depende de las particiones del conjunto de datos $\{C_1, \dots, C_K\}$, es decir:

$$F: P_k(X) \longrightarrow \mathbb{R}, \tag{1}$$

donde $P_K(X)$ son las particiones del conjunto de datos X= $\{x_1, \dots, x_n\}$ en K grupos no vacíos. x_i es un vector $n - dimensional$ (objeto) del conjunto de datos X. El algoritmo K-Means converge a un mínimo local, utilizando la función criterio F , de la sumatoria de las distancias L^2 entre cada objeto y su centroide más cercano. A este criterio normalmente se le denomina error cuadrático. Que puede ser expresado como sigue:

$$F(\{C_1, \dots, C_K\}) = \sum_{i=1}^{K} \sum_{j=1}^{p_i} \| x_{ij} - \bar{C}_i \|, \tag{2}$$

donde K es el número de grupos, p_i es el número de objetos del grupo i, x_{ij} es el $j - ésimo$ objeto del $i - ésimo$ grupo y $\bar{C}i$ es el centroide del $i - ésimo$ grupo el cual es calculado de la siguiente manera:

$$\bar{C}_i = \frac{1}{p_i} \sum_{j=1}^{p_i} x_{ij} \, , \; i = 1, \dots, K. \tag{3}$$

Paso 1. Selecciona los K centroides iniciales $\{C_i, \dots, C_K\}$.

Paso 2. Asigna los objetos x_i del conjunto de datos X, a su centroide más cercano.

Paso 3. Recalcula los nuevos centros, regresa al paso 2, hasta que el algoritmo converge.

Fig. 1. Pseudo-código del algoritmo K-means.

En la Figura 1, se presenta el pseudocódigo del algoritmo K-Means. El algoritmo inicia seleccionando o calculando los centroides iniciales, dependiendo del criterio de selección de centroides, posteriormente asigna los objetos a su centroide más cercano,

para después recalcular los nuevos centroides esto lo realiza hasta que el algoritmo converja (paso 3).

3. Índices validación de agrupamiento

Los índices de validación de agrupamiento son utilizados para evaluar los resultados de los algoritmos de agrupamiento. Existen tres enfoques para el estudio de esta validación: criterio externo, relativo e interno. El criterio interno evalúa los resultados de un algoritmo de agrupamiento considerando información propia del conjunto de datos (como por ejemplo la matriz de proximidad), sin considerar información adicional [7]. En este trabajo se utilizaron los siguientes índices internos para comparar los resultados obtenidos.

3.1. Índice de validación S_Dbw

El índice de validación S_Dbw fue propuesto por M. Halkidi [7]. Este índice evalúa los resultados de un algoritmo de agrupamiento, en función de la densidad y separación de los grupos. Para lo cual mide la varianza intra-grupo e inter-grupo, representada por la Ec. 4:

$$S\_Dbw = Scatt + Dens\_bw. \tag{4}$$

3.2. Índice de validación PS

El índice PS fue propuesto por ChienHsing Chou [3], el cual calcula un promedio de la distancia simétrica hacia los otros centros, con la Ec. 5:

$$PS(C) = \frac{1}{K}\sum_{i=1}^{K}\left[\frac{1}{p_i}\sum_{j=1}^{p_i}\frac{d_c(x_i,\bar{C}_i)}{d_{\min}}\right], \tag{5}$$

donde $d_c(x_i,\bar{C}_i) = d_s(x_i,\bar{C}_i)d_e(x_i,\bar{C}_i)$; $d_e = (x_i,\bar{C}_i)$ es la distancia Euclidiana entre el punto x_i y el centroide $\bar{C}_i$ y d_s es la distancia simétrica [13].

3.3. Índice de validación CS

El índice CS fue propuesto por Chien-Hsing Chou [4]. Este índice evalúa la calidad del agrupamiento tomando en cuenta la densidad y tamaño de los grupos. Se define con la Ec. 6:

$$CS(C) = \frac{\frac{1}{K}\sum_{i=1}^{K}\left\{\frac{1}{|C_i|}\sum_{x_j\in C_i}\max_{x_k\in A_i}\left\{d(x_j,x_k)\right\}\right\}}{\frac{1}{K}\sum_{i=1}^{K}\left\{\min_{j\in c,\, j\neq i}\left\{d(\overline{C_i},\overline{C_j})\right\}\right\}}, \tag{6}$$

donde $\{C_i, \dots, C_K\}$ son los centroides del conjunto de datos X , de los grupos encontrados por el algoritmo de agrupamiento y d es una función de distancia, esta medida es una función de la relación de las sumas de la dispersión dentro del grupo y la separación entre grupos.

4. Trabajos relacionados

M. Emre Celebi [13], presenta una comparación de métodos de inicialización del algoritmo K-Means de orden lineal con respecto al número de objetos: El método Forgy, MacQueen, maximin, Bradley y Frayyad, k-means++, greedy k-means++,Var-Part y PCA-Part. Esta comparación fue realizada con conjuntos de datos reales y sintéticos. Además la comparación se realizó en función de la calidad de agrupamiento y la velocidad de procesamiento, para lo cual se emplearon diferentes índices de validación internos y externos (calidad) el tiempo de CPU y número de iteraciones (velocidad). Concluyendo que los métodos de inicialización del algoritmo K-Means que dieron malos resultados fueron: Forgy, Macqueen y maximin.

El método PCA-Part [14], utilizó un método de agrupamiento jerárquico de tipo divisivo, para lo cual hicieron uso de PCA (Principal Component Analysis). El método inicia, considerando un solo grupo, que contiene a todos los objetos de la muestra en estudio y selecciona iterativamente el grupo con el mayor SSE (suma de cuadrados debida al error), el cual es seleccionado para ser dividido en dos grupos, considerando el hiperplano que pasa por el centro del grupo y que es ortogonal a la dirección del eigenvector de la matriz de covarianza. Este procedimiento se repite hasta que se encuentran los K grupos deseados. Así los centros de los grupos obtenidos son considerados como centroides iniciales del algoritmo K-Means.

El Algoritmo K-Means++, propuesto por Arthur [1], es una modificación del clásico K-means, esta diferencia radica en el método de selección de los centroides iniciales. El método de inicialización de centros, inicia seleccionando aleatoriamente el primer centro C_1 del conjunto de datos X, después escoge el siguiente centro $C_i = x' \in X$ con probabilidad $\frac{D(x')^2}{\sum_{x \in X} D(X)^2}$ proporcional a $D(x')^2$, donde $D(x')^2$ es la distancia más corta entre el punto x y los centros seleccionados previamente. El proceso de selección se repite hasta que $C_i = K$.

Onoda et al. [11], realizan un análisis exhaustivo del método K-means ++ y el método KKz [9], además proponen un nuevo método de selección de centroides, este método se basa en el análisis de componentes independientes (ICA), el cual consiste básicamente de dos pasos:

1.- Obtiene los K componentes independientes $IC_1, \dots, IC_k$ del conjunto de datos X.

2.- Selecciona los K centros iniciales C_i $(i = 1, \dots, k)$ de $C_i = x' \in X$ con un valor mínimo de $\frac{IC_i \cdot x'}{|IC_i||x'|}$.

5. Método propuesto

Sea $X = x_1, x_2, .., x_N$ un conjunto de objetos de tamaño N, donde $x_i \in \mathbb{R}^n$, es decir N objetos (vectores) con n características. El algoritmo RENTOL divide el conjunto X en K particiones, $C = \{c_j, .., c_{K'}\}$, tal que $\cup_{i=1}^{K} c_i = X$, $c_i \cap c_j = \emptyset$ para $1 \leq i \neq j \leq K$ .Cada una de las c_i particiones o grupos tendrán un centroide $\bar{c}_j$.

5.1. Algoritmo propuesto: RENTOL

El algoritmo propuesto requiere como parámetro de entrada el número de grupos K a formar, al igual que el algoritmo K-means. El algoritmo inicia calculando 2 centroides, estos centros son los objetos más alejados del conjunto de datos X (paso 2).

El algoritmo RENTOL requiere como entrada el número de grupos a formar, K

Paso 1. Inicia considerando $K' = 2$

Paso 2. Encuentra los dos objetos más alejados entre sí del conjunto de datos X , que serán los dos centros iniciales

Paso 3. Asigna cada $x_i \, \epsilon X$ para $|i = 1, .., N|$ al centroide más cercano $C = \{c_j, .., c_{K'}\}$, $|j = 1, .., K'|$ obteniéndose los grupos $C = \{c_1, \dots, c_{K'}\}$

Paso 4. Calcula los nuevos centros de los grupos $C = \{c_1, \dots, c_{K'}\}$, como,

$$\bar{c}_j = \frac{1}{|c_j|} \sum_{i=1}^{|c_j|} x_{ij}$$

Paso 5. Repite los pasos 3 y 4 utilizando los nuevos centroides (obtenidos en el paso anterior) $C = \{\bar{c}_j, .., \bar{c}_{K'}\}$, hasta que los objetos no cambien de grupo.

Paso 6. Si $K' = K$ el algoritmo termina

Paso 7. Escoge el siguiente centro $x_i \in X$, que sea el más alejado de los centros $C = \{\bar{c}_j, .., \bar{c}_{K'}\}$

Paso 8. Incrementa en uno el número de centros $K' = K' + 1$

Paso 9. Regresa al paso 3 con K'

Fig. 2. Pseudo-código del algoritmo RENTOL.

En el paso 3, los objetos del conjunto X son asignados a su centroide más cercano. En el paso 4 se recalculan los centros y se repite el paso 3 con los nuevos centros. Los pasos 3 y 4 se repiten hasta que los objetos no cambien de grupo. En el paso 6 se verifica si el número de grupos actual (K') e igual al número de grupos deseado K, sí es así, entonces el algoritmo termina. En caso contrario se calcula el siguiente centroide (paso 7).

Este centro será el objeto más alejado de los centros previamente calculados, en el paso 8 se incrementa en uno el valor K'. En el paso 9 el algoritmo regresa al paso 3 con el valor actualizado de K'. En la Fig. 2 se presenta el pseudocódigo del algoritmo.

6. Resultados experimentales

6.1. Descripción de los conjuntos de datos

Para comprobar los resultados obtenidos por el algoritmo propuesto se utilizaron 27 conjuntos de datos sintéticos, con características bidimensionales estos conjuntos describen figuras geométricas (círculos, elipses, cuadros y rectángulos) de diferentes tamaños, también describen formas irregulares. Además se utilizaron conjuntos de datos reales que fueron obtenidos Irvine, CA: University of California, School of Information and Computer Science[1] (ver Tabla 1).

Table 1. Descripción de los conjuntos de datos.

Conjunto de datos	Número de Registros	Número de Características	Número de clases
Iris	150	4	3
Wine	178	13	3
Diabetes	768	8	2
Glass	214	9	7

6.2. Pruebas realizadas

Para cada una de la pruebas se ejecutaron, los algoritmos K-means, K-means++ y RENTOL, haciendo variaciones del valor de K, desde $K = 2, \ldots, 9$. Se utilizó la distancia Euclidiana, además se midió la calidad de agrupamiento utilizando índices de validación internos: PS, S_Dbw, CS.

6.3. Resultados

En las tablas 2 y 3 se muestran los resultados obtenidos de las pruebas realizadas, cuando se utilizaron las muestras con los algoritmos K-means, K-means++ y RENTOL.

[1] http://ics.uci.edu/mlearn/MLRepository.html.

En el último renglón aparece el porcentaje de aciertos por cada algoritmo de acuerdo al índice de validación.

Table 2. Resultados con datos reales.

	K-Means	**K++Means**	**Rentol**
Iris	2	5	3
Wine	2	2	3
Diabetes	2	2	2
Glass	2	8	9
Porcentaje	25%	25%	75%

Por otro lado también se realizaron pruebas con 27 conjuntos de datos sintéticos, donde se obtuvieron los porcentajes de eficiencia de acuerdo al índice PS, ver Tabla 3.

Table 3. Resumen de resultados con datos sintéticos.

Algoritmo	**PS (Promedio %)**
K-Means	74.00
K++Means	70.03
Rentol	88.88

Es importante resaltar que en las corridas de los experimentos se utilizaron los índices mencionados anteriormente, sin embargo solo se presenta los obtenidos por el índice PS ya que fueron los mejores. Aunque los porcentajes de los otros índices fueron menores se mantuvo la tendencia, es decir el algoritmo Rentol siempre obtuvo porcentajes mayores.

7. Conclusiones

En este artículo se propone un algoritmo de agrupamiento llamado RENTOL. Los resultados fueron comparados con el algoritmo K-means y K-means++, para medir la eficiencia se utilizaron los índices de validación de agrupamiento PS, CS y S_Dbw. En todas las pruebas se utilizó la distancia Euclidiana. Los experimentos se realizaron utilizando datos sintéticos y datos reales.

Los resultados obtenidos, revelan que el algoritmo propuesto obtuvo mejores resultados que los algoritmos K-Means con inicialización aleatoria y K-means++ de acuerdo a los índices de validación PS, CS y S_Dbw. Además elimina la dependencia de la selección de centros iniciales. Aunque el algoritmo que se propone tiene una complejidad mayor al K-means. Sin embargo en un trabajo próximo se trabajará en eliminar esta desventaja.

En el futuro se realizarán experimentos para la segmentación de imágenes médicas e imágenes de un microscopio electrónico.

Referencias

1. Arthur, D., Vassilvitskii, S., K-means++: The advances of careful seeding. In: Proc. of the 18th annual ACM-SIAM symposium on discrete algorithms, pp.1027–1035 (2007)
2. Celebi, M., Emre, K., Hassan, A., Vela, A.: A Comparative study of efficient initialization methods for k-means clustering algorithm. Expert System with Applications, 40, pp. 200–210 (2013)
3. Chow, C. H, Su, M. C., Lai, E.: Symmetry as a new measure for Cluster Validity. In: 2th WSEAS Int. Conf. Scientific Computation and Soft Computing, Crete, Greece, pp. 209–213 (2002)
4. Chow, C. H., Su, M. C., Lai, E.: A new Validity Measure for Clusters with Different Densities. Pattern Anal. Applications, 7, pp. 2005–2020 (2004)
5. Jain, A. K.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31(8), pp. 651–666 (2010)
6. Jain, A. K., Murty, M. N., Flynn, P. J.: Data Clustering: A Review. ACM Computing Surveys, pp. 651–666 (1999)
7. Halkidi, M., Vazirgiannis, M.: Quality scheme assessment in the clustering process. In: Proc. PKDD (Principles and Practice of Knowledge in databases), Lyon, France, Lecture Notes in Artificial Intelligence, Spring-Verlag, Vol. 1910, pp. 265–279 (2000)
8. Kantardzic, M.: Data mining: Concepts, Models, Methods, and Algorithms. Wiley Inter-Science (2001)
9. Katsavounidis, I., Kuo, C. J., Zhang, Z. J., Zhang, Z.: A new initialization technique for generalized Lloyd iteration. IEEE Signal Processing Letters, Vol. 1, No. 10, pp. 144–146 (1994)
10. Kaufaman, L., Rousseeuw, P. J.: Finding groups in data. Wiley Inter-Science (1990)
11. Onada, T., Sakai, M., Yamada, S.: Careful seeding method base on independent components analysis for k-means clustering. Journal of Emerging Technologies in Web Intelligence, 4(1), pp. 51–59 (2012)
12. Peña, J. M., Larranaga, P.: An empirical comparison of four initialization methods for K-means. Pattern Recognition Letters, 20(10), pp. 1027–1040 (1999)
13. Su, M. C., Chow, C. H.: A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry. IEEE Trans. Pattern Anal. and Machine Intelligence, Vol. 23, No. 6, pp. 674–680 (2001)
14. Su, T., Dy, J. G.: In search of deterministic methods for initializing K- means and Gaussian mixture clustering. Intelligent Data Analysis, 11(4), pp. 319–338 (2007)