Standard Ant Clustering Algorithm (SACA) no Agrupamento de Dados Ambientais

Samuel L. Ghellere[1], Ruano M. Pereira[1], Gabriel Felippe[1], Maicon B. Palhano[1], Kristian Madeira[1], Carlyle T. B. Menezes[2], Merisandra C. M. Garcia[1]

[1]Grupo de Pesquisa em Inteligência Computacional Aplicada do Curso de Ciência da Computação– Universidade do Extremo Sul Catarinense (UNESC)
Criciúma- Brasil

[2]Grupo de Pesquisa em Gestão de Recursos Hídricos e Restauração de ambientes Alterados do Curso de Engenharia Ambiental – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma - Brasil

{samuel_ghellere, ruanopereira, gabrielheavy}@hotmail.com, {kma, cbm, maiconpalhano, mem}@unesc.net

Abstract. *Data Mining is a technology that uses various algorithms with the purpose of discovering knowledge in databases, among them algorithms have become the Standard Ant Clustering (SACA), which is based on a model of collective behavior of ant species, more accurately in the organization of cemeteries. This collective behavior it is used by SACA to form groups of similar data, on this work was applied environmental indicators of the quality of water resources in the River Basin Urussanga.*

1. Introdução

O agrupamento é uma tarefa básica no processo de data mining, que auxilia na estruturação e compreensão do conjunto de dados original, identificando os grupos existentes em um conjunto de objetos [Han e Kamber 2006]. Dentre os métodos de agrupamento, tem-se os baseados em inteligência de enxames, no qual, possuem características de robustez, confiabilidade e auto-organização levando a algoritmos eficientes, o que os tornam atraentes para pesquisas científicas [Dorigo e Stützle 2004].

O estudo do comportamento coletivo que levam a espécie de formiga *Pheidole Pallidula* a organizar seus corpos mortos em grupos de itens semelhantes resultou no algoritmo SACA, desenvolvido por Lumer e Faieta (1994). Este algoritmo consiste em uma simulação de colônia de agentes de formigas que se move aleatoriamente em uma grade toroidal bidimensional, onde os objetos estão posicionados, não necessitando de nenhuma informação inicial a respeito da massa de dados que será particionada [Dorigo e Stützle 2004], [Lumer e Faieta 1994].

Neste trabalho implementou-se o algoritmo SACA em uma ferramenta de data mining, denominada Shell Orion Data Mining Engine, e aplicou-se uma base de dados da área ambiental, referente a informações sobre a qualidade das águas do Rio Urussanga na região carbonífera catarinense. A base com os dados pré-processados é composta por 425 registros coletados ao longo de pontos de monitoramento espalhados por essa bacia.

2. O algoritmo SACA no Agrupamento de Dados Ambientais

O algoritmo SACA foi implementado no módulo de agrupamento da Shell Orion, por meio da linguagem de programação Java e do ambiente de programação integrado NetBeans 7.0.1. Os parâmetros a serem informados (número de formigas, total de iterações, memória, *alfa*[0,1], *Kp*[0,1], *Kd*[0,1], campo de visão e número de passos) estão com valores padrão, porém o usuário pode alterá-los conforme a sua necessidade. Após a parametrização, espalham-se os dados e as formigas aleatoriamente na grade e o algoritmo é inicializado.

Na realização dos testes, para a definição dos parâmetros empregaram-se vários valores até se encontrar a parametrização ideal para a base de dados. Na análise dos resultados obtidos os parâmetros utilizados foram sempre os mesmos a fim de se comparar os resultados. Os resultados obtidos pelo algoritmo SACA na base de dados da área ambiental foram analisados por meio da quantidade e da qualidade dos *clusters* formados, empregando-se neste último os índices de validação C-Index e Índice de Dunn. A variação dos resultados foi demonstrada por meio do desvio padrão, média e coeficiente de variação.

3. Considerações Finais

Nos testes observou-se que a partir de um determinado momento o algoritmo SACA não forma *cluster*, ele apenas movimenta até o fim das iterações os objetos que não conseguiu agrupar, visto que ele não se adapta conforme a formação dos grupos, o que aumenta o tempo de processamento. A variação no número de *clusters* está relacionada as ações aleatórias do algoritmo, encontrando-se para esta base de dados uma grande variação conforme o coeficiente de 32,04%. O índice *C-index*, que avalia internamente cada *cluster*, obteve valores próximos de zero indicando que mesmo com a variação no número de *clusters*, estes foram formados por objetos similares. O SACA, se corretamente parametrizado, retorna grupos com índices de validação satisfatórios, porém é importante salientar que este é um método ainda em estudo.

Referências

Dorigo, M. e Stützle, T. (2004), Ant Colony Optimization, MIT Press.

Han, J. e Kamber, M. (2006), Data Mining Concepts and Techniques, Morgan Kaufmann.

Lumer, E.D. e Faieta, B. (1994) “Diversity and Adaptation in Population of Clustering Ants”,In: Third International Conference on Simulation of Adaptative Behaviour, p. 501-508.