# Clusterização de uma Base de Dados Médica pelo Algoritmo Gustafson-Kessel

**José Márcio Cassettari Junior[1], Merisandra Côrtes de Mattos[1], João Manuel M. De Carlo[1], Priscyla Waleska Targino de Azevedo Simões[2], Cristian Cechinel[3]**

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

[2]Grupo de Pesquisa em Inteligência Computacional Aplicada – Curso de Ciência da Computação – Curso de Medicina – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma – SC

[3]Curso de Engenharia da Computação – Unipampa/Bagé – Universidade Federal de Pelotas (UFPEL) – Pelotas – RS

{capitalll, joaomdecarlo}@gmail.com, {mem, pri}@unesc.net, ccechinel.unipampa@ufpel.edu.br

***Abstract.*** *The improvement of storage models made possible the creation of great databases, making necessary the use of techniques that assist the exploration of these information. Amongst the available ones, it is distinguished application of data mining concepts. Considering this, this article demonstrates the application of the fuzzy logic algorithm Gustafson-Kessel for the clustering task, using for this a database contends medical registers of patients with sepsis, in order to make possible the identification of different groups in accordance with the characteristics of knowledge discovery in this database.*

***Resumo.*** *O avanço dos modelos de armazenamento possibilitou a criação de grandes bases de dados, tornando-se necessário a utilização de técnicas que auxiliem a exploração destas informações. Dentre as disponíveis, destaca-se a aplicação das tarefas e métodos de data mining. Considerando isto, este artigo demonstra a aplicação do algoritmo de lógica fuzzy Gustafson-Kessel para a tarefa de clusterização, utilizando para isto uma base de dados contendo registros médicos de pacientes com sepse, a fim de possibilitar a identificação de diferentes grupos de acordo com as características de descoberta de conhecimento nesta base.*

## 1. Introdução

O grande volume de dados gerado pelas organizações necessita ser transformado em conhecimento a fim de beneficia-las. Neste caso, é comum estas instituições utilizarem métodos estatísticos, porém apenas estes cálculos matemáticos não são suficientes para descobrir informações e gerar conhecimento.

Considerando isto, surgiu o conceito de *data mining*, que implementa técnicas de Inteligência Artificial, Estatística, Banco de Dados e Aprendizado de Máquina, para descoberta de conhecimento relevante em diferentes bases de dados. O processo de *data mining* envolve tarefas e métodos que possuem características específicas e são

aplicados conforme os objetivos da descoberta de conhecimento, em uma determinada base de dados [Kantardzic, 2003].

Estas tarefas e métodos de *data mining* encontram-se implementados em ferramentas, denominadas de *Shells*, as quais são utilizadas a fim de auxiliar na descoberta de padrões e relações relevantes. No entanto, grande parte destas aplicações são comerciais, tornando-se inviável para determinadas organizações [Berry e Linoff 2004].

Dentre estas ferramentas, existe em desenvolvimento a *Shell* Orion *Data Mining Engine*, projeto o qual objetiva a construção de uma ferramenta de *data mining* gratuita. Até o momento, a *Shell* Orion possui implementados os módulos referentes as tarefas de associação pelo algoritmo *APriori*, de classificação pelos algoritmos ID3 e CART e de clusterização pelos algoritmos *K-Means* e *Kohonen*.

A tarefa de clusterização, também conhecida como agrupamento, objetiva particionar a base de dados em grupos, denominados de *clusters*, onde cada elemento integrante seja similar aos outros que se encontram no mesmo grupo, diferenciando-os dos demais *clusters* [Goldschmidt e Passos 2005].

No entanto, quando a tarefa de clusterização é aplicada em grandes bases de dados, parte da informação pode estar localizada entre dois *clusters*. Nesta situação, existe a possibilidade destes dados serem forçados a pertencer a um grupo o qual não possua suas características, o que prejudica a execução da tarefa e por conseqüência, gera ambigüidade dos resultados obtidos [Bezdek et al 2005].

A fim de solucionar este problema existe, dentre os métodos de clusterização, o de lógica *fuzzy*, onde os elementos de um *cluster* podem pertencer a outros grupos ao mesmo tempo, dependendo dos graus de pertinência envolvidos, considerando que nesta pesquisa foi implementado o método de lógica *fuzzy* pelo algoritmo Gustafson-Kessel.

## 2. O Algoritmo Gustafson-Kessel

Em 1979, na *IEEE Conference on Decision and Control* na cidade de San Diego, Califórina, Donald E. Gustafson e William C. Kessel publicaram o artigo *Fuzzy Clustering with Fuzzy Covariance Matrix*. Neste, eles descrevem uma modificação do tradicional algoritmo *Fuzzy C-Means* (FCM). A modificação descrita no artigo foi intitulada como Gustafson-Kessel (GK), devido ao nome de seus autores.

A principal alteração em relação ao FCM foi a troca da distância euclidiana por outra que encontra com maior precisão os grupos existentes. Considerando isto, os autores adotaram a distância de Mahalanobis, a qual implementa uma matriz de covariância entre os atributos disponíveis na base de dados. Esta matriz possui a função de calcular a relação entre as diferentes propriedades, a fim de possibilitar maior flexibilidade ao determinar os grupos encontrados [Bezdek et al 2005].

A matriz de covariância permite ao algoritmo Gustafson-Kessel encontrar *clusters* de formas geométricas independentes, ou seja, cada grupo possui suas próprias características de dimensões. Por isso, os resultados gerados pelo GK são, em geral, superiores em relação aos algoritmos tradicionais e ao FCM [Gustafson e Kessel 1979].

Este algoritmo implementa o parâmetro *fuzzyficador* (*m*), o qual determina a *fuzzyficação* entre os elementos e seus protótipos. Se o valor de *m* for 1, não existirá

esta relação de *fuzzyficação* entre dados e grupos, o que gera uma clusterização tradicional, onde cada elemento pertence exclusivamente a um *cluster*. Normalmente, é utilizado o valor 2 para este parâmetro, pois desenvolve uma *fuzzyficação* satisfatória entre elementos e *clusters* [Cox 2005].

O algoritmo Gustafson-Kessel permite apenas a utilização de valores numéricos, pelo fato de todo o processo de clusterização realizar somente operações matemáticas. Se for necessário a utilização de atributos nominais, estes devem ser convertidos em valores decimais [Bezdek et al 2005].

A fim de possibilitar sua implementação na Shell Orion, além do levantamento bibliográfico, foi realizado a modelagem matemática do algoritmo, com o objetivo de compreender seu funcionamento e realizou-se testes para avaliar os grupos gerados pelo método.

## 3. O Algoritmo Gustafson-Kessel na *Shell* Orion *Data Mining Engine*

O algoritmo Gustafson-Kessel foi implementado no módulo de clusterização da *Shell* Orion *Data Mining Engine* por meio da linguagem de programação Java e do ambiente de programação Netbeans 6.0.1.

A clusterização por meio do algoritmo de lógica *fuzzy* Gustafson-Kessel necessita que alguns parâmetros de entrada sejam definidos, a fim de possibilitar a determinação correta dos grupos (Figura 1). Após executar o algoritmo, os resultados podem ser visualizados de diferentes maneiras. A Figura 2 demonstra o resumo da clusterização pelo algoritmo Gustafson-Kessel, onde são apresentados os parâmetros e atributos de entrada, informações sobre os grupos, os centros dos *clusters* e o atributo de saída, que pode ser alterado por meio da opção atributo de saída.

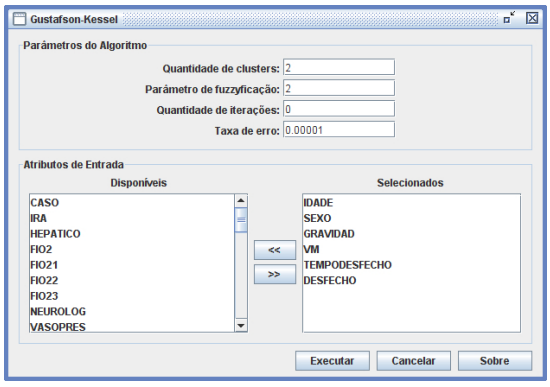


**Figura 1. Parâmetros de entrada**

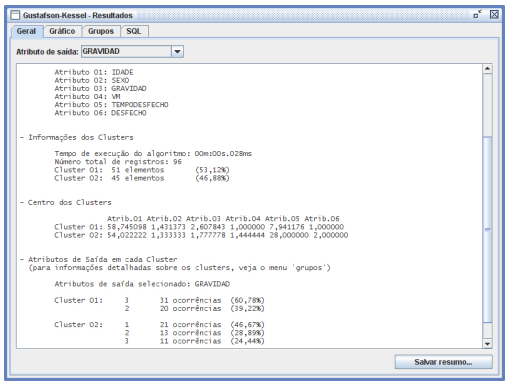


**Figura 2. Resumo da clusterização**

Os resultados também podem ser visualizados em forma gráfica (Figura 3), onde é demonstrado os grupos encontrados por meio de *Principal Component Analysis* (PCA). Este método realiza a transformação de uma base de dados contendo n dimensões em uma matriz de duas dimensões, por meio de sucessivas decomposições dos dados, com o objetivo de possibilitar a projeção dos elementos em um gráfico.

## 4. Resultados

O algoritmo implementado na *Shell* Orion *Data Mining Engine* foi testado com uma base de dados contendo 96 registros e 45 atributos de pacientes com sepse da UTI do

Hospital de Clínicas de Porto Alegre, Rio Grande do Sul. Utilizando para os parâmetros de entrada número de *clusters* e parâmetro *fuzzyficador* o valor 2, selecionou-se alguns atributos da base a fim de obter-se os resultados do algoritmo.

O objetivo deste teste foi determinar o desfecho dos pacientes, de acordo com os atributos de entrada *idade*, *sexo*, *gravidad* e *tempodesfecho*. Considerando isto, o atributo de saída selecionado foi *desfecho*. Os resultados apresentados pelo algoritmo foram precisos, considerando que os dois grupos foram encontrados corretamente. Os dados dos grupos estão descritos na Tabela 1 e estes são demonstrados em um gráfico de duas dimensões (Figura 3), onde o primeiro (azul) agrupa os pacientes curados e o segundo (vermelho) os pacientes que não se curaram.

**Tabela 1. *Clusters* distintos encontrados pelo algoritmo Gustafson-Kessel**

| *Cluster* | Elementos | Porcentagem | Desfecho |
|---|---|---|---|
| 1 | 45 | 46.88% | 2 (cura) |
| 2 | 51 | 53.12% | 1 (óbito) |

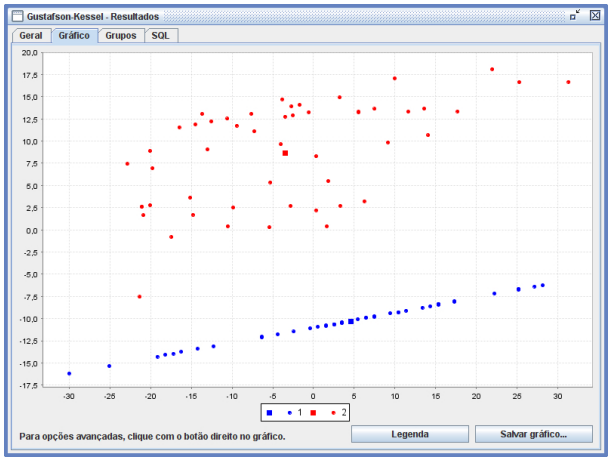


**Figura 3. Gráfico contendo dois grupos distintos**

Em outro teste realizado com os mesmos parâmetros, porém com atributos de entrada diferentes (*ira*, *vm*, *vasopres*, *apacheII*, *tba1* e *carbonyl*), tentou-se determinar o desfecho do paciente. No entanto, os resultados (Tabela 2) não foram precisos como os obtidos anteriormente, pelo fato dos atributos selecionados serem muito variáveis, pois se referem a exames feitos pelos pacientes. Além disso, a Figura 4 demonstra os dois grupos encontrados e pode-se perceber que os dados estão espalhados pelo gráfico, o que dificulta a identificação dos *clusters*.

**Tabela 2. *Clusters* encontrados pelo algoritmo Gustafson-Kessel**

| *Cluster* | Elementos | Porcentagem | Desfecho |
|---|---|---|---|
| 1 | 20 | 20.83% | 2 (cura) |
| 2 | 51 | 53.13% | 1 (óbito) |
| | 25 | 26.04% | 2 (cura) |

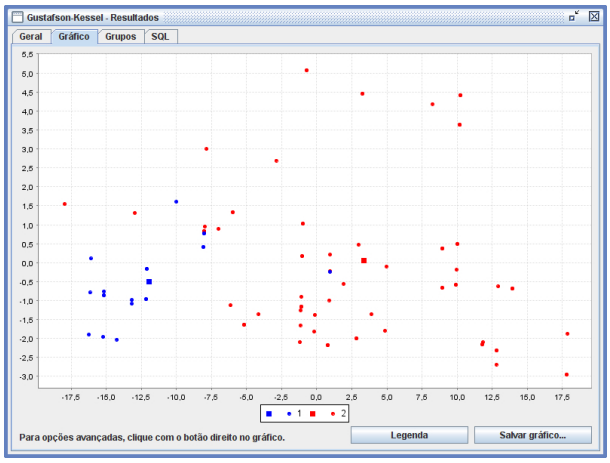


**Figura 4. Gráfico contendo grupos de difícil identificação**

## 5. Conclusões

A complexa tarefa de descoberta de novos conhecimentos em bases de dados pode ser simplificada utilizando-se os conceitos de *data mining*. Considerando suas diversas tarefas e métodos, esta técnica possibilita as organizações vantagem competitiva no que se refere a exploração de diferentes informações e no auxílio a tomada de decisão.

Entre as tarefas existentes, esta pesquisa fundamentou-se no entendimento da clusterização, especificamente sobre o algoritmo Gustafson-Kessel, o qual possibilita a identificação de grupos de diferentes formas e dimensões, por implementar a teoria da lógica *fuzzy*, que permite aos elementos pertencerem a diversos grupos simultaneamente, aumentando a precisão no particionamento dos dados.

Após a compreensão do método e a implementação do algoritmo Gustafson-Kessel na *Shell* Orion, foram executados alguns testes os quais demonstraram seu funcionamento correto e que alguns atributos de entrada podem influenciar diretamente nos resultados do algoritmo e deve-se determinar estes atributos com exatidão, conforme os objetivos de descoberta de conhecimento envolvidos.

## Referências

Berry, M. J. e Linoff, G. (2004), Data mining techniques: for marketing, sales, and customer relationship management, Wiley Publishing.

Bezdek, J., Keller, J., Krisnapuram, R. e Pal, N. (2005), Fuzzy models and algorithms for pattern recognition and image processing, Springer.

Cox, E. (2005), Fuzzy modeling and genetic algorithms for data mining and exploration, Morgan Kaufmann.

Goldschmidt, R. e Passos, E. L. (2005), Data Mining: um guia prático, Elsevier.

Gustafson, D. E. e Kessel, W. C. (1979), “Fuzzy clustering with fuzzy covariance matrix”, Proceedings of the IEEE Control and Decision Conference, San Diego, p. 761-766.

Kantardzic, M. (2003), Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons.