Whale Optimziation Algorithm aplicado al problema de clustering

Yábir García Benchakhtir

19 de junio de 2020

Índice

1. Introducción al problema **3**

2. Whale Optimziation Algorithm **3**

2.1. Descripción de la metaheurística 3

2.2. Adaptación de la metaheurística al problema 5

1. Introducción al problema

Consideramos el problema del agrupamiento con restricciones. En este problema contamos con un conjunto no vacio $P \subset \mathbb{R}^n$ de puntos y nos planteamos cómo podríamos agruparlos de manera que exista una relación entre los puntos de un mismo grupo. A cada grupo lo denominaremos *cluster* y notaremos $\mathcal{C}$ al conjunto de todos los clusters.

Sobre esta base imponenmos restricciones en la manera en la que se realizan las agrupaciones. En primer lugar existe un subconjunto de pares de puntos ML definido como

$$ML = \{(a, b) \in P \times P \mid \quad a \in K \iff b \in K \text{ para } K \in \mathcal{C}\}$$

es decir, el conjunto de puntos que han de estar en el mismo cluster. De manera similar existe otro conjunto CL de pares de puntos que no pueden pertenecer al mismo conjunto.

$$CL = \{(a, b) \in P \times P \mid \quad a \in K \iff b \notin K \text{ para } K \in \mathcal{C}\}$$

Notaremos por $R = ML \cup CL$ al conjunto de restricciones del problema.

Nos concentraremos en encontrar soluciones bajo restricciones *fuertes* y restricciones *débiles* a este problema. Bajo restricciones fuertes todas las restricciones deben cumplirse y bajo restricciones débiles intentaremos encontrar soluciones que minimicen el conjunto restricciones violadas.

2. Whale Optimziation Algorithm

La metaheurística elegida para resolver el problema ha sido Whale Optimziation Algorithm (WOA) propuesta por Seyedali Mirjalili y Andrew Lewis en 2016. Esta metaheurística basa su comportamiento en las técnicas depredadoras de la ballena jorobada y su comportamiento social.

Utilizando el especial comportamiento que tiene este animal cuando colabora con otros de su misma especia se pretende conseguir una metaheurística que proporcione buenos resultados en problemas de optimización de funciones reales intentando preservar un equilibrio entre exploración y explotación.

Más concretamente la técnica de caza consiste en crear una espiral entorno a la presa y levantar un *muro* de burbujas de aire, haciendo que esta se desoriente para posteriormente acercarse y atacar.

2.1. Descripción de la metaheurística

Para modelar el problema los autores de la metaheurística proponen un modelado matemático del comportamiento de la ballena jorobada que se pueda adaptar adaptar a la optimización de una función real.

Los agentes X que participan en nuestro algoritmo (y que representan a las ballenas) se van a representar como N vectores de dimensión d donde cada componente del vector del agente representa una coordenada en nuestra solución

$$X_i(t) =< X_{i1}(t), X_{i2}(t), \ldots, X_{id}(t) >$$

como deja entrever esta notación el estado de la ballena depende de una variable temporal t que represneta el instante de tiempo en el que nos encontramos y que está limitado por una constante T que fijamos nosotros. En cada instante t la mejor solución vendrá represnetada por X^*.

El movimiento de caza se represneta por la modificación del vector agente mediante la expresión

$$X_i(t+1) = X^*(t) - A \cdot D_1$$

donde $\cdot$ representa el producto componente a componente, D viene dado por la expresión

$$D_i^1 = ||CX^*(t) - X_i(t)||$$

y A y C se obtinen como

$$\begin{aligned} A &= 2a \cdot r - a \\ C &= 2r \end{aligned}$$

con $r \in [0,1]^d$ un vector aleatorio y $a \in [0,2]^d$ constante que se hace decrecer de manera lineal a lo largo de los distintos pasos del algoritmo.

La técnica de *caza* se basa en combinar este movimiento que nos proporciona un componente de exploración junto a la creación de una espiral mediante la actualización del agente de acuerdo a la expresión

$$\begin{cases} X_i(t+1) & = e^{bl} cos(2\pi l) D_2 \oplus X^* \\ D_i^2 & = ||X^*(t) - X_i(t)|| \end{cases}$$

siendo $b \in \mathbb{R}$ constante y $l \in [-1,1]$ aleatorio de manera que nos definen un radio para la espiral en cada instante. La notación $\oplus$ se define como

$$\begin{aligned} \oplus \colon & \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n \\ & (a, < x_1, x_2, \ldots, x_n >) \mapsto < ax_1, ax_2, \ldots, ax_n > \end{aligned}$$

El movimiento de caza natural mezcla tanto los desplazamientos en linea recta como el comportamiento en espiral por lo que se introduce un factor aleatorio $p \in [0,1]$ que decide que tipo de movimeinto se va a realizar

$$X_i(t+1) = \begin{cases} X^*(t) - A \cdot D_i^1 & p < \frac{1}{2} \\ e^{bl} cos(2\pi l) D_i^2 + X^* & p \geq \frac{1}{2} \end{cases}$$

2.2. Adaptación de la metaheurística al problema

Durante el desarrollo de la asignatura hemos trabajado con una representación de la solución que se centraba en la asignación de cada punto a un cluster y se recalculaba en cada caso los centros de cada cluster. El espacio de búsqueda era

$$\{< x_1, x_2, \dots x_n >: x_i \in [0, k] \cap \mathbb{N}, k > 0\}$$

donde n representa el número de puntos que intervienen en el problema y k el número de clusters que consideramos.

Para adaptar la metaheurística he decidido variar mi enfoque del problema y, en lugar de modificar las asignaciones que hago de los puntos, pensar que cada agente representa las coordenas de los centroiedes del problema.

Así cada agente (ballena) queda definido como

$$X_i =<< c_{01}, c_{02}, \dots c_{0d} >, \overset{(k)}{\dots}, < c_{k1}, c_{k2}, \dots c_{kd} >>$$