Configuração de Parâmetro no Algoritmo de Agrupamento de Dados baseado em Formigas

Carina Moreira Costa[a,c], Rosangela Villwock[b,c]

[a] *Universidade de Trier, Trier, Alemanha*
[b] *Universidade Estadual do Oeste do Paraná, Cascavel, Paraná, Brasil*
[c] *carinamath5@gmail.com, rosangela.villwock@unioeste.br*

Resumo

O método de Agrupamento de Dados baseado em Colônia de Formigas é um método bio-inspirado no comportamento de colônias de formigas reais, o qual ainda precisa de muita investigação para se tornar uma ferramenta madura para mineração de dados. Este algoritmo possui a grande vantagem de não solicitar o número de grupos e uma desvantagem, a dificuldade de configuração de outros parâmetros. Nesse sentido, o objetivo deste trabalho foi melhorar a eficácia do algoritmo, propondo uma estratégia para configuração de um destes parâmetros, considerado o parâmetro mais difícil de ser calibrado. Para isso, foram realizados experimentos utilizando bases de dados reais e públicas, e com análises da matriz de dissimilaridade foi possível estabelecer um procedimento para definição do parâmetro. Ao final deste estudo, observou-se a utilidade e validade da estratégia proposta, pois se trata de uma forma simples de se definir o parâmetro e o desempenho do algoritmo foi razoável se comparado a execuções nas quais não se sabe como defini-lo ou nas quais se usa uma configuração padrão.

Abstract

The Ant Colony based Data Clustering method is a method bio-inspired on the behavior of real ant colonies, which still requires a lot of research to become a mature tool for data mining. This algorithm has the great advantage of not requiring the number of clusters and a disadvantage, the difficulty of setting other parameters. Therefore, the goal of this work was to improve the effectiveness of the algorithm by proposing a strategy to set up one of those parameters, considered to be the most difficult parameter to be calibrated. To this end, experiments were conducted using real-world and public data sets, and with analyses of the dissimilarity matrix it was possible to establish a procedure for defining the parameter. With this study, the usefulness and validity of the proposed strategy was verified, since it is a simple way to define the parameter and the performance of the algorithm was reasonable when compared with the experiments in which it is not known how to define the parameter or in which a default configuration is used.

Palavras-chaves: Mineração de Dados, Metaheurística, Matriz de dissimilaridade

Manuscrito submetido ao

1. Introdução

O algoritmo de Agrupamento de Dados baseado em Colônia de Formigas, diferente da maioria dos algoritmos de agrupamento de dados, possui a vantagem de não requerer como parâmetro o número de grupos [1] . Entretanto, este algoritmo possui outros parâmetros que são difíceis de serem determinados e alguns dependem da base de dados. O ajuste dos parâmetros é crucial para obter um bom agrupamento [2].

Estudos envolvendo este algoritmo são relevantes dado que trata-se de uma metaheurística relativamente nova e que tem recebido atenção especial, principalmente porque ainda exige muita investigação para melhorar seu desempenho, estabilidade e outras características, que fariam de tal algoritmo uma ferramenta madura para mineração de dados [3].

De acordo com Lauro [1], o problema da convergência do algoritmo é muito citado na literatura e já existem algumas propostas para resolvê-lo. Porém, estas propostas fogem um pouco do paradigma inicial do algoritmo, que tem como virtude a simplicidade. Conforme Faria, Stephany e Becceneri [2], a definição e calibração dos parâmetros do algoritmo para obter uma execução com bom desempenho têm dificultado e afastado os usuários deste.

Assim, o objetivo deste trabalho foi buscar a melhoria da eficácia do algoritmo de Agrupamento de Dados baseado em Colônia de Formigas por meio de um melhor ajuste de um de seus parâmetros.

2. Agrupamento de Dados baseado em Colônia de Formigas

No Agrupamento de Dados baseado em Colônia de Formigas proposto por Deneubourg et al. [4] as formigas foram representadas como agentes simples que se moviam aleatoriamente em uma grade quadrada. Os padrões foram dispersos nesta grade e poderiam ser carregados, transportados e descarregados pelas formigas. Estas operações são baseadas na similaridade e na densidade dos padrões numa vizinhança local. Após um número estipulado de iterações, se obtém uma separação espacial dos padrões na grade e os grupos formados tendem a formar regiões densas na grade.

Na proposta de Deneubourg et al. [4], as decisões de carregar e descarregar padrões são tomadas pelas probabilidades P_{pick} e P_{drop}, dadas pelas Equações (1) e (2). Nestas equações, $f(i)$ é uma estimativa da fração de padrões na vizinhança que são semelhantes ao padrão atual i da formiga e k_p e k_d são constantes reais. Os parâmetros k_p e k_d determinam a influência que a função densidade $f(i)$ exerce no cálculo da probabilidade de um agente carregar ou descarregar um objeto em determinada posição. A função de vizinhança $f(i)$ tem portanto grande influência nas decisões de carregar e descarregar padrões.

Lumer e Faieta [5] incluíram modificações ao modelo que permitiram a manipulação de dados numéricos com o uso de diferentes tipos de objetos e melhoraram a qualidade da solução e o tempo de convergência do algoritmo. Uma das mudanças é a da Equação (3), apresentada a seguir. Nesta equação, $d(i, j)$ é uma função de dissimilaridade entre padrões i e j pertencente ao intervalo $[0, 1]$, o parâmetro $\alpha \in (0, 1]$ vai escalar a dissimilaridade e L é a vizinhança local de tamanho igual a σ^2. Usualmente, $\sigma^2 \in \{9, 25\}$.

Handl, Knowles e Dorigo [6] também modificaram o algoritmo, especificamente o cálculo das probabilidades de carregar e descarregar padrões, bem como incluíram uma restrição na função de vizinhança dada pela Equação (4). Com esta restrição dissimilaridades elevadas são penalizadas, o que segundo os autores, melhora a separação espacial entre os grupos e acelera o tempo de execução do algoritmo.

$$P_{\text{pick}} = \left(\frac{k_p}{k_p + f(i)} \right)^2 \tag{1}$$

$$P_{\text{drop}} = \left(\frac{f(i)}{k_d + f(i)} \right)^2 \tag{2}$$

$$f(i) = max \left\{ 0,\ \frac{1}{\sigma^2} \sum_{j \in L} \left[1 - \frac{d(i,j)}{\alpha} \right] \right\} \tag{3}$$

$$\bar{f}(i) = \begin{cases} \frac{1}{\sigma^2} \sum_{j \in L} \left[1 - \frac{d(i,j)}{\alpha} \right], & \text{se} \quad \forall j \ \left(1 - \frac{d(i,j)}{\alpha} \right) > 0 \\ 0, & \text{caso contrário.} \end{cases} \tag{4}$$

3. Metodologia

A implementação do método utilizado neste trabalho foi a de Boscarioli et al. [7], na ferramenta YADMT - *Yet Another Data Mining Tool*. Maiores detalhes da implementação podem ser vistos em Costa [8]. O parâmetro σ foi escolhido empiricamente, para cada base de dados, dentre valores encontrados na literatura. O parâmetro escolhido para estudo neste trabalho foi o limitante de dissimilaridade, α, que vai "escalar" ou "limitar" a dissimilaridade na função de vizinhança $f(i)$. O funcionamento do algoritmo é crucialmente dependente deste parâmetro, sendo esse o parâmetro mais difícil de ser calibrado. A escolha de um valor muito pequeno para α impede a formação de grupos na grade, por outro lado, a escolha de um valor muito grande resulta na fusão de grupos [6].

A ideia foi usar a matriz de dissimilaridade normalizada no intervalo $[0, 1]$ e a restrição na função de vizinhança proposta por Handl, Knowles e Dorigo [6], dada pela Equação (4). Esta matriz apresenta as distâncias entre todos os padrões, e a partir da análise desta é possível estimar o parâmetro α. Cada base de dados terá um valor para α.

Para a definição deste parâmetro, foram calculadas as porcentagens de distâncias entre padrões maiores do que um certo valor (candidato a ser o valor de α), observado no intervalo [0,1; 0,9] considerando incremento de 0,01. O valor de α_{inicial} foi definido como o valor (a distância) que melhor diferenciava os dados na matriz de dissimilaridade (0,3 para a base de dados IRIS pois aproximadamente 56% das distâncias entre padrões são maiores que 0,3). Assim, de acordo com a restrição que penaliza dissimilaridades elevadas na Equação (4), temos que $d(i, j) > \alpha_{\text{inicial}}$ para 56% das distâncias e, portanto, os padrões que atendem a esta condição são considerados dissimilares, o que resulta em uma separação boa dos dados

já no início do algoritmo. Para uma estimativa do parâmetro $\alpha_{\max}$ foi definido um intervalo de valores observando a matriz de dissimilaridade. O limite inferior foi definido como o valor cuja porcentagem de distâncias entre padrões maiores que ele era aproximadamente 40% e o limite superior foi definido como o valor cuja porcentagem de distâncias entre padrões maiores que ele era aproximadamente 20%. Para uma estimativa do parâmetro α_{controle} foi definido um intervalo de valores. O limite inferior foi definido como o valor cuja porcentagem de distâncias entre padrões maiores que ele era aproximadamente 80% e o limite superior foi definido como o valor cuja porcentagem de distâncias entre padrões maiores que ele era aproximadamente 70%. O parâmetro $\alpha_{\min}$ foi definido com o mesmo valor de α_{inicial}. Os percentuais acima foram definidos previamente.

4. Resultados e Discussão

Para a escolha do parâmetro α adotou-se um critério para a definição de um intervalo de busca e, após isso, os valores foram definidos empiricamente por meio de experimentos. Os resultados obtidos considerando-se as configurações definidas para cada base de dados podem ser vistos em Costa [8].

A Tabela 1 apresenta a comparação das médias das medidas de avaliação (Índice Aleatório - R, Medida F e porcentagem de acerto) obtidas neste trabalho com as que foram obtidas aplicando o algoritmo utilizando a configuração padrão da ferramenta utilizada (YADMT). Comparando os resultados obtidos observa-se a superioridade dos resultados deste trabalho, revelando a efiácia da estratégia proposta.

Tabela 1: Comparação dos resultados médios da aplicação do algoritmo com a estratégia proposta e resultados obtidos utilizando-se a configuração padrão da ferramenta.

Conjunto	Métrica	Estratégia proposta	Configuração padrão
IRIS	R	0,790	0,581
	F	0,785	0,489
	% de acerto	78,3	45,2
WINE	R	0,785	0,555
	F	0,799	0,477
	% de acerto	78,4	46,2
ZOO	R	0,830	0,746
	F	0,672	0,493
	% de acerto	56,6	38,6
PIMA	R	0,552	0,509
	F	0,663	0,585
	% de acerto	66,1	55,5

5. Considerações Finais

O algoritmo de Agrupamento de Dados baseado em Colônia de Formigas tem despertado interesse dos pesquisadores em mineração de dados pelos resultados promissores que oferece. Entretanto, a dificuldade de configuração e calibração dos parâmetros afasta alguns usuários deste. A estratégia proposta neste trabalho se mostrou válida e util, pois é uma maneira simples de se definir o parâmetro e o desempenho do algoritmo foi razoável se comparado a execuções nas quais não se sabe como definir ou nas quais se tem uma configuração padrão oferecida pela ferramenta em uso.

Referências

[1] LAURO, A. L.. Agrupamento de Dados Utilizando Algoritmo de Colônia de Formigas. Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2008.

[2] FARIA, G. F.; STEPHANY, S.; BECCENERI, J. C.. Uma Nova Estratégia Acoplada De Inicialização e Ajuste adaptativo do Parâmetro de Similaridade num Algoritmo de Agrupamento Baseado em colônia de formigas. 10a Conferência Brasileira de Dinâmica, Controle e Aplicações, São Paulo, p.415-419, ago. 2011.

[3] BORYCZKA, U.. Finding groups in data: Cluster analysis with ants. Applied Soft Computing 9, 61-70, 2009.

[4] DENEUBOURG, J. L.; GOSS, S.; FRANKS, N.; SENDOVA-FRANKS, A.; DETRAIN, C.; CHRÉTIEN, L. The dynamics of collective sorting: Robot-like ants and ant-like robots. Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 1, 356–365, 1991.

[5] LUMER, E.; FAIETA, B. Diversity and adaptation in populations of clustering ants. Proceedings of the Third International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3, 501–508, 1994.

[6] HANDL, J.; KNOWLES, J.; DORIGO, M. Ant-Based Clustering and Topographic Mapping. Artificial Life v. 12, 35-61, 2006.

[7] BOSCARIOLI, C.; TEIXEIRA, M. F.; VILLWOCK, R.; FAINO, T. M. O módulo de agrupamento de dados da ferramenta YADMT. In: V EPAC - Encontro Paranaense de Computação. V EPAC - Encontro Paranaense de Computação, Cascavel, 2013.

[8] COSTA, C. M. Algoritmo de Agrupamento de Dados Baseado em Colônia de Formigas. Monografia — Universidade Estadual do Oeste do Paraná, Cascavel, 2016.