Um método de agrupamento em dois estágios combinando mapas auto-organizáveis e *ant* k-médias

Jefferson R. Souza, Teresa B. Ludermir e Leandro M. Almeida

Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Caixa Postal 7851 – 50732-970 – Cidade Universitária, Recife – PE, Brazil

{jrs2, tbl, lma3}@cin.ufpe.br

Abstract. *This paper proposes a clustering method SOMAK, which is composed by Self-Organizing Maps (SOM) followed by the Ant K-means (AK) algorithm. SOM is an Artificial Neural Network (ANN), which has as one of its characteristics nonlinear projection from a high-dimensional space of sensory. AK is based in the Ant Colony Optimization (ACO), which is a recently proposed metaheuristic approach for solving hard combinatorial optimization problems. The AK algorithm modifies the K-means on locating the objects in a cluster with the probabilities which in turn, is updated by the pheromone. The SOMAK has a good performance when compared with some clustering techniques and to reduce the computational time.*

Resumo. *Este artigo propõe um método de agrupamento SOMAK, que é composto pelos Mapas Auto-Organizáveis (SOM) seguido do algoritmo Ant K-médias (AK). SOM é uma Rede Neural Artificial (RNA), que tem como uma de suas características projeção não-linear a partir de uma alta dimensionalidade do espaço sensorial. AK é fundamentado na Otimização baseada em Colônia de Formigas, que é uma abordagem meta-heurística recentemente proposta para resolver problemas de difícil otimização combinatória. O algoritmo AK modifica o k-médias localizando os objetos em seguida agrupando-os de acordo com as probabilidades que por sua vez, é atualizada pelo feromônio. O SOMAK tem um bom desempenho quando comparado com algumas técnicas de agrupamento e reduz o tempo computacional.*

1. Introdução

Com a diminuição substancial do custo de armazenamento de dados, e uma enorme melhoria no desempenho do computador e popularização das redes de computadores, grandes volumes de informações de dados estão sendo produzidos todos os dias em toda parte. Assim, grande número e escala de base de dados têm levado a necessidade de desenvolver algumas técnicas de processamento de dados úteis para o agrupamento de dados ou mineração de dados [Everitt et al 2001].

Análise de agrupamento está sendo utilizada em vários campos, tais como: Estatística, Reconhecimento de Padrões, Aprendizagem de Máquina e Mineração de Dados. Ela particiona um conjunto de dados ou objetos dentro de agrupamentos (ou grupos, classes).

SOM [Kohonen 1998] é uma RNA que permite a visualização de dados de alta dimensionalidade, também o mesmo, implementa um mapeamento ordenado de uma distribuição de alta dimensão dentro de uma grade regular de baixa dimensão. Esta grade ordenada pode ser utilizada como uma visualização conveniente para mostrar diferentes características da rede SOM.

AK é baseado na abordagem meta-heurística, proposta recentemente para solucionar problemas de difícil otimização combinatória denominada Otimização baseada em Colônia de Formigas ou *Ant Colony Optimization* (ACO) [Dorigo and Stützle 2000]. ACO é um método de otimização bio-inspirado que imita o comportamento das formigas, de forma a identificar caminhos mais curto entre o ninho e a fonte de alimento.Deve-se ressaltar que o objetivo neste artigo não é o de encontrar um agrupamento ótimo para os dados, mas obter uma visão sobre a estrutura de agrupamento dos dados, utilizando SOMAK.

No caso dos experimentos, uma comparação entre os resultados dos agrupamentos de dados direto (SOM e K-médias) e os métodos de agrupamentos baseados em dois estágios: a rede SOM seguida de K-médias (SOMK) e SOMAK é realizado.

2. Trabalhos Relacionados

Kuo et al. utilizou AK na análise de agrupamentos [Kuo et al. 2005]. O algoritmo AK modifica o K-médias localizando os objetos em seguida agrupando-os de acordo com as probabilidades, que por sua vez, é atualizada pelo feromônio, de acordo com o total de variância dentro do agrupamento ou t*otal within cluster variance* (TWCV). Os resultados experimentais mostraram que AK é melhor do que outros dois métodos, SOMK e SOM seguido pelo algoritmo K-médias Genético [Kuo et al. 2005]. O único problema para AK é que o número de agrupamentos é requerido, ou seja, é necessário fornecer o número de agrupamentos ao algoritmo AK para ele ser inicializado.

Juha Vesanto e Esa Alhoniemi combinaram SOM e K-médias [Vesanto and Alhoniemi 2000] para solucionar o problema de agrupamento. Em particular, o uso do agrupamento aglomerativo hierárquico e o agrupamento partitivo utilizando K-médias são investigados. O procedimento é composto de dois estágios, primeiro utilizando a SOM para produzir os protótipos, o qual são posteriormente agrupados no segundo estágio pelo K-médias. Os resultados do agrupamento utilizando a SOM como uma etapa intermediária foi computacionalmente eficaz, além de comparar os resultados obtidos diretamente a partir dos dados; tendo em vista as dificuldades originais provenientes das propriedades do algoritmo K-médias.

Tentando solucionar as necessidades dos algoritmos que foram vistos e descritos acima, é necessário desenvolver algumas técnicas de processamento de dados úteis para melhorar a solução do agrupamento de dados ou mineração de dados. Então, este artigo propõe um método de agrupamento baseado em dois estágios, combinando SOM e *Ant* K-médias para a análise de agrupamento.

3. Métodos de Agrupamentos

Nesta seção, há um estudo prévio sob métodos em dois estágios baseados na rede SOM, K-médias e *Ant* K-médias que será descrita com detalhes a seguir.

3.1. K-médias

O método de agrupamento K-médias é um dos mais simples algoritmos de aprendizagem não-supervisionado para solucionar o problema de agrupamento. O objetivo é dividir o conjunto de dados dentro de *k* agrupamentos fixados a priori. O algoritmo consiste em dois estágios: um estágio inicial e um estágio iterativo. O estágio inicial envolve a definição dos *k* centróides, um para cada agrupamento. O segundo estágio iterativo repete a assinatura de cada ponto de dados para o centróide mais perto, e *k* novos centróides são calculados de acordo com a nova assinatura [Mitchell 1997]. Esta iteração pára quando certo critério é encontrado; por exemplo, número de iterações. Dado um conjunto *nPad*, suponha que nós queremos classificar os dados dentro de *k* (grupos), o algoritmo tende a minimizar uma função de erro, tal como um erro médio quadrático definido como:

$$E = \sum_{k=1}^{C} \sum_{i=1}^{nPad} \left\| x_i - c_k \right\|^2 \quad (1)$$

Onde *C* representa o número de agrupamentos, *nPad* o número de amostras, *x* a entrada de cada amostra e c_k é o centro do agrupamento *k*.

3.2. Otimização baseada em Colônias de Formigas

O ACO foi proposto por Dorigo [Dorigo and Stützle 2000]. Quando nos referimos às colônias de formigas, observar-se que as formigas comunicam entre si apenas em uma forma indireta, através de seu ambiente, pela substância chamada feromônio. Caminhos com maiores níveis de feromônio serão mais prováveis de serem escolhidos e, portanto, reforçados, enquanto que a intensidade de feromônio sobre os caminhos que não são escolhidos é reduzida pela evaporação. Esta forma de comunicação indireta é conhecida, e prevê a colônia de formiga a capacidade de encontrar o menor caminho ou percurso [Martens et al 2007].

Existem alguns trabalhos relacionados aos algoritmos de agrupamento baseado em ACO. Yuqing et al. propôs algoritmo de agrupamento K-médias baseado na densidade e na Colônia de Formigas [Yuqing et al. 2003]. Este algoritmo é um novo algoritmo K-médias baseado na densidade e teoria de formigas, o qual solucionou o problema do mínimo local pelas formigas aleatórias, além de manipular os parâmetros iniciais do K-médias. Handl et al. propôs agrupamento baseado em formigas [Handl et al. 2003].

3.3. *Ant* K-médias

A escolha deste algoritmo é porque este produziu resultados satisfatórios, no que diz respeito ao problema de agrupamento. Neste método, é necessário fornecer o número de agrupamentos tal como no algoritmo K-médias convencional para o algoritmo AK. Seja

$$E = \{O_1, O_2, ..., O_n\}$$

o conjunto *n* dados ou objetos, onde O representa os objetos coleção a partir da base de dados, em que cada objeto tem k atributos, onde $k \succ 0$. Abaixo alguns parâmetros importantes, tais como:

• α : A importância relativa da trilha: $\alpha \geq 0$.
• β : A importância relativa da visibilidade: $\beta \geq 0$.
• ρ : O parâmetro de decaimento feromônio: $0 \prec \rho \prec 1$.
• Q : Uma constante.
• n : Número de objetos.
• m : Número de formigas.
• nc : Número de agrupamentos.
• T é o conjunto que inclui objetos usados. O número máximo de armazenamento pelo vetor T será de n, ou seja,

$$T = \{O_a, O_b, ..., O_t\}$$

Onde $a, b, ..., t$ são os pontos que as formigas têm visto ou visitado.
• T_k : O conjunto T é realizado por k formigas.
• $O_{centro}(T)$: O objeto o qual é o centro de todos os objetos em T, ou seja,

$$O_{center}(T) = \frac{1}{nT} \sum_{i=1}^{n} O_i \quad (2)$$

Onde nT é o número de objetos em T.
• $TWCV$: Total de variância dentro do agrupamento, ou seja,

$$\sum_{k=1}^{nc} \sum_{i=1}^{n} (O_i - O_{centro}(T_k))^2 \quad (3)$$

O Algoritmo 1 mostra o procedimento *Ant* K-médias em detalhes a seguir.

Procedimento Pertubação
Cada Formiga começa com o objeto aleatório e escolhe o centróide aleatoriamente do agrupamento para mover todas as k formigas.
Calculando $O_{centro}(T_k)$ onde k = *1, 2,..., nc* e TWCV.
Procedimento *Ant* K-médias
Entrar com o número de agrupamentos e os centróides correspondentes, e o conjunto de parâmetros α, β, ρ, número de iterações e formigas.
Estabeleça igualdade de feromônio para cada caminho.
Enquanto (o número de iterações não é satisfeito)
Faça
Atualizando o feromônio por $\tau_{ij} \leftarrow \tau_{ij} + \frac{Q}{TWCV}$.
Cada k formiga escolhe o centróide para mover com P, ou seja,

$$P = \frac{\tau_{kc}^{\alpha} . \eta_{kc}^{\beta}}{\sum_{i=1}^{nc} (\tau_{ki}^{\alpha} . \eta_{ki}^{\beta})}.$$

Calcular $O_{centro}(T_k)$ onde k = *1, 2,..., nc*.
Calcular TWCV (Total de Variância dentro do Agrupamento).
Enquanto (TWCV não for modificado)
Se TWCV menor do que o menor TWCV, substitua.
Procedimento Pertubação

Algoritmo 1. O procedimento *Ant* K-médias [Kuo et al. 2005]

3.4. Métodos em dois estágios baseados na rede SOM

Um método proposto de agrupamento baseado em dois estágios é útil para melhorar as principais desvantagens de um método de agrupamento partitivo; por exemplo, K-médias, devido à sua sensibilidade aos protótipos iniciais e da dificuldade de determinar um número apropriado de *k* agrupamentos.

Geralmente, um método de dois estágios baseados na rede SOM tem duas possíveis formas de trabalhar. Na primeira, a SOM é inicialmente utilizada para determinar o número de grupos e os centros dos grupos iniciais para o *Ant* K-médias. O centro inicial de um grupo pode ser obtido a partir do vetor peso correspondente aos centros dos grupos sobre a topologia da rede SOM. Na segunda forma, os mapas iniciais da rede SOM apresentam um grande conjunto de dados de escala sobre a sua topologia e gera as coordenadas topológicas dos protótipos para agrupamentos futuros no segundo estágio. O método utilizado na segunda fase é o procedimento *Ant* K-médias. A principal vantagem de um método de dois estágios baseados na rede SOM é a redução do tempo computacional pelos métodos de agrupamento hierárquico ou partitivo, para os conjuntos de dados grandes e complexos [Vesanto and Alhoniemi 2000].

Para mostrar esta característica foi necessário treinar uma rede SOM utilizando o algoritmo de treinamento seqüencial para o conjunto de dados[1]. Os mapas foram treinados em duas fases: um treinamento *rough* com largura da vizinhança inicial e taxa de aprendizagem grande e outra fase chamada *fine-tuning* com largura da vizinhança inicial e taxa de aprendizagem pequena. Ver Tabela 1.

Tabela 1. Parâmetros de treinamento SOM [Vesanto and Alhoniemi 2000]

Mapa	$\sigma_1(0)$ *	$\sigma_2(0)$ *
19 x 17	10	2

*Os parâmetros $\sigma_1(0)$ e $\sigma_2(0)$ são as larguras de vizinhanças iniciais para as fases *rough* e *fine-tuning*, respectivamente.

A largura da vizinhança diminui linearmente para 1 e a função de vizinhança utilizada foi a gaussiana. O tamanho de treinamento das duas fases foram de 3 e 10 épocas e taxas de aprendizagem iniciais foram 0.5 e 0.05, respectivamente.

4. SOMAK

O método proposto neste artigo, SOMAK, pode ser visto na Figura 1. SOMAK utiliza a rede SOM como classificador de características sobre os dados de entrada ao invés de agrupar os dados diretamente. Primeiro, um conjunto grande de protótipos é formado utilizando a SOM. Os protótipos podem ser interpretados como "proto-agrupamentos", que são na próxima etapa combinados para formar o verdadeiro agrupamento. Cada vetor de dados do conjunto de dados original pertence ao mesmo agrupamento como seu protótipo mais próximo.

[1] Treinamento da rede SOM, livremente disponível no pacote *Matlab SOM Toolbox* o qual foi utilizado na implementação do método proposto. Para maiores informações, ver URL http://www.cis.hut.fi/projects/somtoolbox/.

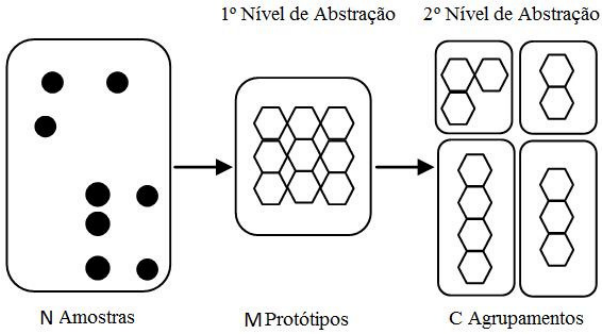

Figura 1. Primeiro nível de abstração é obtido através da criação de um conjunto de vetores protótipos usando a SOM. Algoritmo SOMAK cria o segundo nível de abstração realizando o agrupamento dos M protótipos [Vesanto and

No estudo atual, o número de agrupamentos e o centróide de cada agrupamento são gerados a partir da rede SOM. A fim de validar a solução de análise de agrupamento, o *framework* Monte Carlo [Milligan 1985] foi utilizado neste artigo.

O SOMAK utiliza a SOM para determinar o número de agrupamentos e os pontos iniciais e em seguida usa o procedimento *Ant* K-médias para encontrar a solução final.

O benefício desta abordagem é a redução do custo computacional. A segunda vantagem é a redução do tamanho de agrupamentos. A redução do ruído é outro benefício. Os protótipos são a média local dos dados e, portanto, menos sensíveis às variações aleatórias do que os dados originais.

Por este motivo, é conveniente agrupar um conjunto de protótipos, ao invés dos dados diretamente [Vesanto and Alhoniemi 2000]. Considere N amostras de dados utilizando o algoritmo *Ant* K-médias que está descrito na seção 3.3. Isto envolve fazer tentativas de agrupamento com diversos valores para o número de protótipos que foram obtidos pela rede SOM. A complexidade computacional é proporcional a $\sum_{k=2}^{C_{\max}} Nk$, onde $C_{\max}$ é o número máximo pré-estabelecido de agrupamentos e k representa o número de agrupamentos iniciais. Quando um conjunto de protótipos é utilizado em uma etapa intermediária (Figura 1 – 1º Nível de Abstração), a complexidade total é proporcional a $NM + \sum_{k} Mk$, onde M é o número de protótipos obtidos. Com $C_{\max} = \sqrt{N}$ e $M = 5\sqrt{N}$, a redução do custo computacional é baseada em $\frac{\sqrt{N}}{15}$, ou cerca de seis *folds* para $N = 10000$ [Vesanto and Alhoniemi 2000]. Em nosso caso, utilizamos dois *folds* e $N = 1000$ para a realização dos experimentos. Evidentemente, que esta é uma estimativa muito grosseira, visto que é uma estimativa em cima de outra; e muitas considerações práticas e experimentais são ignoradas.

5. Material e Métodos

Os experimentos realizados foram: dados sintéticos, dados reais, o método Monte Carlo, com o propósito de verificar a eficiência dos quatro métodos de agrupamento. Para realização dos experimentos foi utilizada uma máquina *Intel (R) Core (TM) 2 Quad*, processador 2.40GHz Q6600, memória RAM de 3,00 GB, sistema operacional *Microsoft Windows XP Professional* versão 2002 *Service Pack* 3. As bases de dados utilizadas neste artigo estão descritas em mais detalhes, na próxima seção.

5.1. Conjunto de Dados

Neste artigo foram utilizados cinco conjuntos de dados: *Lines, Banana, Highleyman* sendo estes classificados como dados sintéticos e *Contraceptive Method Choice* e *Glass* como dados reais.

A base ***Lines*** é composta por 1000 pontos de dados agrupados em 10 segmentos. As duas outras bases de dados[2] sintéticas, estão configuradas da seguinte forma: *A* representa um conjunto de dados de duas classes em duas dimensões; e *N* representa o número de amostras do vetor gerado com o número de amostras por classe. *N* = [500, 500] tendo um total de 1000 pontos de dados. Ainda a base de dados ***Banana***, retrata que os pontos de dados estão distribuídos em uma distribuição normal em forma de uma banana, com desvio padrão *S* = 1, em todas as direções. Já o terceiro conjunto de dados ***Highleyman***, além da configuração citada anteriormente, está dividida em duas classes: a **1° classe** contém 500 pontos de dados para cada uma das duas gaussianas com média 1 e 0 e variâncias 0 e 0.25, a **2° classe** contém 500 pontos de dados para cada uma das duas gaussianas com média de 0.01 e 0 e variâncias 0 e 4.

Os dados reais utilizaram o repositório *UCI* [Aha 2009]. A ***Contraceptive Method Choice*** **ou CMC** representa o problema de prever a escolha do método contraceptivo atual de uma mulher com base nas suas características econômicas e sócio-demográficas. O número de instâncias é de 1473, divididos em três classes. O número de atributos é 10, incluindo a classe. A segunda base real é a ***Glass***, esta base de dados tem como objetivo determinar se um vidro pertence a um tipo "*float*" ou não. O estudo de classificação deste tipo de vidro foi motivado por uma investigação criminológica, em que foram feitos vários testes sobre o vidro. O número de instâncias é de 214, divididos em seis classes. O número de atributos é 11, incluindo a classe.

Todos os dados sintéticos e as partições dos dados reais foram obtidos através do gerador de números aleatórios (Monte Carlo). Depois disto, utilizou-se a validação cruzada estratificada com dois *folds* sobre as bases de dados, fornecendo os conjuntos de treinamento e teste para todos os métodos de agrupamento. Assim, é razoável aceitar a confiabilidade do gerador de números aleatórios. Por fim, foram realizadas 30 execuções sobre o projeto.

5.2. Parâmetros

Os parâmetros considerados neste artigo são aqueles que afetam direta ou indiretamente as técnicas de agrupamento que já foram descritas nas seções anteriores para solucionar o problema de agrupamento. De acordo com [Dorigo and Stützle 2000], existem diversas combinações para determinar os parâmetros como aplicado ao algoritmo *ant colony system* ou sistema de colônia de formigas. Normalmente, os parâmetros são $\alpha = \{0,0.5,1,2,5\}$, $\beta = \{0,1,2,5\}$, $\rho = \{0.3,0.5,0.7,0.99,0.999\}$ e $Q = \{1,100,10000\}$. Existem 300 combinações de parâmetros, os resultados mostraram em [Berkhin 2009], que $\alpha = 0.5$, $\beta = 1$, $\rho = 0.9$ e $Q = 1$ neste método tem a menor variância. Onde m = 2 obteve melhores resultados comparado com m = 4 sugerido por Marco Dorigo [Dorigo and Stützle 2000]. A Tabela 2 mostra os parâmetros das técnicas de agrupamento.

[2] Conjunto de Dados, livremente disponível no pacote *Matlab PRTools:Toolbox for Pattern Recognition* foi utilizada. Para maiores informações, ver URL http://prtools.org/academic.html.

Tabela 2. Parâmetros principais das Técnicas de Agrupamento

Técnicas Agrupamento	Parâmetros
SOM	Número (Núm.) de atributos = quantidade dos padrões de entrada, taxa de aprendizagem inicial = 0.5 e final = 0.99, tamanho das linhas mapa = 19 e colunas = 17, raio inicial = 10 e final = 2, função de vizinhança = Gaussiana formato da vizinhança = hexa, tipo de treinamento = épocas, tamanho do treinamento para fase *rough* = 3 e a fase *fine-tuning* = 10.
K-médias	k = Núm. de agrupamentos iniciais, inicialização (inicializ.) dos centros = k.
SOMK	k = Núm. de protótipos SOM, inicializ. dos centros = Núm. de centróides SOM.
SOMAK	$\alpha = 0.5$, $\beta = 1$, $\rho = 0.9$, Q = 1, n = 500, m = 2, nc = Núm. de protótipos SOM.

6. Resultados Experimentais e Discussão

Após, encontrar o número de "proto-agrupamentos" que obtém como resultado 110 através da rede SOM; AK é usada para agrupar 500 amostras de dados sob o conjunto de teste. Tabela 3 mostra uma comparação entre SOMAK e SOMK para obter um menor número de agrupamentos. O número de agrupamentos e seus centróides são obtidos pela rede SOM e, em seguida, utiliza o AK para encontrar as soluções definitivas. SOMAK tem a melhor eficiência em comparação com SOMK, que é também o método composto de dois estágios.

Tabela 3. Resultados do tamanho de agrupamentos obtido pelo conjunto de teste

Dados	Agrupamentos iniciais	SOMK	SOMAK
Lines (I)	10	8	3
Banana (II)	2	6	4
Highleyman (III)	2	11	4
CMC (IV)	3	4	4
Glass (V)	6	4	3

SOMAK é aplicado como uma técnica de agrupamento para o estudo de caso porque obteve um valor menor de agrupamentos. A avaliação do processo do método proposto de agrupamento baseado em dois estágios inclui a comparação entre SOM, K-médias, SOMK e SOMAK. Em seguida, será apresentada uma medida de erro chamada Erro Médio Quadrático ou *Mean Squared Error* (MSE), que mostrou um valor menor para SOMAK quando comparado a SOMK; os parâmetros Min, Max, Med e Std representam respectivamente Mínimo, Máximo, Média e Desvio padrão, que podem ser vistos na Tabela 4. A medida de erro é expressa na Equação 4.

$$MSE = \sum_{k=1}^{C} \sum_{i=1}^{nPad} \frac{\left\| x_i - c_k \right\|^2}{nPad} \quad (4)$$

O parâmetro Std relatado na Tabela 4 apresentou um valor pequeno para a maioria dos métodos de agrupamento menos para SOMAK. Assim, a Tabela 4 mostrou também uma grande variabilidade destacada no parâmetro desvio padrão, resultando em uma desvantagem para o método proposto SOMAK.

Na grande maioria dos experimentos o método SOMAK mostrou um erro médio quadrático menor nos parâmetros de mínimo (Min), máxima (Max) e média (Med), quando comparado com os métodos de agrupamento SOM, K-médias e SOMK para todos os conjuntos de dados. A Tabela 4 também mostra o tempo computacional para todas as técnicas de agrupamento utilizadas nos experimentos.

Tabela 4. Resultados dos métodos com 30 execuções cada para obtenção do MSE e o Tempo Computacional (segundos)

Dados	Métodos	Resultados do MSE		Tempo Computacional	
		Min|Max|Med	Std	Min|Max|Med	Std
I	SOM	28.390|32.257|30.839	0.894	3.574|3.710|3.633	0.030
	K-médias	3.196|3.438|3.326	**0.051**	0.347|0.370|0.354	0.006
	SOMK	1.771|**2.108**|1.902	0.069	4.492|4.628|4.554	0.031
	SOMAK	**0.182**|2.767|**1.173**	0.845	4.276|4.409|4.334	0.030
II	SOM	6580.510|7285.590|6880.000	197.171	3.660|3.757|3.722	0.025
	K-médias	98.345|115.556|105.570	4.591	0.264|0.402|0.278	0.023
	SOMK	31.649|35.389|33.360	**1.014**	4.574|4.686|4.642	0.028
	SOMAK	**24.599|29.964|27.472**	1.291	4.387|4.484|4.448	0.026
III	SOM	560.414|716.573|635.081	37.154	3.590|3.701|3.626	0.019
	K-médias	9.930|14.207|11.823	1.042	0.263|0.295|0.278	0.009
	SOMK	2.867|3.387|3.068	**0.153**	4.516|4.616|4.553	0.022
	SOMAK	**0.906|1.577|1.265**	0.180	4.332|4.443|4.367	0.021
IV	SOM	26.022|28.244|27.313	0.520	20.619|23.101|22.119	0.564
	K-médias	0.465|0.538|0.508	0.015	0.278|0.300|0.286	0.005
	SOMK	0.251|0.292|0.273	0.011	21.792|24.265|23.322	0.562
	SOMAK	**0.050|0.058|0.055**	**0.002**	21.656|24.138|23.156	0.564
V	SOM	0.274|0.387|0.335	0.027	4.030|4.147|4.085	0.023
	K-médias	0.036|0.047|0.042	0.002	0.290|0.319|0.298	0.006
	SOMK	0.002|0.007|0.005	0.001	4.623|4.743|4.671	0.026
	SOMAK	**0.001|0.002|0.002**	**0.000**	4.175|4.294|4.231	0.024

K-médias tem sido sempre o mais rápido computacionalmente, porque é um algoritmo simples, ou de um único estágio. No entanto, este mesmo algoritmo apresentou um *MSE* alto, visto na Tabela 4, quando comparado com os métodos SOM, SOMK e SOMAK. SOMAK teve um tempo maior que o K-médias, e obteve resultados mais satisfatórios quando comparados com o SOMK, tanto no que se refere ao erro médio quadrático quanto ao tempo computacional.

Concluiu-se que os resultados experimentais são estatisticamente independentes, de acordo com a aplicação do teste t (teste de hipótese). Foi aplicado tanto para o erro médio quadrático quanto ao tempo computacional, respectivamente vistos na Tabela 4, e com 5% de grau de significância observou que o SOMAK é melhor do que o SOMK.

7. Conclusão e Trabalhos Futuros

O objetivo deste artigo foi propor um método de agrupamento, SOMAK, composto de dois estágios combinando SOM e *Ant* K-médias. O método SOMAK é capaz de reduzir o tamanho de agrupamentos, encontrando um bom desempenho quando comparado com algumas outras técnicas de agrupamento (SOM, K-médias e SOMK), e reduzir o tempo computacional dos experimentos.

Os algoritmos de agrupamento descritos anteriormente foram testados tanto para os dados diretamente quanto para os dados treinados pela rede SOM. Foi utilizada a rede SOM como uma etapa intermediária, além de realizar uma comparação dos resultados obtidos diretamente a partir dos dados.

Os resultados para os dados gerados pelo método Monte Carlo mostraram que SOMAK é melhor do que o SOMK, porque houve uma redução do tamanho de agrupamentos para o conjunto de teste (Tabela 3), por ter encontrado um melhor desempenho quando comparado com o SOMK visto (Tabela 4) e, finalmente, a Tabela 4

mostra também que a SOMAK reduziu o tempo computacional, em comparação com o SOMK para resolver o problema de agrupamento de dados.

Portanto, o método proposto é um método de agrupamento robusto. Pode ser aplicado a muitos tipos diferentes de problemas de agrupamento ou combinados com algumas outras técnicas de mineração de dados para obter resultados mais promissores.

Para trabalhos futuros, a idéia é reajustar o algoritmo SOMAK com o propósito de reduzir o tempo computacional deste quando comparado aos métodos descritos neste artigo. O primeiro método que será observado é o ABSOM [Sheng-Chai and Chih-Chieh 2008]. Este tem melhor desempenho do que a SOM e também funciona muito bem na análise de agrupamento em dois estágios quando é utilizada como uma técnica de pré-processamento. Assim, este método é composto de dois estágios para a análise dos dados, onde tem mostrado ser útil e eficaz.

8. Agradecimento

Este trabalho foi apoiado pelo CNPq e FACEPE.

Referências

AHA, D. (2009). UCI machine learning repository, 1987. http://archive.ics.uci.edu/ml/. Acesso em 16 de jan. 2009.

BERKHIN, P. (2009). Survey of Clustering Data Mining Techniques, Accrue Software. Available: http://www.accrue.com/. Acesso em 07 de jan. 2009.

DORIGO, M. and STÜTZLE, T. (2000). The Ant Colony Optimization Metaheuristic: Algorithms, Applications, and Advances, Technical Report IRIDIA.

EVERITT, B. S., LANDAU, S., and LEESE, M. (2001). Cluster Analysis, Edward Arnold, London.

HANDL, J., KNOWLES, J. and DORIGO, M. (2003). Ant-Based Clustering: A Comparative Study of its relative performance with respect to k-means, average link and 1D-SOM, IRIDIA-Technical Report Series.

KOHONEN, T. (1998). The self-organizing map, Neurocomputing, vol. 21, pp. 1-6.

KUO, R. J., WANG, H. S., TUNG-LAI HU and CHOU, S. H. (2005). Application of Ant K-Means on Clustering Analysis, Computers & Mathematics with Applications. vol. 50, nº 10-12, pp. 1709-1724.

MARTENS, D., DE BACKER, M. and HAESEN, R. (2007). Classification with Ant Colony Optimization, IEEE Transactions on Evolutionary Computation, vol. 11, nº 5, pp. 651-665.

MILLIGAN, G. W. (1985). An Algorithm for generating Artificial Test Clusters, Psychometrika, Springer New York, vol. 50, n° 1, pp. 123-127.

MITCHELL, T. (1997). "Machine Learning". McGraw-Hill, 352p.

SHENG-CHAI, C. and CHIH-CHIEH, Y. (2008). A Two-stage Clustering Method Combining Ant Colony SOM and K-means, Journal of Information Science and Engineering, vol. 24, pp. 1445-1460.

VESANTO, J. and ALHONIEMI, E. (2000). "Clustering of the Self-Organizing Map," IEEE Transactions on Neural Networks, vol. 11, nº 3, pp. 586-600.

YUQING, P., XIANGDAN, H. and SHANG, L. (2003). "The k-means clustering algorithm based on density and ant colony," IEEE Intelligent Neural Networks and Signal Processing, vol. 1, pp. 14-17.