Artículo
Revista digital Matemática, Educación e Internet (www.cidse.itcr.ac.cr/revistamate/). Vol. 12, Nº 1. Agosto – Febrero 2012.

Determinación de grupos de datos binarios mediante el comportamiento de hormigas

Alejandra Jiménez R.
ajimenez@itcr.ac.cr
Escuela de Matemática
Instituto Tecnológico de Costa Rica

Resumen.

En este artículo se presenta el concepto de Inteligencia de Enjambre (*Swarm Intelligence*) así como las características de ciertas especies de hormigas, utilizadas para modelar la solución de problemas de clasificación de datos. Se utiliza un algoritmo de clasificación y se presentam los resultados obtenidos en grupos de datos binarios.

Palabras claves: colonia de hormigas, clasificación de datos, *clustering*, enjambres de partículas.

Abstract.

This article introduces the concept of Swarm Intelligence as well as characteristics of certain species of ants, used to model the solution of problems of data classification.. It uses a cluster algorithm and presents the results in binary data sets.

KeyWords: Ants colony, data classification , swarms of particles

1.1 Introducción

Las acciones grupales de agentes simples que culminan en el cumplimiento de metas complejas, imposibles de realizar para sólo un agente, han motivado el estudio de tales comportamientos para aplicarlos a la solución de problemas de optimización, clasificación y enrutamiento de grafos, entre otros. El objetivo de este arículo es presentar los resultados de un algoritmo de clasificación de datos, basado en el comportamiento de cierta especie de hormigas. Se presenta una breve introduccón al comportamiento de enjambre así como al problema de clasificación de datos.

1.2 Comportamiento de Enjambre

El comportamiento de grupos de individuos simples, con el que se logran metas complejas, ha motivado a muchos investigadores en el campo de los sistemas artificiales a aplicar estos procedimientos en la búsqueda de la solución de problemas. Estos grupos de individuos que realizan acciones colectivamente inteligentes, reciben el nombre de enjambres [4], los culaes están estructurados por agentes que interactúan entre sí. Entre estos grupos se pueden citar las colmenas, las colonias de hormigas, el tráfico vehicular, las muchedumbres, el sistema inmunológico, las parvadas y la economía, cuyos agentes son respectivamente abejas, hormigas, autos, personas, células y moléculas, aves y agentes económicos. Dado que no existe un conjunto de instrucciones sobre cómo actúan estos agentes, la interacción colectiva dentro del sistema a menudo conduce a algún tipo de comportamiento colectivo o de inteligencia. Este tipo de inteligencia se utiliza para explorar, distribuyendo la solución de un problema entre todos los agentes, sin tener una estructura de control centralizado.
Los enjambres poseen características comunes, tales como estar compuesto de agentes simples, ser descentralizado, no tener un plan global siendo su comportamiento emergente, ser robusto en el sentido de cantidad de agentes que lo forman, las acciones se completan sin depender de un único individuo suceptible a fallar, puede responder a cambios externos y modificar su comportamiento debido a una percepción del entorno.
En el caso de los grupos de insectos, los individuos interactúan de forma social de forma que las desiciones de cada uno dependen de él mismo y de la información disponible mediante los demás agentes. Los insectos siguen reglas simples y se comunican en forma simple, siendo limitado su repertorio de acciones. Ningún individuo tiene acceso al estado completo de la colonia con lo que no es posible realizar una división efectiva de las tareas a cumplir. Aunque cada individuo, según la deficinión de enjambre, no se considera inteligente ni fiable y es una parte simple del grupo, la colonia lleva a cabo acciones de nivel complejo, inteligentes y fiables.

1.2.1 Colonias de Hormigas

El comportamiento de las colonias de hormigas ha sido uno de los modelos más populares de inteligencia colectiva. Individualmente una hormiga podría parecer que actúa al azar y sin un objetivo evidente, sin embargo cuando la acción se realiza en conjunto, surge el objetivo buscado tal como construir el camino más corto desde la colonia hasta la fuente de alimento, mantener limpia la colonia, organizar las larvas, construir estructuras, etc.
Entre sus características, usadas para solución de problemas, está el rastro químico (feromona) que cada individuo deja en el espacio de búsqueda, el cuál es volátil, siendo de gran importancia al modelar soluciones a problemas de grafos o rutas. Otra característica importante es la clasificación y agrupamiento de cuerpos así como el acomodo de larvas, las cuales se pueden observar [1] en las especies *Messor sancta* y *Lepthothorax unifasciatus* respectivamente, estas características son útiles en la búsqueda de soluciones a problemas de clasificación de datos.
El acomodo de larvas ha sido observado en la especie *Lepthothorax unifasciatus*, en la cual las hormigas obreras apilan las larvas según su tamaño. El acomodo de las larvas en un grupo, es de forma tal que en el centro del conjunto se ubican las microlarvas y huevos, dándoles un pequeño espacio individual.A medida que aumenta el tamaño de la larvas, estas son ubicadas en la periferia, aumentando el espacio entre ellas. Si existen pupas, estas se ubican entre las larvas medianas y grandes. Una explicación para este comportamiento puede ser la agrupación de individuos con necesidades similares en un mismo radio, con lo que el cuido es más eficiente y no así al tamaño de los elementos a clasificar.
En la especie *Lassius niger* las obreras construyen pilas de hormigas muertas a manera de cementerio a fin de mantener ordenado el nido [1]. Particularmente, con la especie *Messor sancta* se han realizado experimentos

en los cuales se distribuyen en forma aleatoria cuerpos o partes de estos en una región. Las obreras entonces clasifican estos cuerpos en varios cementerios por medio de su similitud.

1.3 Clasificación de datos

El problema de clasificación de datos consiste en crear grupos de individuos similares, donde a cada uno se le ha realizado observaciones sobre características comunes. Así, dada una colección de objetos, se desea clasificarlos en clases diferenciadas, de acuerdo con las disimilitudes entre los mismos, de forma tal que las clases sean homogéneas internamente [2]. Para reconocer estos grupos de individuos homogéneos se han utilizado métodos jerárquicos, piramidales, clasificación difusa, cruzada, etc.

Teniendo entonces n individuos con m atributos, cada uno se representa en un espacio de dimensión $n \times m$ y se desea proyectar el espacio de las variables en un espacio de dimensión menor a fin de que la distancia entre los atributos de individuos en el mismo grupo sea menor que la distancia entre los atributos de individuos en distintos grupos.

Como parte de la métodos de clasificación se tiene la *clasificación automática*, en la que la clasificación debe realizarse sin intervención de un agente externo y la clasificación y el número de grupos son desconocidos a priori. Por otra parte, mediante las técnicas de *clasificación supervisada* se clasifica los individuos en cierta cantidad dada de grupos.

El comportamiento de las hormigas *Lassius niger* puede modelarse [3] como una técnica de clasificacón automática. El acomodo de cuerpos similares puede representarse mediante las acciones simples de recogerlos, trasladarlos y dejarlos en otro lugar, en el que se hallen objetos similares. Para esto debe medirse la similitud de los objetos en un vecindario en torno a la ubicación de cada hormiga, donde se ha trasladado objeto. En las hormigas, la percepción de similitud puede hacerse por medio señales de tacto o químicas.

Suponiendo que todos los objetos a clasificar inicialmente son iguales para una hormiga, ésta eligirá uno al azar. La probabilidad p_r de que una hormiga libre (sin que cargue un objeto) recoja un objeto puede representarse [3] por

$$p_r = \left(\frac{k_r}{k_r + f}\right)^2$$

donde f es la cantidad percibida de objetos similares al objeto transportado, en el vecindario de la hormiga (densidad local) y k_r es un umbral constante, elegido de forma que si f tiene un valor mucho menor a k_r, entonces p_r sea cercana a 1, siendo alta la probabilidad de que el objeto sea recogido. Por otro lado, p_r es cercana a 0 cuando f es mucho mayor a k_r, de modo que en este caso la densidad local de objetos similares en el vecindario es alta, haciendo baja la probabilidad de que sea trasladado del grupo.

La probabilidad p_d de que un objeto que está siendo trasladado, sea ubicado en un nuevo grupo puede medirse por [3]

$$p_d = \left(\frac{f}{k_d + f}\right)^2$$

con k_d un umbral constante. De este modo, si f es mucho menor a k_d entonces p_d es cercana a 0 mientras que si f es mucho menor a k_d, resulta que p_d es cercana a 1.

Las hormigas utilizan rastros químicos y los sentidos para dirigir su comportamiento y medir la similitud entre objetos, por lo cual debe modelarse una función que tome en cuenta estos factores. Dados dos objetos i y j, su similitud s o disimilitud d debe ser medida de forma que cuanto mayor sea el valor de la función de similitud, mayor será el parecido entre ellos y entre más cercano a cero, menor será el parecido. Una manera de definir f [5] es considerando la disimilitud d entre objetos. Así, dados i y j, se dice que son idénticos si $d(i,\ j) = 0$ y $d(i,\ j) = 1$ si no son idénticos ($0 \leq d \leq 1$).Esta disimilitud puede ser una distancia.

Para el caso de individuos con atributos medidos en forma binaria con valores 0 ó 1, se puede utilizar [2] el índice de similitud de Jaccard, calculado mediante $s(i,j) = \frac{p_{ij}}{p_{ij}+q_{ij}}$ o el índice de Russel y Rao medido por

$s(i,j) = \frac{p_{ij}}{p}$, entre otros, donde p_{ij} es el número de atributos que los objetos i y j poseen a la vez (similitudes), q_{ij} es el número de atributos que presenta sólo uno de ellos (diferencias) y p es el número total de atributos. Considerando s con un valor máximo de 1, la función de disimilitud d es calculada de la forma

$$d(i,j) = 1 - s(i,j)$$

Los agentes que se mueven en el espacio de proyección pueden percibir una región de dimensión $m \times m$ (Neigh$_{(m\times m)}$)en torno a su posición. Se busca una reducción de dimensión, pues al considerar cada objeto a clasificar como una observación con varios atributos, se proyecta el espacio de atributos en un espacio de menor dimensión z para realizar grupos de datos con distancias intra−clases menores. La implementación del siguiente algoritmo resuelve la clasificación de datos binarios, dando como resultado las ubicación de los individuos en un plano en $\mathbb{R}^2$. Se representan las hormigas en un espacio discreto, con vecindarios P de dimensión $m \times m$ al rededor del sitio r en el que se ubica la hormiga. La densidad local f con respecto al objeto i en el momento t puede calcularse [5] de la forma:

$$f(i) = \begin{cases} \frac{1}{m^2} \sum_{j \in P_{m\times m}(r)} \left(1 - \frac{d(i,j)}{\alpha}\right) & \text{si } f > 0 \\ 0 & \text{de otro modo} \end{cases}$$

de modo que $f(i)$ medirá la similitud promedio del objeto i con los objetos j que se encuentran en el vecindario de i. El valor α define la escala de similitud que determina la medida en la que se acepta que dos objetos se ubiquen en el mismo grupo, esto es que sean similares.

Algoritmo 1.1: Clasificación mediante Colonias de Hormigas

```
1  Ubicar cada objeto i aleatoriamente en una celda;
2  Ubicar cada hormiga k aleatoriamente en una celda libre;
3  iter = 1;
4  while iter < maxiter do
5      for n = 1,2,...TotalHormigas do
6          if Hormiga desocupada y Celda ocupada por objeto i then
7              Calcule f(i) y Calcule p_r(i);
8          if Hormiga ocupada con objeto i y Celda desocupada then
9              Calcule f(i) y Calcule p_d(i);
10             Dejar objeto i con probabilidad p_d(i);
11         Mover Hormiga aleatoriamente a un vecindario escogido al azar y a una celda desocupada;
12     iter = iter +1;
13 return Ubicaciones de objetos;
```

1.3.1 Datos y resultados

El algoritmo 1.1 sugerido en [1] se implementó utilizando una región de 100×100 ubicaciones en las que se distribuyeron aleatoriamente los individuos (asignando coordenadas a cada uno), utilizando vecindarios de tamaño 3×3. El resultado del programa, *Ubicaiones de objetos*, luego de una cantidad definida de iteraciones es un conjunto de coordenadas del plano cartesiano que representado gráficamente muestra la similitud de los individous, mediante la cercanía en la gráfica de cada punto que los representa. Con esto es posible identificar la cantidad de grupos de datos similares presentes en el conjunto de datos.

A continuación se presentan los resultados obtenidos en dos grupos de datos: uno con datos simulados y otro con datos estudiados previamente en [6] donde se sugiere su agrupamiento.

1.3.2 Datos Simulados

Este conjunto de datos consiste de 20 con 5 atributos, mostrados en el cuadro 1.1. Existen cuatro grupos definidos, pues en cada uno los individuos poseen las mismas características.

Grupo	Individuo	AI: Atributos					
1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1
	3	1	1	1	1	1	1
	4	1	1	1	1	1	1
	5	1	1	1	1	1	1
2	6	1	1	1	0	0	0
	7	1	1	1	0	0	0
	8	1	1	1	0	0	0
	9	1	1	1	0	0	0
	10	1	1	1	0	0	0
3	11	0	0	0	1	1	1
	12	0	0	0	1	1	1
	13	0	0	0	1	1	1
	14	0	0	0	1	1	1
	15	0	0	0	1	1	1
4	16	0	0	0	0	0	0
	17	0	0	0	0	0	0
	18	0	0	0	0	0	0
	19	0	0	0	0	0	0
	20	0	0	0	0	0	0

Tabla 1.1 Grupo de datos binarios simulados.

Mediante 10^4 iteraciones, 20 hormigas, con valores de las constantes: $k_r = 0.1,\ k_d = 0.15,\ \alpha = 0.3$ y el índice de similitud de Jaccard, el algoritmo genera las coordenadas mostradas en el cuadro 1.2.

Individuo	Coordenada	Individuo	Coordenada
1	(32,57)	2	(33,55)
3	(33,56)	4	(32,56)
5	(34,56)	6	(95,21)
7	(95,23)	8	(95,22)
9	(94,21)	10	(96,22)
11	(83,29)	12	(84,28)
13	(83,30)	14	(83,31)
15	(84,29)	16	(73,54)
17	(73,55)	18	(74,55)
19	(72,55)	20	(72,54)

Tabla 1.2 Coordenadas de ubicación, generadas por el programa para el grupo de datos simulados: 20 hormigas, 10^4 iteraciones, $k_r = 0.1,\ k_d = 0.15,\ \alpha = 0.3$

En la figura 1.1 se muestra la gráfica de las coordenadas generadas por el programa, mostradas en el cuadro 1.2.

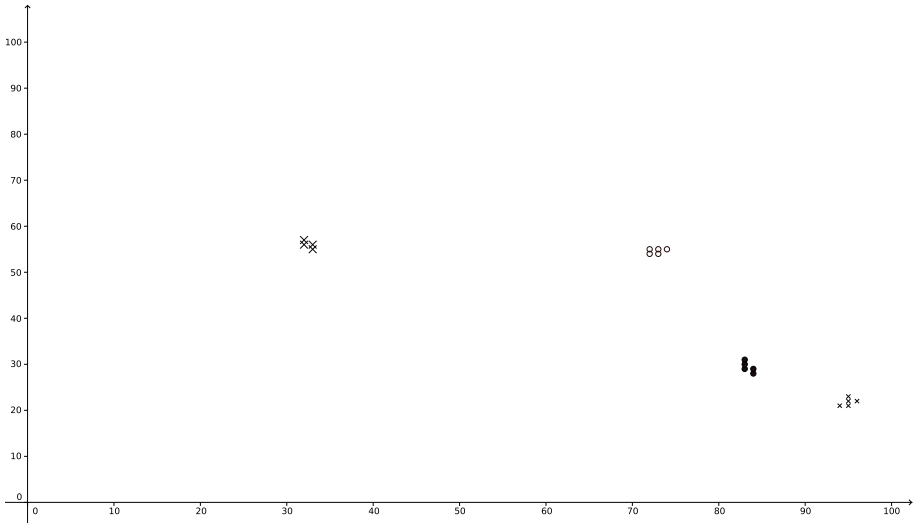

Figura 1.1 Clasificación de datos simulados, dados en el cuadro 1.2. Simbología: $\times$ - grupo 1, $\times$ - grupo 2, $\bullet$ - grupo 3, $\circ$ - grupo 4.

1.3.3 Datos de plantas de pejibaye

El segundo grupo de datos analizado proviene de un estudio realizado por [6] a 6 distintas poblaciones de plantas de pejibaye (*Bactris gasipaes*) ubicadas en Brasil, Perú, Bolivia, Colombia, Panamá y Costa Rica. Este conjunto consta de 87 individuos (fragmentos de plantas de pejibaye) descritos por 60 variables binarias (rastro genético). La solución óptima y natural es la clasificación de los datos en seis grupos, correspondientes a los países de origen, de la siguiente forma:

Grupo1$= \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15\}$,

Grupo2$= \{16,17,18,19,20,21,22,23,24,25,26,27,28,29,30\}$,

Grupo3$= \{31,32,33,34,35,36,37,38,39,40,41,42,43\}$,

Grupo4$= \{44,45,46,47,48,49,50,51,52,53,54,55,56,57\}$,

Grupo5$= \{58,59,60,61,62,63,64,65,66,67,68,69,70,71,72\}$,

Grupo6$= \{73,74,75,76,77,78,78,80,81,82,83,84,85,86,87\}$.

El programa genera las coordenadas mostradas en el cuadro 1.3 para la clasificación de los individuos luego de 10^6 iteraciones, usando 20 hormigas con los valores $k_r = 0.1$, $k_d = 0.15$, $\alpha = 0.3$ y el índice de similitud de Jaccard.

Grupo	Coordenada por Individuo
1	$\{(17,91),(16,92),(18,91),(15,93),(16,91),(17,90),(20,92),(18,90),$ $(20,90),(17,92),(17,89),(20,91),(19,91),(19,92),(16,90)\}$
2	$\{(84,54),(83,53),(83,54),(84,53),(84,52),(83,50),(82,53),(82,54),$ $(24,26),(85,51),(85,54),(83,52),(82,56),(83,55),(81,55)\}$
3	$\{(95,92),(100,95),(96,93),(91,89),(98,94),(94,89),(98,97),(99,96),$ $(100,96),(92,89),(99,94),(92,90),(96,94)\}$
4	$\{(16,40),(14,44),(14,43),(16,39),(17,38),(17,42),(14,41),(17,41),$ $(11,40),(6,14),(16,38),(12,44),(17,39),(17,39)\}$
5	$\{(22,23),(22,24),(24,24),(23,24),(23,23),(19,20),(18,24),(24,23),$ $(20,20),(19,25),(20,19),(21,18),(23,25),(19,24),(19,21)\}$
6	$\{(42,93),(45,90),(46,90),(46,99),(45,96),(44,97),(45,100),(45,98),$ $(49,90),(48,89),(45,99),(46,95),(48,92),(50,89),(47,89)\}$

Tabla 1.3 Coordenadas de ubicación, generadas por el programa para el grupo de datos simulados: 20 hormigas, 10^4 iteraciones, $k_r = 0.1,\ k_d = 0.15,\ \alpha = 0.3$

En la figura 1.2 se grafican las coordenadas dadas en el cuadro 1.3 y en la figura 1.3 se muestra la gráfica de coordenadas generada en una segunda corrida del programa en la cual se observa nuevamente los seis grupos de datos, en este caso se omiten las coordenadas generadas por el programa. Cabe destacar que el agoritmo al no ser determinístico, no genera una única solución, sin embargo aunque las coordenas para los indivuduos son diferentes se clasifican en forma similar. Con esto se comprobó que el programa realiza una clasificación adecuada de los datos mediante una adecuada escogencia de los parámetros constantes.

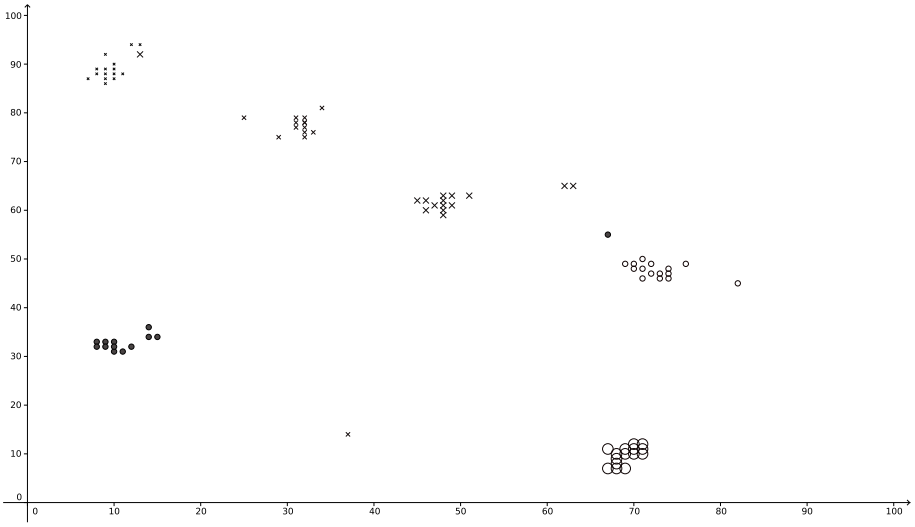

Figura 1.2 Clasificación de datos de pejibaye dados en el cuadro 1.3. Simbología: ○ - grupo 1, × - grupo 2, ● - grupo 3, × - grupo 4, × - grupo 5, ◯ - grupo 6.

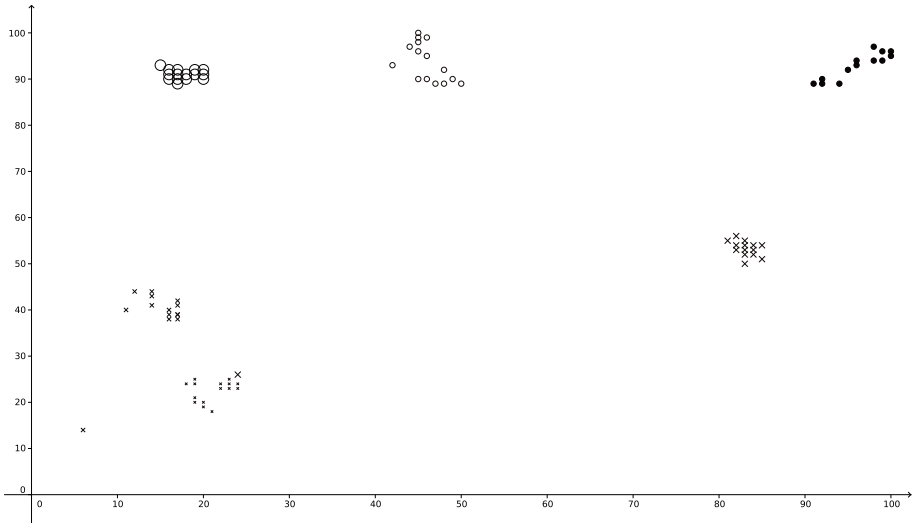

Figura 1.3 Clasificación de datos de pejibaye. Simbología: ○ - grupo 1, × - grupo 2, ● - grupo 3, × - grupo 4, ╳ - grupo 5, ◯ - grupo 6.

1.4 Conclusiones

A partir de los resultados obtenidos se observa que el algoritmo agrupa correctamente los datos simulados y realiza una adecuada clasificación de los datos de pejibaye, mostrando que existen 6 grupos de datos, los cual concuerda con el estudio realizado en [6]. Al aumentar la cantidad de atributos se debió aumentar la cantidad de iteraciones para que los resultados fueran veraces. Al no ser un algoritmo determinístico los resultados pueden variar. Se recomienda realizar una variación en los parámetros para los distintos grupos de datos, esto a fin de favorecer el agrupamiento. El algoritmo se basa en similitud dentro de un vecindario (intraclase) y no la mide entre vecindarios (interclase), con esto podrían obtenerse grupos de datos similares representados en varios subgrupos, por tanto se recomienda analizar la similitud entre grupos.

Bibliografía

[1] E. Bonabeau, M. Dorigo, G. Theraulaz. Swarm Intelligence. Santa Fe Institute Studies in the Sciences of Complexity. New York, Oxford, 1999.

[2] Castillo E., William. González V., Jorge. Trejos Z., Javier. Anális Multivariado de Datos: Métodos y Aplicaciones.Escuela de Matemática, Universidad de Costa Rica.

[3] Deneubourg, J.L., Goss, S., Franks, N., Sendova-Franks A., Detrain, C., Chretien, L. The Dynamic of Collective Sorting Robot-like Ants and Ant-like Robots. Cambridge, MA: MIT Press, 1991.

[4] J. Kennedy, R. Eberhart. Swarm Intelligence. Morgan Kaufmann Publishers, 2001.

[5] Lumer E. D., Faieta B. Diversity and Adaptation in Populations of Clustering Ants. Cambridge, MA: The MIT Press/Bradford Books, 1994.

[6] Murillo, A., Piza, E., Trejos, J. Combinatorial Optimization Heuristics in Partitioning with non Euclidean Distances. CIMPA, Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica