

Algoritmos Genéticos Aplicados al Problema de "Clustering" basado en Densidad

Edwin Sneyder Gantiva Ramos¹

Universidad Nacional de Colombia, Bogotá, Colombia,
I.esgantivar@unal.edu.co

Abstract. El presente documento tiene como objetivo presentar una aplicación de los algoritmos genéticos para el problema de determinar los centroides de clusters, en esta propuesta no es necesario conocer la cantidad de clusters que se quieren encontrar, basa su búsqueda y optimización de centroides en la densidad. Este trabajo fue desarrollado haciendo uso de algoritmos genéticos con operadores genéticos auto-adaptativos y que fueron presentados en [3], además fue necesario aplicar una técnica para optimización mutimodal que está inspirada en el principio natural del niching, los anteriores presentados en [2] y [7].

Keywords: Optimización, Algoritmos Genéticos, Clustering

1 Introducción

El problema de clustering en simples términos busca encontrar dentro de una base de datos conjuntos de datos similares entre ellos, a esto se le conoce como clusters. Muchos de los métodos conocidos se caracterizan por necesitar de un parámetro inicial que fija la cantidad de estos clusters que se desean encontrar. En el campo de estudio de la computación evolutiva se han presentado trabajos acerca del problema de clustering aplicando algoritmos genéticos en [5] y [8]. Sin embargo estos trabajos hacen uso de parámetros supremamente complejos que salen del común entendimiento y del conocimiento de los autores de este documento. Por tal motivo se ha propuesto una simple técnica que se basa en la detección de buenos centroides para los clusters usando el concepto de la densidad.

Este documento está dividido de la siguiente manera, una sección muestra las definiciones básicas que son necesarias para poder entender el problema que se desea abordar en donde también se introduce el concepto de algoritmo genético, otra sección hace una presentación formal de la propuesta que se expone en este documento y para finalizar se muestran algunos resultados que se encontraron en el proceso de experimentación.

2 Marco Teórico

El análisis de cluster (Clustering) divide los datos en grupos (Clusters) que son significativos, útiles o ambos. Si los grupos significativos son el objetivo, entonces los clusters deberían capturar la estructura subyacente de los datos. En algunos casos, sin embargo el clustering solamente es útil para ser un punto de partida para otros propósitos. [9].

2.1 Algoritmos Genéticos

El algoritmo genético convencional es el concepto base para el desarrollo de este documento, este en términos sencillos busca dada una función de fitness y un conjunto de población inicial busca evolucionar el rendimiento del conjunto de los individuos sobre la función de fitness generación tras generación.

Algorithm 2.1: Algoritmo Genetico($f, \mu, CondicionDeTerminacion$)

```

1 Inicializar la Población Inicial  $P_0$ ;
2 Evaluar( $P_0, f$ );
3 while  $CondiciondeTerminacion$  do
4    $descendientes = GENERACION(P_t, f, SELECCION)$ ;
5    $P_t = FUNCION\_REEMPLAZO(descendientes, P_{t-1})$ 

```

Operadores Genéticos Auto Adaptativos:

Los operadores genéticos auto-adaptativos son un concepto que fue presentado en [3]. La idea principal de estos es que dado un conjunto de tamaño n de operadores genéticos, se codifica dentro del cromosoma de cada individuo de la población las ratas de probabilidad que están asociadas a cada operador. Estas ratas son actualizadas al final de cada operación del operador sobre el individuo y se caracteriza por recompensar o castigar, según sea el caso a la rata del operador. Estas ratas de los operadores son heredadas a sus posibles descendientes.

Deterministic Crowding:

Dentro de los algoritmos genéticos existe un problema que es conocido como el problema de la recombinación [6], este problema consiste en que cuando no existe un control de los individuos que se pueden procrear la final del proceso de evolución todas las especies de individuos han convergido a un único óptimo local.

Para dar solución se han presentado diferentes estrategias, sin embargo dado trabajos previos se ha concluido que una buena estrategia para controlar este problema se basa en una política de restricción de apareamiento basada en una medida de distancia. A esta técnica se le conoce como Deterministic Crowding,

los conceptos para la implementación de esta técnica son tomados de [2].

3 Propuesta

El objetivo de este trabajo es encontrar los mejores candidatos a ser centroides para los problemas de clustering usando algoritmos genéticos, para esto es necesario definir cual sera el cromosoma del problema, en este caso se tratara de un cromosoma con genes de tipo real codificados en un arreglo n-dimensional.

Al tomar esta codificación pueden ser aplicados operadores genéticos bien conocidos y que anteriormente han proveído buenos resultados, los operadores usados son:

- **Mutación Gaussiana:** Esta mutación se basa en una distribución normal con σ dado como parámetro, y dado lo descrito en [1] se genera un σ_i para cada uno de los genes que serán objetivo de mutación.
- **Mutación Uniforme:** Esta mutación se basa en una distribución uniforme, que de forma aleatoria escoge si mutar el gen hacia el límite superior o el límite inferior de acuerdo a un parámetro generado aleatoriamente siguiendo la distribución de probabilidad mencionada.
- **LinealXOver (Combinación Lineal):** En este operador se genera un parámetro aleatorio s siguiendo una distribución de probabilidad uniforme. Ya que el operador tiene una aridad 2 necesita de dos parientes y genera dos sucesores aplicando la siguiente operación a cada gen de los parientes así: $gs_i^1 = s * gp_i^1 + (1 - s) * gp_i^2$ y $gs_i^2 = s * gp_i^2 + (1 - s) * gp_i^1$.

Función de Fitness:

Para el desarrollo de la propuesta es importante mencionar la función de fitness, en esta ocasión esta definida como el conteo de puntos que pertenecen al dataset y que están dentro de un radio que esta definido como parámetro inicial, con este concepto se mide la densidad del centroide candidato con respecto al dataset.

Nota: Los desarrollos presentados anteriormente se encuentran disponibles para su libre uso y consulta en Repositorio en GitHub

4 Resultados

Los resultados de la experimentación se han hecho sobre tres datasets, el primero se trata de un conjunto de datos artificiales con distribución gaussiana, cuyos resultados son presentados en las figuras: 1, 2 y 3. Sobre este conjunto de datos la propuesta presenta buenos resultados.

Los siguientes casos se aplicaron sobre conjuntos de datos bien conocidos que son presentados en [4], y que se caracterizan por no ser clusters con formas regulares. Los resultados son presentados en las figuras: 4, 5, 6, 7, 8 y 9.

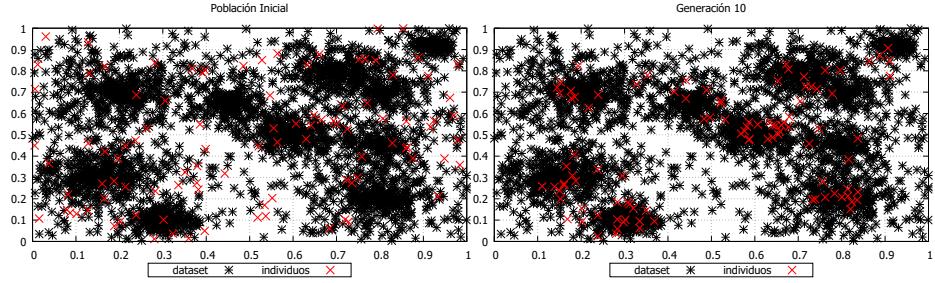


Fig. 1. Del lado izquierdo la distribución de la población inicial, Del lado derecho la distribución de la población en la Generación 10

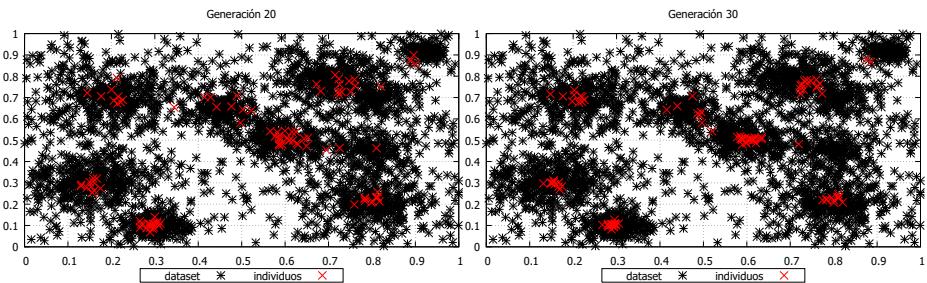


Fig. 2. Del lado izquierdo la distribución de la población en la generación 20, Del lado derecho la distribución de la población en la Generación 30

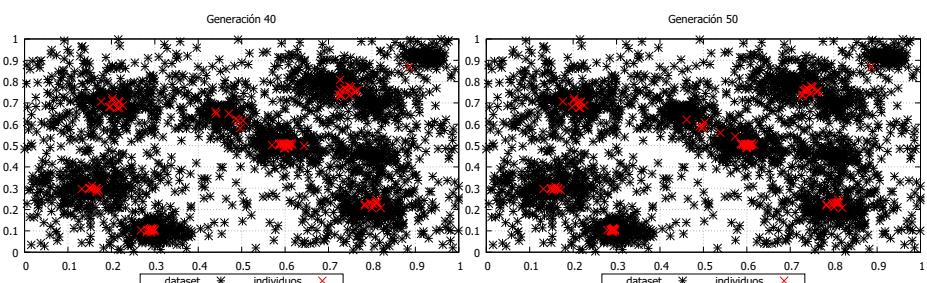


Fig. 3. Del lado izquierdo la distribución de la población en la generación 40, Del lado derecho la distribución de la población en la Generación 50

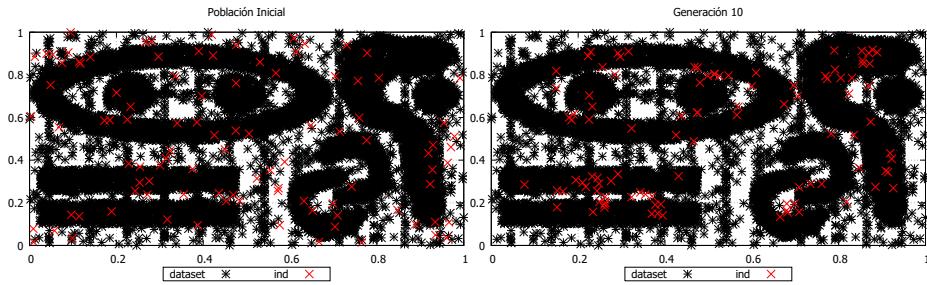


Fig. 4. Del lado izquierdo la distribución de la población inicial, Del lado derecho la distribución de la población en la Generación 10

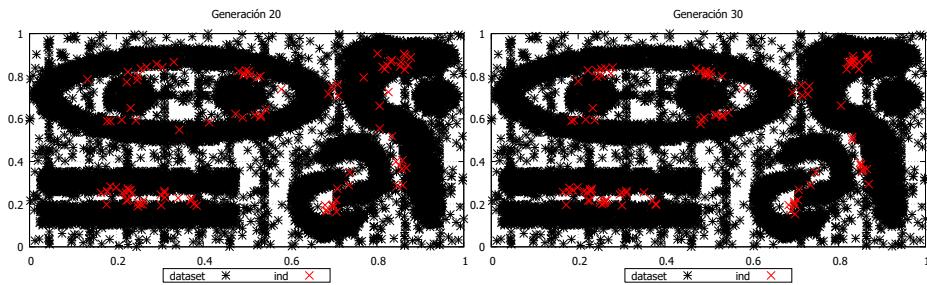


Fig. 5. Del lado izquierdo la distribución de la población en la generación 20, Del lado derecho la distribución de la población en la Generación 30

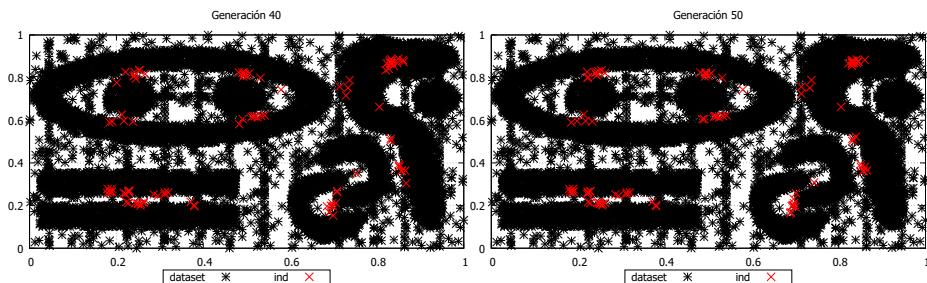


Fig. 6. Del lado izquierdo la distribución de la población en la generación 40, Del lado derecho la distribución de la población en la Generación 50

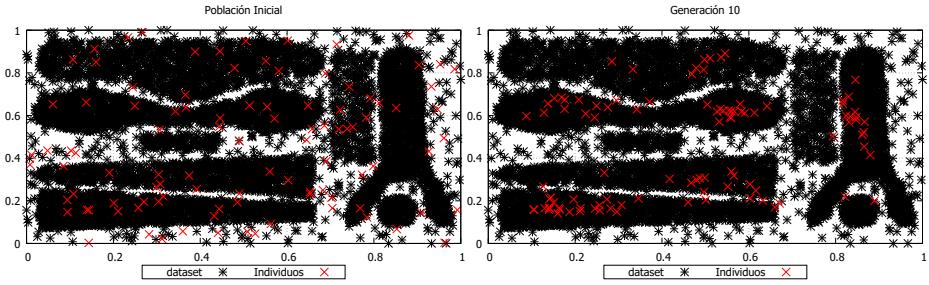


Fig. 7. Del lado izquierdo la distribución de la población inicial, Del lado derecho la distribución de la población en la Generación 10

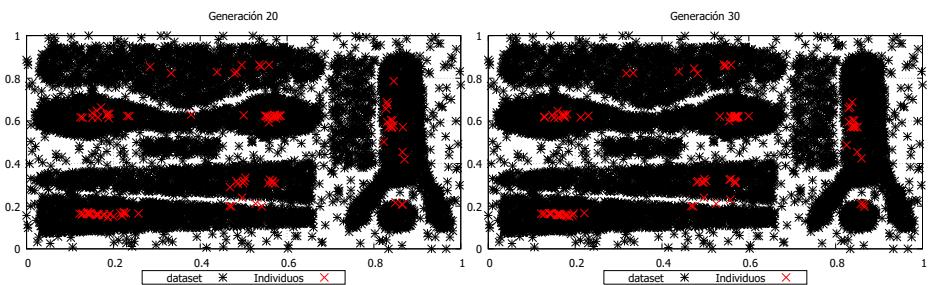


Fig. 8. Del lado izquierdo la distribución de la población en la generación 20, Del lado derecho la distribución de la población en la Generación 30

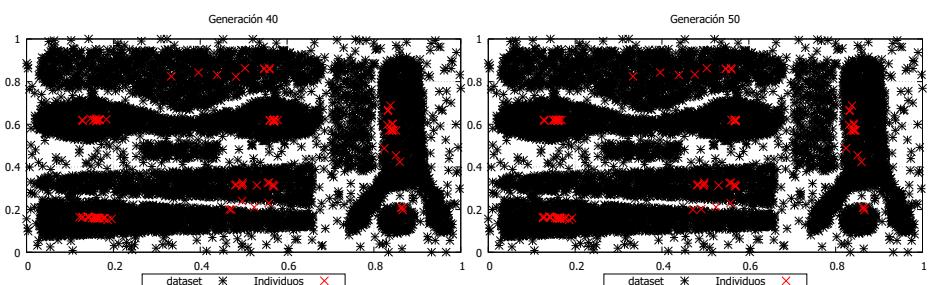


Fig. 9. Del lado izquierdo la distribución de la población en la generación 40, Del lado derecho la distribución de la población en la Generación 50

5 Conclusiones

- La propuesta realizada en este trabajo puede ser usada como una estrategia para la estimación del numero de clusters para algoritmos que solicitan este parametro.
- La propuesta desarrollada muestra buenos resultados para distribuciones de puntos con forma gaussiana. Sin embargo para conjuntos de datos que no tienen esta característica presenta buenos resultados que pueden ser objeto de un proceso de refinamiento para mejorar los resultados.
- La propuesta proporciona buenos resultados a pesar de que se fijan parámetros estáticos del umbral del radio.

Trabajo Futuro

Se recomienda implementar un estimador robusto del radio de los centroides, para que este junto con el cromosoma sea objeto de evolución, esto es conocido como memetico o evolución cultural. Se pueden encontrar propuestas para estos trabajos en [5]

References

1. Hans-Georg Beyer and Hans-Paul Schwefel, *Evolution strategies—a comprehensive introduction*, Natural computing **1** (2002), no. 1, 3–52.
2. Jonatan Gomez, *Self adaptation of operator rates for multimodal optimization*, Evolutionary Computation, 2004. CEC2004. Congress on, vol. 2, IEEE, 2004, pp. 1720–1726.
3. _____, *Self adaptation of operator rates in evolutionary algorithms*, Genetic and Evolutionary Computation—GECCO 2004, Springer, 2004, pp. 1162–1173.
4. George Karypis, Eui-Hong Han, and Vipin Kumar, *Chameleon: Hierarchical clustering using dynamic modeling*, Computer **32** (1999), no. 8, 68–75.
5. Elizabeth Leon, Olfa Nasraoui, and Jonatan Gomez, *Ecsago: Evolutionary clustering with self adaptive genetic operators*, Evolutionary Computation, 2006. CEC 2006. IEEE Congress on, IEEE, 2006, pp. 1768–1775.
6. Samir W Mahfoud, *A comparison of parallel and sequential niching methods*, Conference on genetic algorithms, vol. 136, Citeseer, 1995, p. 143.
7. Brad L Miller and Michael J Shaw, *Genetic algorithms with dynamic niche sharing for multimodal function optimization*, Evolutionary Computation, 1996., Proceedings of IEEE International Conference on, IEEE, 1996, pp. 786–791.
8. Olfa Nasraoui, Elizabeth Leon, and Raghu Krishnapuram, *Unsupervised niche clustering: Discovering an unknown number of clusters in noisy data sets*, Evolutionary Computation in Data Mining, Springer, 2005, pp. 157–188.
9. Tan Pang-Ning, Michael Steinbach, Vipin Kumar, et al., *Introduction to data mining*, Library of Congress, 2006, p. 74.