Estudo de caso do algoritmo bioinspirado Ant-Clustering

Victor Eduardo Requia

[1]Universidade do Estado de Santa Catarina – UDESC
Joinville – Brasil

victorrequia@gmail.com

Resumo. *Com o advento da era digital, a explosão de dados na internet tornou-se realidade. Alavancado pelo baixo custo de armazenamento e pela necessidade de manter informação. Uma das áreas que está em grande crescimento é a mineração de dados (Data Mining) que busca identificar padrões para extrair conhecimento implícito de um vasto conjunto de dados. Esse trabalho, explora a teoria e a implementação do algoritmo Ant-Clustering, que tem como objetivo, agrupar dados semelhantes utilizando um pequeno conjunto de organismos, baseado em colônias de formigas, em um ramo da Inteligência Artificial chamado Swarm Intelligence (SI) no qual, agentes simples que interagem entre si e com o ambiente, apresentam comportamento emergente.*

1. Introdução

Com a grande quantidade de dados armazenados na internet, várias áreas estão surgindo para obter informações relevantes sobre esse montante de dados. A ascensão da inteligência artificial juntamente com a evolução dos componentes de hardware, tem facilitado a incorporação dessas técnicas em diversos campos.

Para o problema de agrupamento de dados semelhantes (Clustering), diversos algoritmos tem sido propostos, entre eles, os algoritmos bioinspirados. Esses algoritmos são inspirados em fenômenos naturais, especificamente no comportamento coletivo para modelar e solucionar problemas do mundo real. Uma das subáreas dos algoritmos bioinspirados é a de Inteligência de Enxame (Swarm Intelligence) que, têm se destacado como promissores nos resultados na qual são aplicados [Jafar and Sivakumar 2010].

Dentro da área de Swarm Intelligence, existem mais duas subáreas de algoritmos. Colônias de Formigas (ACO) e Algoritmos de Enxames de Partículas (PSO). Nesse trabalho vamos focar no algoritmo de Colônia de Formigas para resolver o problema de agrupamento de dados. Conforme evidenciado por pesquisas recentes, algoritmos baseados em agrupamento por colônias de formigas (ant-based clustering) têm alcançado resultados notáveis para esse tipo de aplicação[Jafar and Sivakumar 2010].

A otimização baseada nas colônias de formiga (ACO) é uma metaheurística global que tem como inspiração no comportamento social das formigas [Melo et al. 2019]. É um algoritmo não supervisionado de classificação de padrões em diferentes grupos [Jafar and Sivakumar 2010], podendo ter elementos heterogêneos e multidimensionais.

Apesar do algoritmo ser muito utilizado e ter grande relevância na área da otimização, uma das limitações é a sensibilidade a parametrização, na qual, dado uma configuração podemos obter resultados significativamente diferentes além de poder levar facilmente a perda da diversidade e à convergência prematura [Melo et al. 2019].

Na próxima seção, será abordada a metodologia de desenvolvimento na qual, será descrita as estratégias utilizadas assim como a justificativa das escolhas. Na seção 3, será feita a descrição de experimentos, as simulações e os resultados obtidos. Na seção 4 será feita a análise dos resultados obtidos. E na última seção será feita a conslusão sobre os resultados do trabalho.

2. Metodologia de Desenvolvimento

Em um ambiente simulado 2D, processado de forma sequencial, com dimensão 80x50, será realizado experimentos com um conjunto de itens, primeiramente homogêneos e depois heterogêneos multidimensionais, espalhados de forma aleatória nesse ambiente, junto com formigas, na qual, o objetivo é agrupar esses dados conforme suas características. Para isso, será desenvolvido um algoritmo bioinspirado baseado no comportamento social de agrupamento de itens por formigas em uma colônia de formigas descrito na seção 1.

O algoritmo consiste em dois principais métodos para atingir o objetivo, probabilidade de pegar o item (Pp) ou probabilidade de largar o item (Pd). Esses métodos são intrínsecos as formigas e são eles que darão a característica de agrupamento dependendo dos parâmetros utilizados.

Para montar o ambiente de simulação, foram utilizadas a linguagem de programação Java, junto com a biblioteca Swing, usada para desenhar o ambiente de forma gráfica e ver os resultados em tempo real. As simulação foram divididas em duas partes: Simulação com dados homogêneos e simulação com dados heterogêneos.

2.1. Dados homogêneos

Nesta simulação, todos os dados são iguais (representados no ambiente como folhas), não existe nenhum tipo de distinção por parte da formiga e todos elas possuem um raio de visão igual a 1. No inicio, 400 itens (folhas) foram colocados no ambiente, como podemos ver um exemplo na figura 1. Foi escolhido a quantidade de 400 itens, pois é a mesma quantidade de itens utilizada em um dos cenários para dados heterogêneos, usado depois para depois fazer comparações na seção Análise dos resultados obtidos entre esses dois cenários.

A formula utilizada pelas formigas para pegar ou largar um item são descritas a seguir.

$$P_{pegar}(x_i) = (1 - \frac{i}{n})$$

$$P_{largar}(x_i) = \frac{i}{n}$$

Nestas equações, i representa a quantidade de itens próximos e n, o número máximo de itens que podem ser vistos no raio da formiga.

2.2. Dados heterogêneos

Para os dados heterogêneos, foram feitas duas simulações. A primeira possuía uma base sintética de 400 itens (dados) heterogêneos com duas dimensões cada. Já a segunda base de dados, foi composta por uma base sintética de 600 itens (dados) heterogêneos de duas dimensões cada. Todas as formigas tinham um raio de visão fixo igual a 1.

Na primeira simulação eram esperadas a formação de 4 grupos distintos no ambiente de simulação e na segunda eram esperadas a formação de 15 grupos distintos no ambiente de simulação.

Ambas as simulações utilizaram a mesma equação que iria geria a regra de pegar o item ou largar o item. Esse conjunto de regras pode ser descrita como

$$f(i) = \max(0, \frac{\sum_j \left(1 - \frac{d(i,j)}{\alpha}\right))}{s^2}$$

A equação acima define a similaridade entre um dado e sua vizinhança. Quanto menor as distâncias tendem a ser, maior a similaridade tende a ser com os dados que pretense-se largar ou relação às células da vizinhança (s x s)

$$P_{pegar}(x_i) = (\frac{K_1}{K_1 + f(x_i)})^2$$

A equação acima define a regra para pegar um item. Sendo K1 um parâmetro constante que pode varia de 0 até 1, sendo K1 maior que 0. O resultado da função pode variar de k1/(k1 + 1) até 1.

$$P_{largar}(x_i) = (\frac{f(x_1)}{K_2 + f(x_i)})^2$$

A equação acima define a regra para largar um item. Sendo K2 um parâmetro constante que pode varia de 0 até 1, sendo K2 maior que 0. O resultado da função pode variar de 0 até 1/(k2 + 1).

$$D_{euclidiana} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

A equação acima é conhecida como distância Euclidiana, neste caso, foi usada para medir a distância entre as dimensões dos dados.

3. Experimentos

Nesta seção, será apresentado os experimentos dos cenários das simulações homogêneas e heterogêneas propostas na seção de metodologia de desenvolvimento.

3.1. Dados homogêneos

Para esse experimento, foi disposto um cenário de 400 itens dentro de uma matriz 80x50 com 15 formigas. Foram realizados 3 teste para fazer a análise do grau de aleatoriedade e de semelhança entre eles na seção Análise dos resultados obtidos. O critério de parada foi uma condição que engloba iterações e quantidade de formigas carregando, na qual, o programa irá parar caso tenham mais de 150000 iterações e todas as formigas largaram os itens. Os experimentos são vistos nas figuras 1,2 e 3.

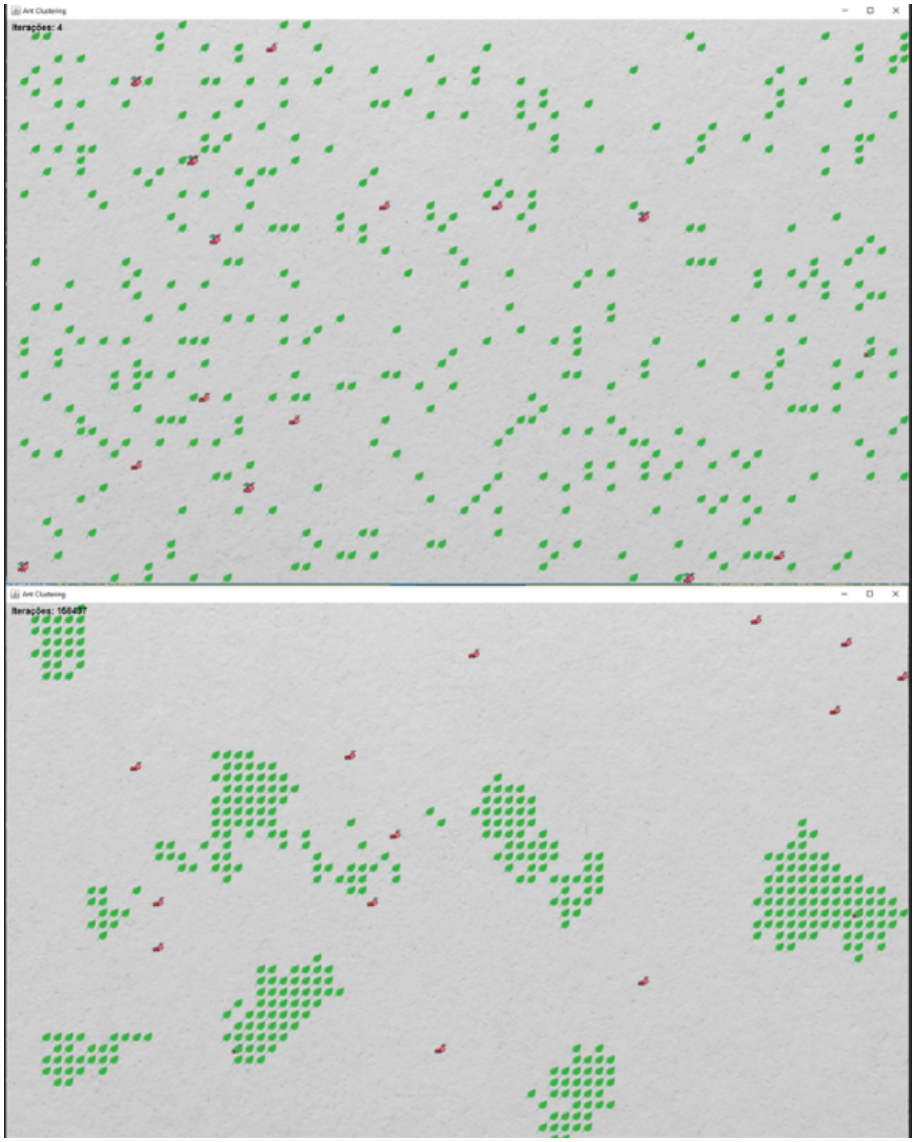

Figura 1. Ambiente inicial e final cenário 1 - 4 e 158437 iterações

3.2. Dados hetetogêneos

Essa subseção será dividida em outras duas tratando das simulações de dados heterogêneos multidimensionais. A subseção 4 Grupos, irá tratar da simulação com 4 grupos distintos e um total de 400 dados. Na subseção 15 grupos, será mostrado os experimentos com 15 grupos distintos com 600 dados diferentes. É importante salientar que as formigas não reconhecem esses grupos, apenas tem as infamações das dimensões. O critério de parada foi o mesmo utilizado para agrupar dados homogêneos.

3.3. 4 Grupos

Para o experimento, foram realizados vários testes modificando os valores das constantes de alfa, k1 e k2. Dentre os testes, a maior variação foi em relação a mudança da constante alfa, utilizada na fórmula da função f(i) mostrada na seção Metodologia e Desenvolvimento, por esse motivo, será disposto 3 cenários com diferentes valores de alfa, como podemo ver nas figuras 4,5 e 6. Os valores de k1 e k2 foram igual a 0.7.

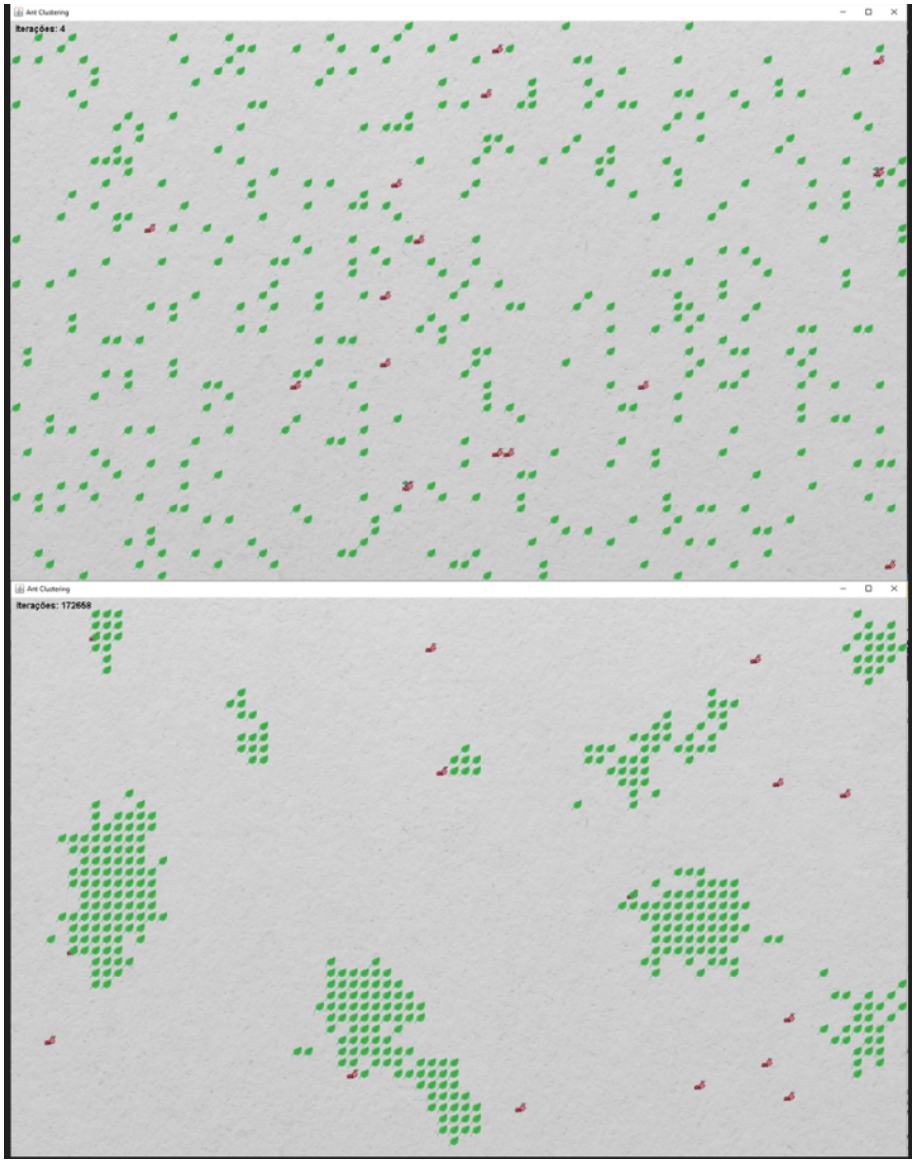

Figura 2. Ambiente inicial e final cenário 2 - 4 e 172658 iterações

3.4. 15 Grupos

Para a simulação com 15 grupos, foram testados diversos cenários alterando os valores das constantes alfa, k1 e k2, porém, mesmo executando durante mais longos períodos, não foi possível obter um resultado significativo em termos de agrupamento global. No melhor teste, com alfa=1, k1=0.1 e k2=0.1, foi disposto o seguinte cenário (figura 7). Podendo ser observado apenas o agrupamento mínimo e local de itens.

4. Análise dos resultados obtidos

Tanto nos cenários de testes dos dados homogêneos como heterogêneo, depois de 50000 iterações, o cenário manteve-se estável, mudando pouco sua forma, mesmo com valores de constantes diferentes nos dados heterogêneos.

Analisando os cenários dos dados heterogêneos e multidimensionais, alterando o valor da constante alfa, podemos concluir que (como citado na seção de introdução) dado

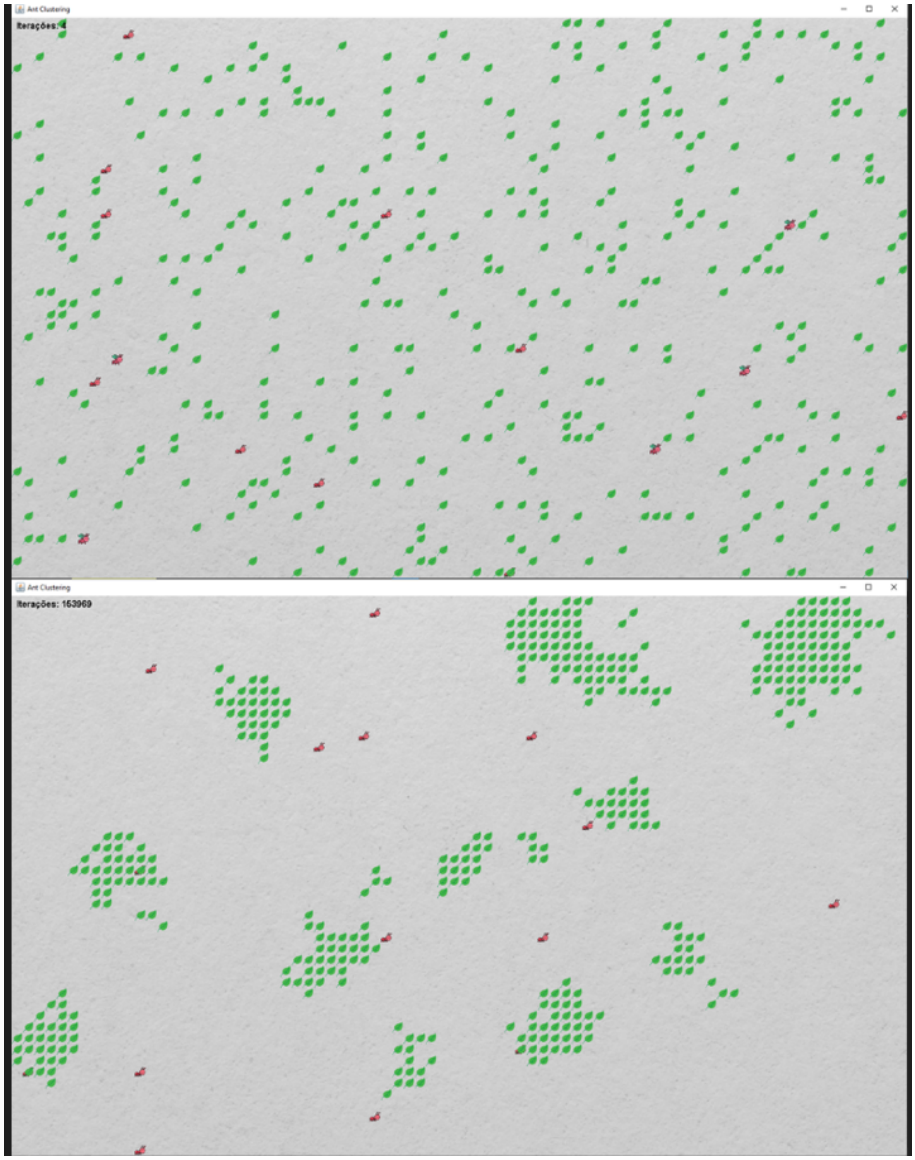

Figura 3. Ambiente inicial e final cenário 3 - 4 e 153969 iterações

uma configuração, podemos obter resultados significativamente diferentes, além de poder levar facilmente a perda da diversidade e convergência prematura. No caso da formação de 15 grupos, as constantes possuíam uma sensibilidade muito maior que em 4 grupos, especialmente a constante alfa.

Podemos notar que, utilizando uma fórmula não linear no agrupamento heterogêneo, foi possível observar um melhor agrupamento dos itens, deixando-os mais consistentes que no agrupamento homogêneo mesmo com a mesma quantidade de itens, como podemos ver nas figuras 1 e 6.

É interessante ver que no agrupamento de itens homogêneos nas imagens 1,2 e 3, todos apresentaram pequenos e grandes grupos e, todos os grupos variavam as posições e o número de iterações necessárias para encerrar o programa. Isso se deve ao fato da estocasticidade do algoritmo.

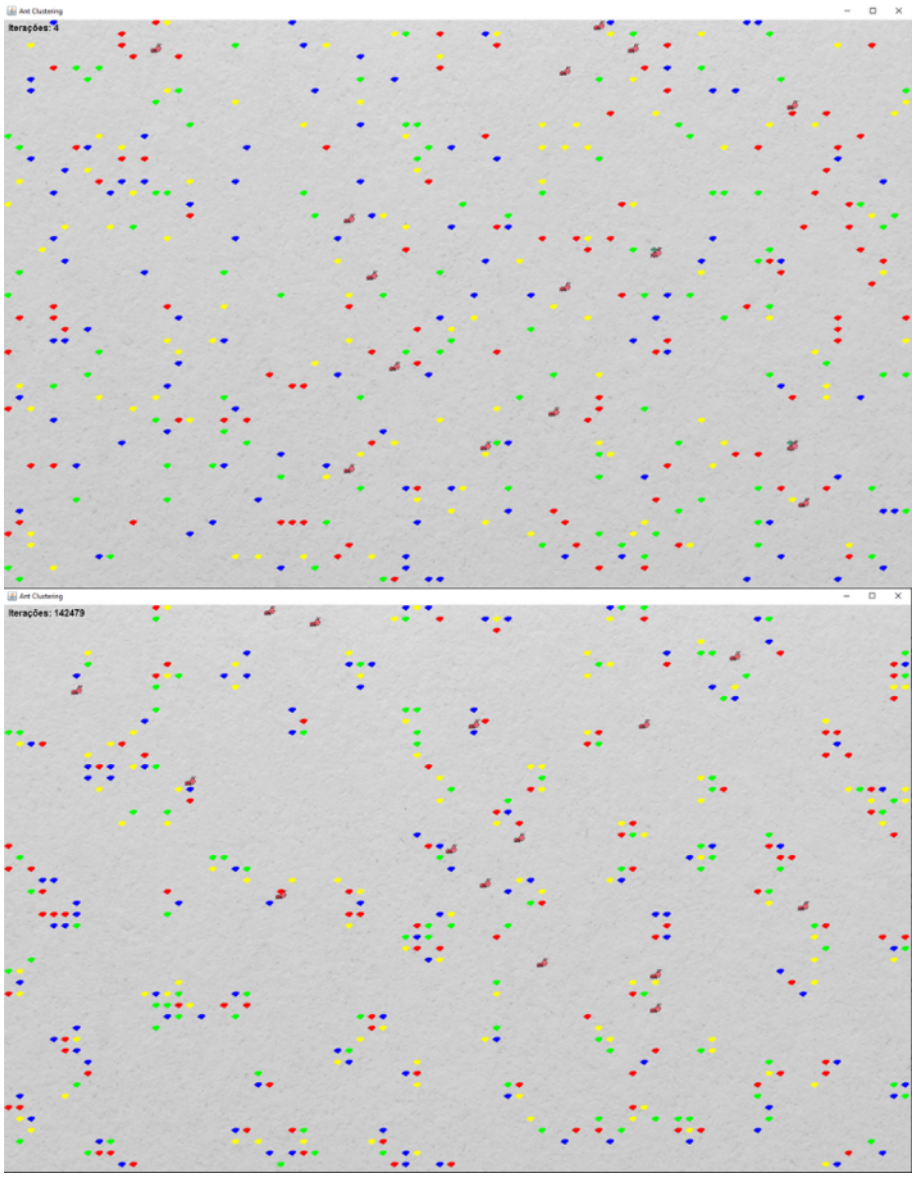

Figura 4. Ambiente inicial e final cenário 1 - alfa = 10, k1 e k2 = 0.7

Inicialmente, para obter um valor inicial para alfa nos dados heterogêneos, foi feito a média das distâncias, obtendo como resultado o valor 21.45 porém, esse valor não teve efeito de agrupamento. Apenas com valores próximos de 50 esse comportamento foi notado.

5. Conclusão

Com esse trabalho, podemos observar a grande capacidade dos algoritmos bioinspirados baseados em enxames, e como um simples comportamento social indireto de formigas, pode ser aplicado por exemplo, no agrupamento de dados semelhantes para resolver problemas complexos no mundo real. Além disso, é importante observar que problemas estocásticos, apesar de serem resolvidos da mesma maneira, com os mesmo parâmetros, o cenário final pode mudar assim como o tempo para resolver determinado problema, como em Big Data e Data Mining.

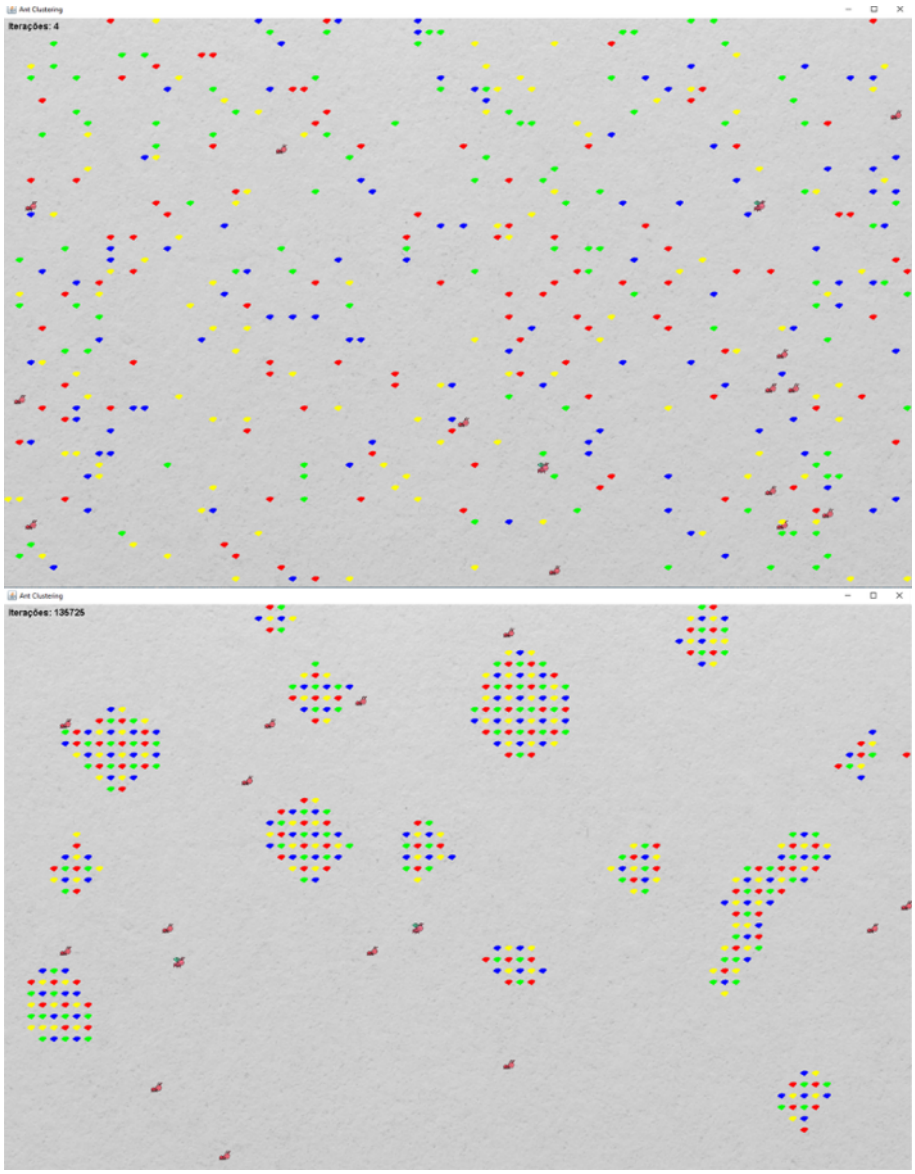

Figura 5. Ambiente inicial e final cenário 2 - alfa = 25, k1 e k2 = 0.7

Como trabalho futuro, seria interessante a presença de uma quantidade maior de dimensões nos dados e um novo desafio, como por exemplo, implementar paralelismo. Também seria interessante observar o comportamento das formigas com um raios maior de visão podendo ser variada ou fixa.

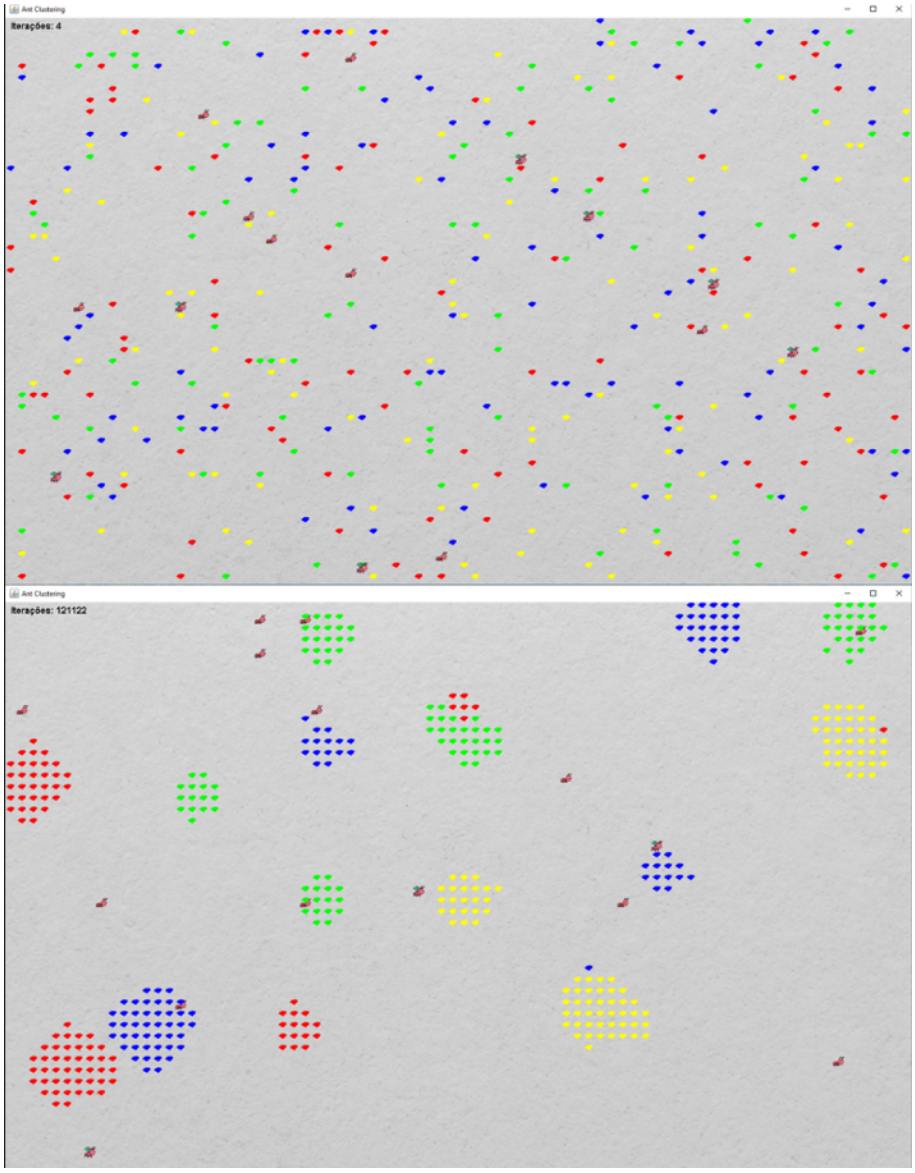

Figura 6. Ambiente inicial e final cenário 3 - alfa = 1, k1 e k2 = 0.7

Referências

Jafar, O. M. and Sivakumar, R. (2010). Ant-based clustering algorithms: A brief survey. *International journal of computer theory and engineering*, 2(5):787.

Melo, L. I. d. A., Costa, E., and Pereira, F. (2019). Self adaptation in ant colony optimisation. Publishing Press.

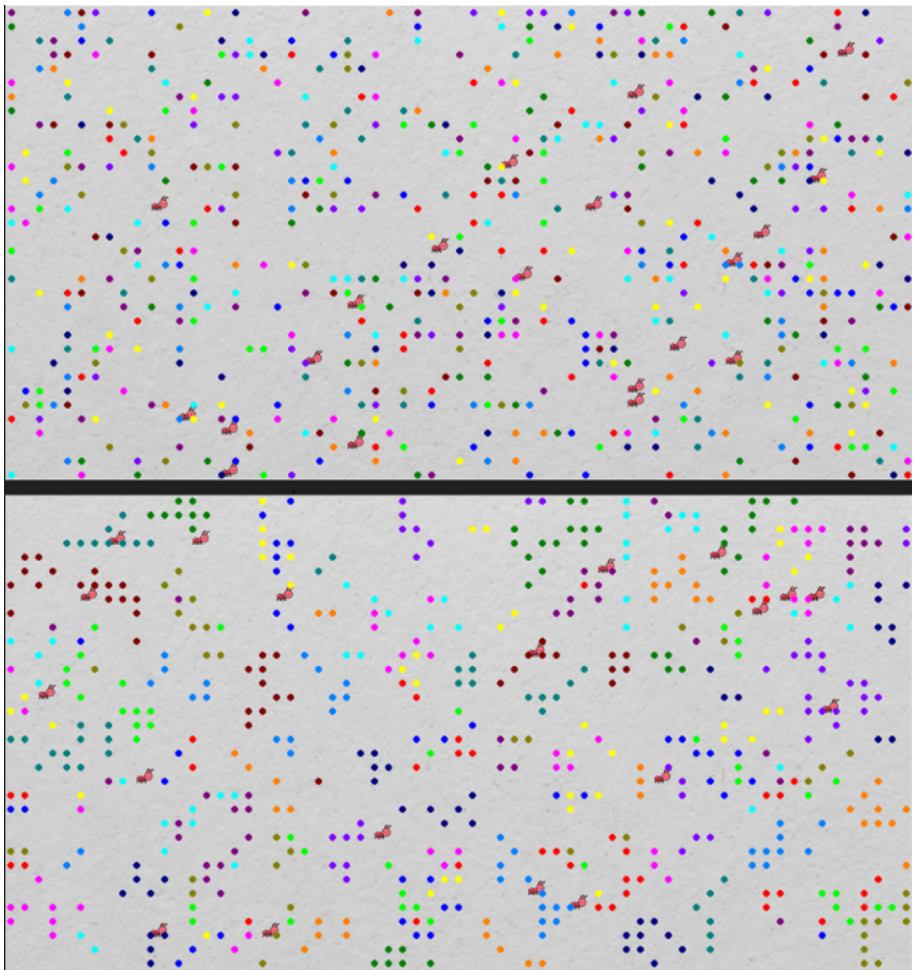

Figura 7. Ambiente inicial e final cenário 3 - alfa = 1, k1=0.1 e k2 = 0.1