Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, Vol. 3, N. 1, 2015.
Trabalho apresentado no XXXV CNMAC, Natal-RN, 2014.

Agrupamento de Dados a partir do *SOM*: uma proposta baseada na extração de componentes conectados

Thiago M. Faino **Rosangela Villwock** **Clodis Boscarioli**
Universidade Estadual do Oeste do Paraná
Campus de Cascavel, PR
E-mail: {thiagofaino, rosangela.unioeste, boscarioli}@gmail.com

RESUMO

Mapa Auto-Organizável (ou *Self Organizing Map* - *SOM*) é um modelo de rede neural desenvolvido por Teuvo Kohonen [7], com aprendizado não supervisionado e competitivo, com habilidade de realizar mapeamentos que preservam a topologia entre os espaços de entrada e de saída, uma propriedade observada no cérebro.

SOM é baseada em uma grade de neurônios, que normalmente é de uma ou duas dimensões. Uma grade de neurônios bidimensional pode apresentar topologia retangular ou hexagonal. Para cada topologia existe uma forma de ligação entre os neurônios, que define o tipo de vizinhança [5]. Segundo [8], sua arquitetura é formada por duas camadas: camada de entrada e a camada de saída. Cada neurônio da camada de saída é completamente conectado com todos os padrões do vetor de entrada. São apresentados os padrões do vetor de entrada à camada de saída e a cada padrão do vetor de entrada apresentado tem-se uma região de atividade na grade. A localização e natureza de uma determinada região variam de um padrão de entrada para outro. Assim sendo, todos os neurônios da rede devem ser expostos a um número suficiente de diferentes padrões de entrada, garantindo assim que o processo de auto-organização ocorra de forma apropriada [6].

Segundo Boscarioli [2], somente o mapeamento topológico do *SOM* não é suficiente para realizar uma análise de agrupamentos. Para a realização da análise de agrupamentos via *SOM*, outras técnicas devem ser utilizadas sobre o mesmo, para que os resultados sejam observáveis e úteis na geração de novos conhecimentos. O algoritmo SL- *SOM* [3], a metodologia de Vesanto e Alhoniemi [9], a metodologia de Boscarioli [2], entre outras encontradas na literatura, são técnicas utilizadas para agrupar dados a partir de um mapa *SOM* treinado.

Neste trabalho, uma metodologia para agrupamento a partir do *SOM* que se utiliza de um algoritmo de extração de componentes conectados foi desenvolvida e implementada [4]. A ideia desta metodologia é utilizar o algoritmo de extração de componentes conectados para a rotulação dos grupos na matriz de densidade de um *SOM* treinado.

A metodologia foi implementada no módulo de agrupamento da ferramenta YADMT - *Yet Another Data Mining Tool* [1], uma ferramenta de *KDD* (*Knowledge Discovery in Databases* ou Descoberta de Conhecimento em Bases de Dados) que está sendo desenvolvida na UNIOESTE.

Para a avaliação da metodologia implementada foram realizados experimentos utilizando três bases de dados reais e públicas (disponíveis no repositório da Universidade da Califórnia em Irvine, "*UCI Machine Learning Repository*"[1]: *Iris Plants Database*, *Pima Indians Diabetes* e *Vehicle Silhouettes*) comparando os resultados com o método de agrupamento baseado em Formigas descrito em [10] e com o K-médias, que também estão presentes na ferramenta YADMT. Utilizou-se como métricas para avaliação a Medida F, o Índice aleatório R e a porcentagem de agrupamento correto.

Analisando os resultados obtidos (Quadro 01) com os experimentos pode-se concluir que a metodologia de agrupamento de dados a partir do *SOM* proposta apresentou um resultado satisfatório, mesmo não tendo as melhores medidas de avaliação para as bases de dados testadas. O K-médias apresentou melhor resultado para duas das três bases de dados (*Iris Plants Database* e *Vehicle Silhouettes*) e o agrupamento por Formigas apresentou melhor resultado (na maioria das medidas de avaliação) para a base de dados *Pima Indians Diabetes*, porém, para

[1] Disponível em: http://archive.ics.uci.edu/ml/datasets.html.

esta base de dados as medidas de avaliação ficaram muito próximas para os diferentes métodos. Além disso, para a base de dados *Vehicle Silhouettes* os resultados não ficaram tão próximos, mas estes foram ruins para todos os métodos. Somente para a base de dados *Iris Plants Database* o K-médias apresentou resultado consideravelmente melhor que a metodologia estudada. A vantagem da metodologia *SOM* proposta em relação ao K-médias é que não há necessidade de informar o número de grupos.

Quadro 01 – Resultados dos métodos Colônia de Formigas, K-médias e *SOM*

Base de Dados	Medidas de Avaliação	Formigas	K-médias	*SOM* Metodologia Proposta
Iris Plants Database	*F*	0,525	0,886	0,49
	R	0,533	0,892	0,33
	Agrupamento correto (%)	41	88	34
Vehicle Silhouettes	*F*	0,316	0,428	0,323
	R	0,319	0,420	0,281
	Agrupamento correto (%)	29	33	27,12
Pima Indians Diabetes	*F*	0,693	0,665	0,653
	R	0,701	0,621	0,625
	Agrupamento correto (%)	65	66	66,1

Como sugestão para trabalhos futuros recomenda-se a comparação da metodologia proposta com outras metodologias para agrupamento de dados a partir de *SOM*.

Palavras-chave*: Mineração de Dados, Agrupamento de Dados, Mapas Auto-Organizáveis*

Referências

[1] E. W. Benfatti; F. N. Bonifacio; A. D. Girardello; C. Boscarioli, "Descrição da Arquitetura e Projeto da Ferramenta YADMT - Yet Another Data Mining Tool", Relatório Técnico nº 01 do Curso de Ciência da Computação, UNIOESTE, Campus de Cascavel, 2010.

[2] C. Boscarioli, "Análise de agrupamentos baseada na topologia dos dados e em mapas auto-organizáveis", 2008, Tese (Doutorado em Engenharia Elétrica) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2008.

[3] J. A. F. Costa, "Classificação automática e análise de dados por redes neurais auto-organizáveis", 1999, Tese (Doutorado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica e Computação, Universidade Estadual de Campinas, São Paulo, 1999.

[4] T. M. Faino, "Agrupamento de Dados a partir de Mapas Auto-Organizáveis na Ferramenta YADMT", Trabalho de Conclusão de Curso – Ciência da Computação, Universidade Estadual do Oeste do Paraná, Cascavel, 2013.

[5] L. Fausett, "Fundamentals of Neural Networks – Architectures, Algorithms, and Applications", New Jersey, Prentice Hall, 1994.

[6] C. A. C. Francisco, "Rede de Kohonen: Uma ferramenta no estudo das relações tróficas entre espécies de peixes", 2004, Dissertação (Mestrado em Métodos Numéricos em Engenharia), Universidade Federal do Paraná, Curitiba, 2004.

[7] T. Kohonen, "Self-Organization and Associative Memory", 3a. Ed., Nova York, USA, Spring-Verlag, 1989.

[8] T. Kohonen, "Self-Organization", 3a. Ed., Nova York, USA, Spring-Verlag, 2001.

[9] J. Vesanto; E. Alhoniemi, Clustering of the Self-Organizing Map, *IEEE Transactions on Neural Networks*, v. 11, n. 3, p. 586–600, May 2000.

[10] R. Villwock, "Técnicas de Agrupamento e de Hierarquização no Contexto de *KDD* – Aplicação a Dados Temporais de Instrumentação Geotécnica-Estrutural da Usina Hidrelétrica de Itaipu", 2009, Tese (Doutorado em Métodos Numéricos em Engenharia), Universidade Federal do Paraná, Curitiba, 2009.

