Trabalho apresentado no XXXIX CNMAC, Uberlândia - MG, 2019.

**Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**

# Clusterização de dados utilizando o algoritmo de enxame de vagalumes

Victor de Mendonça Petta [1]
Ádrea Lima de Sousa [2]
Glécia Bezerra Rocha [3]
Itamar da Silva Pinto [4]
Juan Ferreira Vidal [5]
Orlando Fonseca Silva [6]
Faculdade de Engenharia Elétrica e Biomédica, Universidade Federal do Pará

## 1 Introdução

O avanço da tecnologia proporcionou um grande aumento na quantidade de dados armazenados. Tais dados podem ser dos mais diversos tipos e provenientes das mais variadas fontes. Essa grande quantidade de dados pode ser analisada para que se possa retirar informações implícitas dos bancos de dados. Nesse contexto surge a mineração de dados com o objetivo de extrair essas informações [1]. Uma das tarefas da mineração de dados é o processo de clusterização, que consiste em agrupar dados que não possuem uma classificação prévia (rótulos) no banco. Onde os elementos pertencentes a um mesmo grupo possuem características comuns enquanto diferem de elementos de outros grupos formados.

Existem diversas ferramentas que buscam realizar esse agrupamento, entre elas recentemente tem-se observado a utilização de técnicas de Inteligência Computacional (IC), e dentre as abordagens envolvendo IC, tem-se a Inteligência de Enxame (IE), que utiliza a emulação de comportamentos coletivos apresentados por certos grupos de seres vivos. Dessa forma, este trabalho tem por objetivo mostrar como pode ser obtido o agrupamento de um banco de dados utilizando o algoritmo de enxame de vagalumes (FA, do inglês *Firefly Algorithm*) [2].

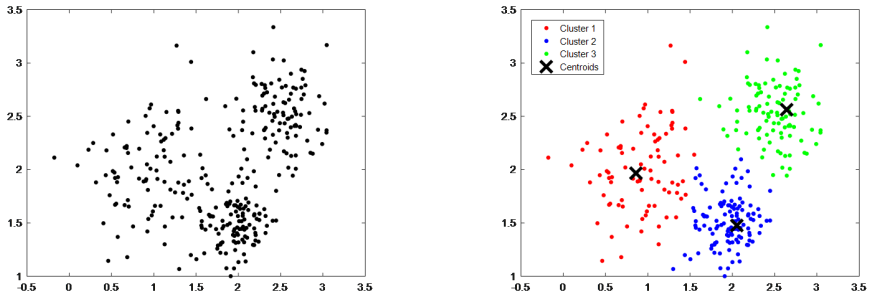
[1]victormpetta@gmail.com
[2]adrealsousa@gmail.com
[3]gleciarocha9@gmail.com
[4]itamarsp11@gmail.com
[5]jfvidal@ufpa.br
[6]orfosi@ufpa.br

## 2 Resultados

As simulações foram feitas utilizando o *software* Matlab. Primeiramente, foi criado um conjunto de dados aleatórios utilizando a Equação 1, sendo $C_n$ o centro onde o dado estará relacionado, $\sigma_n$ é o desvio padrão com o qual os dados serão criados em relação ao centro $n$ e $r$ é um número aleatório, Figura 1(a). Tendo como base os dados de entrada criados, o algoritmo é executado para otimizar a localização dos $k$ centroides que melhor separam os dados. Um vagalume $i$ do enxame será avaliado pela função objetivo representada pela Equação 2, onde o numerador representa a distância euclidiana média entre os dados pertencentes a diferentes clusters e o denominador a distância média entre os dados pertencentes ao mesmo cluster. O algoritmo retorna como solução os $k$ centroides que melhor agrupam os dados. Na Figura 1(b) pode-se ver os resultados obtidos pelo FA na alocação de conjunto de dados com $n = 3$. Como pode ser observado, o algoritmo apresentou um ótimo desempenho em alocar corretamente os centroides dos dados, obtendo apenas uma pequena diferença entre os valores reais de $C_n$ que foram utilizados para criar inicialmente os dados.

$$D_n = r * \sigma_n + C_n \tag{1}$$

$$F_i = \frac{D_{inter}}{D_{intra}} \tag{2}$$




(a) Dados não Alocados (b) Dados Alocados pelo FA

Figura 1: Dados do problema

## Referências

[1] M.Bramer. *Principles of data mining*. Springer, London, 2007.

[2] B. Xing and W. J. Gao. *Innovative computational intelligence: a rough guide to 134 clever algorithms*. Springer, New York, 2014.