# SOM para dados relacionais baseados em múltiplas tabelas de dissimilaridade

**Anderson Berg dos Santos Dantas** [1]

[1]Centro de Informática - Universidade Federal de Pernambuco

absd@cin.ufpe.br

## 1. Justificativa

O projeto de pesquisa aqui apresentado surgiu da necessidade de realizar agrupamento de dados relacionais levando em conta simultaneamente múltiplas matrizes de dissimilaridade. Agrupamento de dados relacionais é mais utilizado em situações onde os dados não podem ser descritos por características numéricas, também é mais prático quando a distância possui alto grau de complexidade computacional ou quando grupos de objetos similares não podem ser representados eficientemente por um único protótipo.

Muitos algoritmos foram adaptados para analisar dados relacionais. [Kaufman and Rousseeuw 1987] apresenta uma adaptação do k-means para dados relacionais. Ainda, [Golli et al. 2004] apresenta um modelo de mapa auto-organizável por lote baseado em dados relacionais. O algoritmo apresentado neste trabalho é capaz de analisar dados levando em consideração a dissimilaridade entre observações. Os dados analisados são matrizes de dissimilaridade contendo a relação entre cada um dos objetos presentes na base de dados. Cada matriz representa uma variável da base e a dissimilaridade é calculada por uma função fixa que é a distância euclidiana entre os objetos.

Diversos métodos de análise de dados se baseiam em dados que podem ser descritos por valores reais, ou seja, por vetores em um espaço dimensional fixo e finito. Entretanto, muitos dados do mundo real requerem estruturas mais complexas para que sejam representados adequadamente. Textos, por exemplo, não são numéricos e possuem uma estrutura interna complexa que é difícil de representar em um vetor.

Em diversas situações, dados relacionais são descritos por múltiplas tabelas de dissimilaridade. Como apontado por [Frigui et al. 2007] muitas aplicações podem se beneficiar de algoritmos de agrupamento baseados em múltiplas matrizes de dissimilaridade. Na categorização de imagens, pode-se ter uma matriz com informações de cor, outra matriz com informação de textura e outra com informação de estrutura.

Porém, diferentes matrizes não são igualmente importantes, algumas podem ser redundantes, outras irrelevantes, ou ainda, podem interferir negativamente na formação dos agrupamentos. Para que haja uma formação adequada dos agrupamentos faz-se necessário o uso de pesos para cada matriz de dissimilaridade, pesos estes que dependem de cada agrupamento. O objetivo da ponderação sobre as matrizes é encontrar graus de relevância e identificar quais características descrevem melhor os dados, tornando o agrupamento mais significativo.

O modelo aqui apresentado utiliza diferentes pesos adaptativos para cada matriz de dissimilaridades. Esses pesos mudam a cada iteração do algoritmo e, além disso, são

diferentes de um agrupamento para outro, ou seja, cada matriz possui uma influência diferente sobre a formação de cada agrupamento. O cálculo dos vetores de pesos neste algoritmo foi inspirado pela abordagem utilizada para calcular pesos para cada variável em cada agrupamento no algoritmo de agrupamento dinâmico baseado em distâncias adaptativas ([Diday and Govaert 1977]). Para encontrar uma partição dos elementos, o método descrito leva em consideração simultaneamente a descrição relacional dos dados dada por múltiplas matrizes de dissimilaridade.

## 2. Revisão da Literatura

Métodos de agrupamento organizam um conjunto de itens em grupos de forma que itens localizados em um mesmo grupo possuem alto grau de similaridade, por outro lado, itens de grupos distintos têm alto grau de dissimilaridade. Estes modelos têm sido amplamente aplicados em campos como taxonomia, processamento de imagem, recuperação de informação e mineração de dados [Jain et al. 1999]. As técnicas mais populares de agrupamento são os métodos hierárquicos e de particionamento [Jain et al. 1999, Xu and Wunsch 2005].

Métodos hierárquicos fornecem uma hierarquia completa, isto é, uma sequência de partições aninhadas a partir dos dados de entrada. Métodos hierárquicos podem ser aglomerativos [Sneath and Sokal 1973, Zhang et al. 1996, Guha et al. 1998, Karypis et al. 1999, Guha et al. 2000] ou divisivos [Lance and Williams 1968, Gowda and Krishna 1978, Kaufman and Rousseeuw 1990, Guenoche et al. 1991, Chavent 1998]. Métodos aglomerativos fornecem uma sequência partições aninhadas iniciando com agrupamento trivial onde cada item está em um único grupo e termina com um agrupamento onde todos os itens estão no mesmo grupo. Um método divisivo inicia com todos os itens no mesmo grupo e executa um procedimento de divisão até que um critério de parada seja alcançado.

Métodos de particionamento buscam obter uma única partição dos dados em um número fixo de agrupamentos. Estes métodos frequentemente buscam uma partição que otimize (geralmente uma solução local) uma função objetivo. Para melhorar a qualidade do agrupamento, o algoritmo é executado diversas vezes com diferentes inicializações e a melhor configuração obtida do total de execuções é usada como saída do algoritmo de agrupamento. Métodos de particionamento podem ser divididos em agrupamentos HARD [Forgy 1965, Huang 1998, Kanungo et al. 2000, Hansen and Mladenoviae 2001, Su and Chou 2001] e agrupamentos fuzzy [Bezdek 1981, Hoeppner et al. 1999, Hathaway et al. 2000, Hung and Yang 2001, Kolen and Hutcheson 2002]. Agrupamentos HARD fornecem uma partição na qual cada objeto do conjunto de dados é atribuído a um e somente um grupo. Agrupamento fuzzy gera uma partição fuzzy que fornece um grau de atribuição a cada padrão em um dado grupo, que dá flexibilidade para expressar que este objeto pertence a mais de um grupo ao mesmo tempo.

Existem duas representações comuns dos objetos nos quais os modelos de agrupamento podem ser baseados: dados caracterizados ou relacionais. Quando cada objeto é descrito por um vetor de valores quantitativos ou qualitativos, os vetores que descrevem os objetos são chamados dados caracterizados. Quando cada par de objetos é representado por uma relação então temos dados relacionais. O modelo mais

comum de dados relacionais é o caso de uma matriz de dissimilaridades $R = [r_{il}]$, onde $r_{il}$ é a dissimilaridade pareada (geralmente uma distância) entre os objetos $i$ e $l$. Agrupamento baseado em dados relacionais é muito útil quando os objetos não podem ser representados por um vetor de valores, quando a medida de distância não tem uma forma definida, etc [Kaufman and Rousseeuw 1990, Lechevallier 1974, De Carvalho et al. 2009, Davenport et al. 1989, Hathaway and Bezdek 1994].

Os mapas auto-organizáveis de Kohonen [Kohonen 1990] (Self organizing maps - SOM) fazem parte do grupo de modelos de rede neurais não-supervisionadas e de aprendizado competitivo. A rede SOM possui propriedades de agrupamento e visualização. O objetivo destes modelos é encontrar uma estrutura lógica entre os dados fornecidos, não existe uma resposta esperada nem uma ação determinada que deva ser realizada. Para descobrir essa estrutura lógica, o algoritmo se utiliza de interações laterais entre os neurônios formando uma vizinhança. O neurônio que obtiver o melhor valor de similaridade para uma dada entrada é atualizado, da mesma forma neurônios vizinhos também são atualizados para representar melhor a entrada, resultando em regiões nas quais os neurônios são mais similares entre si.

O mapa auto-organizável pode ser considerado como um algoritmo que mapeia dados de alta dimensionalidade espacial em um espaço de dimensionalidade reduzida, geralmente uma, duas ou três dimensões. Esta projeção habilita o particionamento dos dados em grupos similares e possui a propriedade de preservar a topologia dos dados. A característica mais importante dos mapas auto-organizáveis é a possibilidade de comparar agrupamentos [Badran et al. 2005]. Cada objeto é afetado a um grupo e cada grupo é projetado em um nó do mapa. Objetos semelhantes são projetados no mesmo nó. A dissimilaridade entre os objetos projetados aumenta com a distância que separa os nós.

[Golli et al. 2004] e [Conan-Guez et al. 2006] propõem uma adaptação dos mapas auto-organizáveis em lote para dados de dissimilaridade. [Frigui et al. 2007] propôs um algoritmo de agrupamento fuzzy para dados relacionais (CARD), que é capaz de particionar objetos levando em conta múltiplas matrizes de dissimilaridade e que ainda calcula pesos que medem a relevância de cada matriz de dissimilaridade para cada um dos grupos. O CARD é baseado em algoritmos de agrupamento fuzzy para dados relacionais bastante conhecidos, o NERF [Hathaway and Bezdek 1994] e o FANNY [Kaufman and Rousseeuw 1990]. Como citado por [Frigui et al. 2007], diversas aplicações podem se beneficiar de algoritmos de agrupamento de dados relacionais baseados em múltiplas matrizes de dissimilaridade. No campo de categorização de base de dados de imagem, a relação entre os objetos pode ser descrita por múltiplas matrizes e a medida de dissimilaridade mais efetiva não possui uma forma definida ou não é diferenciável com respeito aos parâmetros do protótipo.

A ponderação de características deriva da seleção de características e tem sido um tópico de pesquisa importante em algoritmo de aprendizado não-supervisionado. Em [Qiang Wang and Huang 2008], os autores introduzem um algoritmo fuzzy k-means que tem a vantagem de trabalhar com ponderação para objetos e variáveis simultaneamente. [Grozavu et al. 2009] desenvolveu dois modelos usando mapas auto-organizáveis (SOM), que realizam simultaneamente agrupamento e ponderação de variáveis. [Frigui et al. 2007] propõe um algoritmo de agrupamento baseados em múltiplas tabelas de dissimilaridade (CARD) que calcula pesos relacionados à

relevância de cada matriz de dissimilaridade sobre cada agrupamento. Em outro trabalho, [Frigui and Nasraoui 2004] apresenta uma abordagem que realiza agrupamento e ponderação de variáveis simultaneamente.

## 3. Objetivos

A maioria dos métodos de análise de dados busca classificar novos objetos com base no conhecimento adquirido pela observação anterior de objetos semelhantes. Métodos de agrupamento, por outro lado, apenas buscam uma estrutura inerente aos dados, agrupando-os de acordo com as características semelhantes entre si. Existem diversos métodos de agrupamento baseados em dados relacionais, porém grande parte deles leva em consideração apenas uma tabela representando todos as variáveis da base de dados. Em diversas situações na área de análise de dados, a representação destes dados é melhor descrita através de múltiplas matrizes de dissimilaridade.

Pensando nesta limitação, é necessário a criação de modelos que sejam capazes de agrupar objetos levando em consideração a descrição desses dados, contidas em múltiplas tabelas de dissimilaridade, simultaneamente. Este trabalho propõe um modelo que seja capaz de tratar múltiplas tabelas de dissimilaridade simultaneamente, que possa, também, aprender e calcular pesos medindo a relevância de cada tabela na formação dos grupos. Além disso, o método aqui proposto possui propriedades de visualização, pois se baseia no algoritmo de mapas auto-organizáveis.

Como objetivos específicos podemos citar:

1. Desenvolvimento de novos métodos para agrupamento de dados relacionais baseados em múltiplas tabelas de dissimilaridade adaptando o algoritmo de mapa auto-organizável em lote original.
2. Os métodos devem ser capazes de aprender pesos que medem a relevância de cada matriz de dissimilaridade na formação dos grupos. Os pesos aprendidos são calculados para cada matriz e são diferentes de um grupo para outro (estimados localmente) ou podem ser iguais para todos os grupos (estimados globalmente).
3. Realização de experimentos para a análise dos resultados alcançados e validação dos métodos propostos. Nesta fase será necessária a implementação do mapa auto-organizável em lote original para a comparação de resultados.

## 4. Metodologia

Este projeto de pesquisa deverá ser executado considerando as atividades previstas a seguir. O projeto deverá ser concluído num prazo de 24 meses, podendo ser estendido por mais 6 meses.

1. Identificação dos principais trabalhos relacionados;
2. Estudo dos principais trabalhos relacionados;
3. Estudo e implementação de métodos de agrupamento baseados em dados relacionais;
4. Implementação de métodos de agrupamento de dados relacionais baseados em múltiplas tabelas de dissimilaridade;
5. Comparação experimental entre os modelos propostos e os modelos da literatura;
6. Escrita de artigos para conferências e periódicos;

7. Escrita da dissertação.

A ideia é concentrar esforços inicialmente na busca por novos trabalhos relacionados verificando suas relevâncias para o nosso problema. Neste momento, será reunido o estado da arte dos métodos de agrupamento, mais especificamente os mapas auto-organizáveis.

A fase seguinte será de implementação dos modelos propostos e de modelos encontrados na literatura. Serão realizados testes para validação inicial dos métodos. Nesta fase, também, serão reunidas bases de dados com o objetivo de realizar experimentos, que serão executados na fase seguinte.

Após realizados diversos experimentos os modelos implementados serão comparados com o objetivo de determinar sua utilidade e desempenho. De posse dos experimentos, inicia-se a fase de escrita de artigo, muito importante pois ela consolida o resultado do trabalho na comunidade científica. Os resultados obtidos com os métodos desenvolvidos serão relatados e uma documentação será gerada de forma a facilitar o acesso ao método por outros pesquisadores. A escrita da dissertação finaliza o trabalho de pesquisa.

## Referências


Badran, F., Yacoub, M., and Thiria, S. (2005). Self-organizing maps and unsupervised classification. *Neural networks: methodology and applications*, pages 379–442.

Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. *Plenum Press, New York*.

Chavent, M. (1998). A monotetic clustering method. *Pattern Recognition Letters*, 19:989–996.

Conan-Guez, B., Rossi, F., and Golli, A. E. (2006). Fast algorithm and implementation of dissimilarity self-organizing maps. *Neural Networks*, 19:855–863.

Davenport, J., Hathaway, R., and Bezdek, J. (1989). Relational duals of the c-means algorithms. *Pattern Recognition*, 22:205–212.

De Carvalho, F., M.Csernel, and Lechevallier, Y. (2009). Clustering constrained symbolic data. *Pattern Recognition Letters*, 30 (11):1037–1045.

Diday, E. and Govaert, G. (1977). Classification automatique avec distances adaptatives. *R.A.I.R.O. Informatique Comput. Sci.*, 11:329–349.

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics*, 21:768–780.

Frigui, H., Hwanga, C., and Rhee, F. C.-H. (2007). Clustering and aggregation of relational data with applications to image database categorization. *Pattern Recognition*, (40 (11)):3053–3068.

Frigui, H. and Nasraoui, O. (2004). Unsupervised learning of prototypes and attribute weights. *Pattern Recognition*, 37:567–581.

Golli, A. E., Conan-Guez, B., and Rossi, F. (2004). A self organizing map for dissimilarity data. *IFCS 2004 Proceedings*.

Gowda, K. and Krishna, G. (1978). Disaggregative clustering using the concept of mutual nearest neighborhood. *IEEE Transactions on Systems, Man, and Cybernetics*, 8:888–895.

Grozavu, N., Bennani, Y., and Lebbah, M. (2009). From variable weighting to cluster characterization in topographic unsupervised learning. *Proceedings of the International Joint Conference on Neural Networks*, pages 1005–1010.

Guenoche, A., Hansen, P., and Jaumard, B. (1991). Efficient algorithms for divisive hierarchical clustering. *Journal of Classification*, 8:5–30.

Guha, S., Rastogi, R., and Shim, K. (1998). Cure: An efficient clustering algorithm for large databases. In *Proc. ACM SIGMOD Int. Conf. Management of Data*, pages 73–84.

Guha, S., Rastogi, R., and Shim, K. (2000). Rock: A robust clustering algorithm for categorical attributes. *Information Systems*, 25 (5):345–366.

Hansen, P. and Mladenoviae, N. (2001). J-means: A new local search heuristic for minimum sum of squares clustering. *Pattern Recognition*, 34:405–413.

Hathaway, R. and Bezdek, J. (1994). Nerf c-means: non-euclidean relational fuzzy clustering. *Pattern Recognition*, 27 (3):429–437.

Hathaway, R., Bezdek, J., and Hu, Y. (2000). Generalized fuzzy c-means clustering strategies using lp norm distances. *IEEE Transactions on Fuzzy Systems*, 8 (5):576–582.

Hoeppner, F., Klawonn, F., and Kruse, R. (1999). Fuzzy cluster analysis: Methods for classification, data analysis, and image recognition. *Wiley, New York*.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:283–304.

Hung, M. and Yang, D. (2001). An efficient fuzzy c-means clustering algorithm. *Proc. IEEE Int. Conf. Data Mining*, pages 225–232.

Jain, A. K., Murty, M., and Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, 31 (3):264–323.

Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., and Wu, A. (2000). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions in Pattern Analysis Machine Intelligence*, 24 (7):881–892.

Karypis, G., Han, E., and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32 (8):68–75.

Kaufman, L. and Rousseeuw, P. (1990). Finding groups in data. *Wiley, New York*.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. *Statistical data analysis based on the L1-norm and related methods*, pages 405–416.

Kohonen, T. (1990). The self-organizing maps. *Proceedings of the IEEE*, 78:1464–1480.

Kolen, J. and Hutcheson, T. (2002). Reducing the time complexity of the fuzzy c-means algorithm. *IEEE Transactions on Fuzzy Systems*, 10 (2):263–267.

Lance, G. and Williams, W. (1968). Note on a new information statistic classification program. *The Computer Journal*, 11:195–197.

Lechevallier, Y. (1974). Optimisation de quelques criteres en classification automatique et application a l'etude des modifications des proteines seriques en pathologie clinique. *ThÃ¨se de 3eme cycle. Universite Paris-VI*.

Qiang Wang, Y. Y. and Huang, J. Z. (2008). Fuzzy k-means with variable weighting in high dimensional data analysis. *The Ninth International Conference on Web-Age Information Management*, pages 365–372.

Sneath, P. and Sokal, R. (1973). Numerical taxonomy. *Freeman, San Francisco*.

Su, M. and Chou, C. (2001). A modified version of the k-means algorithm with a distance based on cluster symmetry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (6):674–680.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16 (3):645–678.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: An efficient data clustering method for very large databases. In *Proc. ACM SIGMOD Conf. Management of Data*, pages 103–114.

## 5. Cronograma de Atividades

Aqui serão apresentadas as atividades previstas para os próximos seis meses.

Resumo das atividades:

1. Realização de experimentos.
2. Escrita de artigo científico para periódico contendo os resultados dos experimentos.
3. Escrita da dissertação.
4. Defesa da dissertação.

| | Meses | | | | | |
|---|---|---|---|---|---|---|
| Atividades | Mar | Abr | Mai | Jun | Jul | Ago |
| 1 | X | | | | | |
| 2 | X | X | X | | | |
| 3 | X | X | X | X | X | |
| 4 | | | | | | X |