# Algoritmos Sequenciais de Agrupamento e Estratégias de Refinamento Associadas

**Eduardo Machado Real[1], Maria do Carmo Nicoletti[2]**

[1]Universidade Estadual de Mato Grosso do Sul (UEMS)
Nova Andradina – MS – Brasil

[2]Faculdade Campo Limpo Paulista (FACCAMP)
Campo Limpo Paulista – SP – Brasil

eduardomreal@uems.br, carmo@cc.faccamp.br,

***Abstract.*** *This paper presents the main results of the investigation of a family of three sequential clustering algorithms, BSAS, MBSAS and TTSAS, as well as two post-processing strategies, merge and reassignment, to be coupled at the end of a sequential clustering process, aiming at refining its result. Experiments were conducted using the developed computational system SEQ_CLUSTER having as input data from several domains and a comparative performance analysis of the algorithms was carried out.*

***Resumo.*** *Este artigo apresenta os principais resultados da investigação de uma família de três algoritmos sequenciais de agrupamento denominados BSAS, MBSAS e TTSAS, bem como de duas estratégias de pós-processamento, merge e reassignment, a serem acopladas ao final de um processo de agrupamento sequencial, com o objetivo de refinar o seu resultado. Experimentos foram realizados usando o sistema computacional desenvolvido SEQ_CLUSTER, tendo como entrada dados de vários domínios e uma análise comparativa dos desempenhos dos algorithmos foi realizada.*

## 1. Introdução

Este trabalho apresenta uma versão condensada e parcial dos principais resultados obtidos durante a realização de uma pesquisa em nivel de mestrado acadêmico (descrita na íntegra em [Real 2014]). O trabalho de pesquisa investigou (1) algoritmos de Aprendizado de Máquina (AM), particularmente aqueles relacionados à aprendizado de máquina não supervisionado caracterizados como algoritmos de agrupamento; (2) técnicas de pré-processamento de dados e (3) dois índices de validação. Dentre os muitos algoritmos de agrupamento disponíveis na literatura, três deles, identificados como algoritmos sequencias de agrupamento, foram escolhidos. Algoritmos sequencias são razoavelmente simples, produzem um único agrupamento e, via de regra, fazem poucas 'varreduras' dos dados iniciais. Esse artigo descreve apenas o trabalho realizado relacionado aos três algoritmos e a dois outros, caracterizados como possíveis estratégias de refinamento de agrupamentos obtidos.

Como estabelecido em inúmeras publicações acadêmicas (ver [Theodoridis e Koutroumbas 2009], [Jain *et. al.* 1999], [Han e Kamber 2006]), o principal objetivo de um algoritmo de agrupamento é o de organizar um conjunto de dados em grupos, geralmente utilizando como critério a similaridade entre os dados. O conjunto inicial de dados organizado em grupos é o que, nesse trabalho, é identificado como agrupamento.

Tipicamente agrupamentos são usados para categorizar (cada grupo do agrupamento pode ser abordado como uma categoria ou classe), dados que são similares entre si. A investigação e uso de tais algoritmos são essenciais, principalmente diante do vasto número de situações do mundo real existentes; uma situação representada por um volume de dados pode ser melhor organizada e, consequentemente, melhor entendida, se tais dados forem agrupados, usando como critério a similaridade entre eles. Os dados x e y pertencem a um mesmo grupo se eles forem similares. Em [Theodoridis e Koutroumbas 2009] e [Jain *et. al.* 1999] podem ser vistas mais definições de agrupamentos, bem como uma organização de categorias e de classificações para os seus algoritmos.

## 2. Algoritmos Sequenciais de Agrupamento

Os três algoritmos sequenciais investigados foram (1) *Basic Sequential Algorithmic Scheme* (BSAS); (2) *Modified Basic Sequential Algorithmic Scheme* (MBSAS) e (3) *Two-Threshold Sequencial Algorithmic Scheme* (TTSAS). O BSAS é considerado uma generalização da proposta descrita em [Hall 1967]. Tanto o MBSAS quanto o TTSAS são considerados variações do BSAS, propostas com o objetivo de reparar algum pequeno problema do algoritmo BSAS. Os três algoritmos: (1) são razoavelmente simples e rápidos; (2) produzem um único agrupamento; (3) tendem a gerar agrupamentos compactos cujos grupos têm formas esféricas ou elipsóidais; (4) têm um ou poucos passos; (5) fazem poucas 'varreduras' dos dados iniciais. As entradas para os três algoritmos são:

(1) um conjunto de dados E (em que $E = \{E_1, E_2, ... , E_N\}$ ($|E| = N$) e cada dado $E_i$ ($1 \leq i \leq N$) é descrito como um vetor de M atributos ($AT=\{A_1, A_2, ..., A_M\}$);

(2) dois parâmetros defindos pelo usuário, i.e., um ou mais limiares ($\Theta$ para o BSAS e MBSAS; $\Theta_1$ e $\Theta_2$ para o TTSAS) e o número máximo de grupos (q) a serem criados (apenas para o BSAS e MBSAS).

A saída de cada um dos algoritmos é um agrupamento $G = \{G_1, G_2, ..., G_Z\}$ ($1 \leq Z \leq q$). Os grupos são definidos por meio de um cálculo de distância apropriado entre um dado e um grupo, levando em consideração o limiar associado a essa distância. Nos experimentos realizados foi implementada a distância euclidiana e cada grupo de um determinado agrupamento foi representado pelo *representativo*, um vetor no qual cada posição (que representa um determinado atributo) é a média dos valores daquele atributo, considerando todos os elementos do grupo.

A cada interação o algoritmo BSAS considera um próximo dado do conjunto E e, dependendo da distância do dado considerado aos grupos formados até então, executa uma dentre duas possíveis ações: (1) incorpora tal dado a um dos grupos de dados já existentes ou (2) dá início à formação de um novo grupo, incluindo tal dado nele. A decisão de executar (1) ou (2) é tomada antes que o processo de criação de todos os grupos que compõem o agrupamento tenha sido finalizado. Esse fato pode, eventualmente, provocar dois problemas: (1) a atribuição de um dado a um grupo não apropriado, uma vez que o grupo que seria o mais apropriado não foi ainda criado ou (2) um dado não ser alocado a qualquer dos grupos criados.

O MBSAS é considerado uma versão do BSAS na qual o processo de formação de grupos é refinado; o refinamento, entretanto, implica o algoritmo ter que processar o

conjunto de dados duas vezes. Já o TTSAS foi proposto com o intuito de minimizar a dependência de ambos, BSAS e MBSAS, tanto da ordem na qual os dados são processados quanto do valor atribuído ao parâmetro $\Theta$ (valores não adequados de $\Theta$ podem implicar indução de grupos não significativos). O algoritmo busca contornar o problema por meio da definição de uma região duvidosa e, para tal, usa dois limites, $\Theta_1$ e $\Theta_2$ tal que $\Theta_2 > \Theta_1$ – a criação da região implica uma verificação posterior, para a alocação do dado. A descrição completa dos três algoritmos pode ser encontrada em [Real 2014] e [Theodoridis e Koutroumbas 2009] e de parte dos experimentos, em [Nicoletti et al. 2013].

Os resultados dos três algoritmos sequencias investigados podem ser influenciados por dois fatores: (1) ordem de processamento dos dados – a ordem interfere tanto no número de grupos criados pelos algoritmos quanto aos dados que cada um dos grupos agrupa e (2) valores dos parâmetros fornecidos pelo usuário (q e $\Theta$ ou $\Theta_1$ e $\Theta_2$) – considerando o BSAS (ou MBSAS), se o valor atribuído a $\Theta$ for muito pequeno, grupos desnecessários podem ser criados e se for muito grande, um número reduzido de grupos (aquém do número apropriado) será criado.

## 3. *Merge* e *Reassingment* como Estratégias de Refinamento

Duas estratégias de refinamento, descritas em [Theodoridis e Koutroumbas 2009], foram implementadas como um processo pós-agrupamento i.e., a serem acopladas ao final da execução de qualquer dos três algorithmos (BSAS, MSAS e TTSAS), com vistas a melhorar os agrupamentos obtido por esses algoritnos. Os resultados dos três algoritmos podem ser melhorados em situações: (1) em que o agrupamento resultante possui grupos que estão suficientemente próximos para serem unidos em único grupo e (2) que buscam minimizar a sensibilidade à ordem dos dados (embora não tão crítico para o TTSAS).

Uma maneira de lidar com a situação (1) é por meio de um processo de junção dos dois grupos, implementado por meio do procedimento *merge*, que une grupos considerados próximos o suficiente (de acordo com um parâmetro definido pelo usuário, *Close*). A Figura 1 e a correspondente Tabela 1 mostram um exemplo no qual dois grupos, $G_1$ e $G_4$, são unidos, uma vez que estão 'suficientemente próximos', ou seja, a distância entre seus centróides é menor que o valor do parâmetro *Close*. Uma maneira de lidar com a situação (2) é por meio de um processo que redistribui alguns dos dados, para grupos mais apropriados; tal processo foi implementado como o procedimento *reassignment* que reatribui aqueles dados considerados deslocados – i.e., dados que, no agrupamento considerado, poderiam pertencer a outros grupos mais próximos àqueles aos quais pertencem.

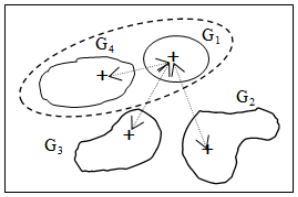


**Figura 1.** Agrupamento $G=\{G_1, G_2, G_3, G_4\}$ e a junção dos dois grupos mais próximos.

**Tabela 1.** Agrupamento $G=\{G_1, G_2, G_3, G_4\}$ e distâncias ($d_{ij}$) entre os pares de centroides.

| $G_i$ | $G_j$ | $d_{ij}$ | valor($d_{ij}$) |
|---|---|---|---|
| $G_1$ | $G_2$ | $d_{12}$ | 6 |
| $G_1$ | $G_3$ | $d_{13}$ | 3 |
| $G_1$ | $G_4$ | $d_{14}$ | 2 |
| $G_2$ | $G_3$ | $d_{23}$ | 5 |
| $G_2$ | $G_4$ | $d_{24}$ | 7 |
| $G_3$ | $G_4$ | $d_{34}$ | 4 |

## 4. Experimentos e Análise dos Resultados

Nos experimentos foram utilizados 6 conjuntos de dados reais, extraídos do repositório UCI Machine Learning Repository [Bache e Lichman 2013] e 3 sintéticos (artificialmente criados). Todos os experimentos foram realizados utilizando o ambiente computacional SEQ_CLUSTER, implementado durante o desenvolvimento do projeto de pesquisa de mestrado. Para cada conjunto de dados foram realizadas dez execuções aleatorizadas, associadas a cada um dos quatro possíveis esquemas investigados (cada esquema utilizado com cada um dos algoritmos (BSAS, MSAS e TTSAS)): (1) uso apenas do algoritmo, sem acoplamento de estratégia de refinamento, (2) uso do algoritmo, acoplado apenas ao *merge*, (3) uso do algoritmo acoplado apenas ao *reassignment* e (4) uso do algoritmo acoplado a ambos, *merge*+*reassignment*.

Os resultados das validações mostraram que, no geral, os algoritmos apresentaram um bom desempenho com relação aos agrupamentos obtidos e, em muitos casos, confirmaram que o uso das estratégias de refinamento e do pré-processamento de dados podem melhorar os resultados obtidos. Particularmente, a estratégia *reassignment* e *merge*+*reassignment* foram as mais eficientes quando usadas com o BSAS e MBSAS, e as estratégias *merge* e *merge*+*reassignment*, quando usadas com o TTSAS. Não pode ser esquecida, entretanto, a forte dependência que tal família de algoritmos tem da ordem na qual os dados são processados, bem como dos valores de parâmetros fornecidos. A descrição completa e detalhada do trabalho, bem como a especificação completa do sistema computacional desenvolvido, o SEQ_CLUSTER, estão em [Real 2014].

## Referências


Bache, K.; Lichman, M. (2013) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Hall, A.V. (1967) "Methods for demonstrating resemblance in taxonomy and ecology", Nature, pp. 830-831.

Han, J. e Kamber, M. (2006) Data Mining: Concepts and Techniques, Elsevier.

Jain, A. K., Murty, M. N. e Flynn, P. J. (1999) "Data clustering: a review", ACM Computing Surveys, (31), n.3, pp. 264-323.

Nicoletti, M.C., Real, E.M. e Oliveira, O.L. (2013) "The impact of refinement strategies on sequential clustering algorithms", In: Proc. of The 13th International Conference on Intelligent Systems Design and Applications (ISDA 2013), pp. 47-52.

Real, E. M. (2014) "Investigação de Algoritmos Sequenciais de Agrupamento com Pré-processamento de Dados em Aprendizado de Máquina", FACCAMP, 2014, 175 p. Dissertação. Programa de Mestrado em Ciência da Computação.

Theodoridis, S. e Koutroumbas, K. (2009) Pattern Recognition, 4ed., USA: Elsevier.