

# **SOM para dados relacionais baseados em múltiplas tabelas de dissimilaridade**

**Anderson Berg dos Santos Dantas**<sup>1</sup>

<sup>1</sup>Centro de Informática - Universidade Federal de Pernambuco

absd@cin.ufpe.br

## **1. Justificativa**

## **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 [Qiang Wang and Huang 2008, 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, Hoepfner 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.

### 3. Objetivos

### 4. Metodologia

#### 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.
- 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.
- 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.
- 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.
- Hoepfner, 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*.
- 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. *Thèse 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