Desempenho do Agrupamento Baseado em Formigas com relação ao Método Ward e aos Mapas de Kohonen Unidimensionais

Rosangela Villwock (UNIOESTE) rosangela@unioeste.br

Maria Teresina Arns Steiner (UFPR) tere@ufpr.br

Paulo Henrique Siqueira (UFPR) paulohs@ufpr.br

Resumo: Os comportamentos coletivo e auto-organizável de insetos sociais inspiraram pesquisadores a reproduzir este comportamento. Métodos inspirados em formigas são uma grande promessa para problemas de agrupamento. Os objetivos do presente trabalho foram propor alterações e melhorias no algoritmo de Agrupamento baseado em Formigas e avaliar o desempenho do algoritmo proposto com relação ao Método Ward e aos Mapas de Kohonen Unidimensionais. As principais modificações foram: substituição do padrão carregado pela formiga; comparação da probabilidade de descarregar um padrão numa determinada posição com a probabilidade de descarregar este padrão em sua posição atual; avaliação da probabilidade de descarregar um padrão para uma nova posição, caso o padrão não tenha sido descarregado na posição de sorteio, mas numa posição vizinha. Para a avaliação do desempenho do algoritmo foram utilizadas três bases de dados. Os resultados mostram que o desempenho do algoritmo proposto neste trabalho foi melhor do que o ACAM (Ant-based Clustering Algorithm Modified) para duas das três bases de dados. Quando comparadas as médias das medidas de avaliação na aplicação dos métodos de agrupamento Ward, Mapas de Kohonen Unidimensionais e Agrupamento proposto baseado em Formigas, os resultados não mostraram superioridade de algum método sobre os demais.
Palavras-chave: Mineração de Dados; Metaheurística; Agrupamento Baseado em Formigas.

1. Introdução

Sociedades de insetos sociais são sistemas distribuídos que apresentam uma organização social altamente estruturada, apesar da simplicidade dos seus indivíduos. Como resultado desta organização, colônias de formigas podem realizar tarefas complexas que, em alguns casos, excedem a capacidade individual de uma única formiga. Na área de pesquisa sobre "algoritmos de formigas", estudam-se modelos inspirados na observação do comportamento de formigas reais e usam-se estes modelos como fonte de inspiração para o desenvolvimento de novos algoritmos para a solução de problemas de otimização e de controle distribuído (DORIGO; STÜTZLE, 2004).

Entre os comportamentos dos insetos sociais, o mais amplamente reconhecido é a habilidade das formigas para trabalhar em grupo com o intuito de desenvolver uma tarefa que não poderia ser executada por um único agente. Também vista na sociedade humana, esta habilidade das formigas é um resultado de efeitos cooperativos. O efeito cooperativo recorre ao fato de que o efeito de dois ou mais indivíduos ou partes coordenadas é mais alto do que o total dos efeitos individuais. Alguns pesquisadores alcançaram resultados promissores em mineração de dados usando uma colônia de formigas artificiais. O número alto de indivíduos em colônias de formigas e a abordagem descentralizada para tarefas coordenadas (executadas de forma simultânea) significam que colônias de formigas mostram graus altos de paralelismo, auto-organização e tolerância a falhas. Estas características são desejadas em

técnicas de otimização modernas (BORICZKA, 2009).

O algoritmo de Agrupamento baseado em Colônia de Formigas foi escolhido para estudo, análise e novas propostas, devido a diversos fatores. Primeiramente, é uma metaheurística relativamente nova e tem recebido atenção especial, principalmente porque ainda exige muita investigação para melhorar seu desempenho, estabilidade e outras características consideradas "chaves", que fariam de tal algoritmo uma ferramenta madura para mineração de dados (BORYCZKA, 2009). Ainda, o referido algoritmo "consegue descobrir", automaticamente, a quantidade de grupos nos padrões.

Os objetivos do presente trabalho foram propor alterações e melhorias no algoritmo de Agrupamento baseado em Formigas originalmente proposto por Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo, 2006) e avaliar o desempenho do algoritmo proposto, comparativamente, com relação ao Método Ward e os Mapas de Kohonen Unidimensionais. O método da área de Estatística Multivariada (Método de Ward) foi utilizado por ser um dos métodos mais consagrados na literatura (JOHNSON; WICHERN, 1998). Os Mapas de Kohonen Unidimensionais foram utilizados porque, assim como o Agrupamento baseado em Formigas, executam as tarefas de agrupamento e mapeamento topográfico simultaneamente.

Este trabalho está estruturado da seguinte forma: na seção 2 é apresentada uma revisão bibliográfica sobre o Agrupamento baseado em Formigas, a descrição do algoritmo, a recuperação do agrupamento e as medidas para a avaliação de agrupamentos; na seção 3 são apresentadas as bases de dados utilizadas, detalhes de implementação dos métodos utilizados e as principais contribuições (modificações e melhorias) para o Agrupamento baseado em Colônias de Formigas; na seção 4 são apresentados os resultados e discussões e, finalmente, na seção 5, são apresentadas as conclusões.

2. Agrupamento Baseado em Formigas

O Agrupamento baseado em Formigas foi proposto inicialmente por Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo, 2006). Neste trabalho, as formigas foram representadas como agentes simples que se moviam aleatoriamente em uma grade quadrada. Os padrões foram dispersos dentro desta grade e poderiam ser carregados, transportados e descarregados pelos agentes (formigas). Estas operações são baseadas na similaridade e na densidade dos padrões distribuídos dentro da vizinhança local dos agentes, padrões isolados ou cercados por dissimilares são mais prováveis de serem carregados e então descarregados numa vizinhança de similares. As decisões de carregar e descarregar padrões são tomadas pelas probabilidades P_{pick} e P_{drop} dadas respectivamente pelas equações (1) e (2), a seguir:

$$P_{pick} = \left(\frac{k_p}{k_p + f(i)} \right)^2 \quad (1)$$

$$P_{drop} = \left(\frac{f(i)}{k_d + f(i)} \right)^2 \quad (2)$$

Nestas equações, *f(i)* é uma estimativa da fração de padrões localizados na vizinhança que são semelhantes ao padrão atual da formiga e k_p e k_d são constantes reais. No trabalho de Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo, 2006), os autores usaram $k_p = 0{,}1$ e $k_d = 0{,}3$. Neste trabalho, os autores obtiveram a estimativa *f*, através de uma memória de curto prazo de cada formiga, onde o conteúdo da última célula da grade analisada é armazenado. Esta escolha da função de vizinhança *f(i)* foi essencialmente motivada pela sua facilidade de

realização por robôs simples.

Lumer e Faieta (1994, *apud* Handl, Knowles e Dorigo, 2006) introduziram um número de modificações ao modelo que permitiu a manipulação de dados numéricos e melhorou a qualidade da solução e o tempo da convergência do algoritmo. A idéia era definir uma medida de similaridade ou dissimilaridade entre os padrões, já que no algoritmo proposto inicialmente, os objetos eram similares se fossem idênticos e dissimilares se não fossem idênticos. No referido trabalho aparece pela primeira vez o mapeamento topográfico.

Segundo Vizine *et al.* (2005), a idéia geral deste algoritmo é ter padrões semelhantes no espaço *n*-dimensional original em regiões vizinhas da grade, ou seja, padrões que são vizinhos na grade indicam padrões semelhantes no espaço original.

No trabalho de Lumer e Faieta (1994, *apud* Handl, Knowles e Dorigo, 2006), a decisão de carregar padrões é baseada na probabilidade P_{pick} dada pela equação (1) anterior e a decisão de descarregar padrões é baseada na probabilidade P_{drop} dada pela equação (3) a seguir, onde *f(i)* é dada pela equação (4).

$$P_{drop}(i) = \begin{cases} 2f & se \quad f(i) < k_d \\ 1 & se \quad f(i) \geq k_d \end{cases} \tag{3}$$

$$f(i) = \max\left\{0, \quad \frac{1}{\sigma^2}\sum_{j \in L}\left[1 - \frac{d(i,j)}{\alpha}\right]\right\} \tag{4}$$

Na equação (4), $d(i, j)$ é uma função de dissimilaridade entre padrões i e j pertencentes ao intervalo [0, 1]; α é um parâmetro escalar dependente dos padrões e pertencente ao intervalo [0, 1]; L é a vizinhança local de tamanho igual a σ^2, onde σ é o raio de percepção. Os autores usaram em seu trabalho $k_p = 0{,}1$, $k_d = 0{,}15$ e $\alpha = 0{,}5$.

Mudanças que melhoram a separação espacial dos grupos e permitem que o algoritmo seja mais robusto foram introduzidas por Handl, Knowles e Dorigo (2006). Uma delas é a restrição na função *f(i)* dada pela equação (5), a seguir, que serve para penalizar dissimilaridades elevadas.

$$f*(i) = \begin{cases} \dfrac{1}{\sigma^2}\displaystyle\sum_{j \in L}\left[1 - \frac{d(i,j)}{\alpha}\right] & se \quad \forall j \quad \left(1 - \frac{d(i,j)}{\alpha}\right) > 0 \\ 0 & caso \quad contrário \end{cases} \tag{5}$$

A separação espacial dos grupos na grade é crucial para que grupos individuais sejam bem definidos, permitindo a sua recuperação automática. A proximidade espacial, quando ocorrer, pode indicar a formação prematura do agrupamento (HANDL; KNOWLES; DORIGO, 2006).

A definição dos parâmetros da função de vizinhança é um fator decisivo na qualidade do agrupamento. No caso do raio de percepção σ, é mais atrativo empregar vizinhanças maiores para melhorar a qualidade do agrupamento e da distribuição na grade. Porém, este procedimento é mais caro computacionalmente (porque o número das células a serem consideradas para cada ação cresce quadraticamente com o raio), e ainda inibe a formação rápida dos grupos durante a fase de distribuição inicial. Um raio de percepção que aumenta gradualmente com o tempo acelera a dissolução de grupos pequenos preliminares (HANDL; KNOWLES; DORIGO, 2006). Um raio de percepção progressivo também foi utilizado por Vizine *et al.* (2005).

Segundo Handl, Knowles e Dorigo (2006), α determina a porcentagem de padrões na grade classificados como semelhantes. A escolha de um valor muito pequeno para α, impede a formação de grupos na grade; por outro lado, a escolha de um valor muito grande para α, resulta na fusão de grupos.

Fixar parâmetro α não é simples e a sua escolha é altamente dependente da estrutura do conjunto de dados. Um valor inadequado é refletido por uma excessiva ou extremamente baixa atividade na grade. A quantidade de atividade é refletida pela freqüência de operações com sucesso da formiga em carregar e descarregar. Com base nestas análises, Handl, Knowles e Dorigo (2006) propuseram uma adaptação automática de α. Já Boryczka (2009) propôs um novo esquema de adaptação para o valor de α.

2.1. O algoritmo básico

No algoritmo básico proposto por Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo, 2006), em sua fase inicial, todos os padrões são aleatoriamente espalhados na grade. Depois, cada formiga escolhe aleatoriamente um padrão para carregar e é colocada em uma posição aleatória na grade.

Na próxima fase, chamada de fase de distribuição, em um laço (*loop*) simples, cada formiga é selecionada aleatoriamente. Esta formiga se desloca na grade executando um passo de comprimento l, numa direção determinada aleatoriamente. Segundo Handl, Knowles e Dorigo (2006), o uso de um tamanho de passo grande acelera o processo de agrupamento. A formiga então decide, probabilisticamente, se descarrega seu padrão nesta posição.

Se a decisão de descarregar o padrão for negativa, escolhe-se aleatoriamente outra formiga e recomeça-se o processo. No caso de decisão positiva, a formiga descarrega o padrão em sua posição atual na grade, se esta estiver livre. Se esta célula da grade estiver ocupada por outro padrão, o mesmo deve ser descarregado numa célula imediatamente vizinha desta, que esteja livre, por meio de uma procura aleatória.

A formiga procura, então, um novo padrão para carregar. Dentre os padrões livres na grade, ou seja, dentre os padrões que não estão sendo carregados, a formiga seleciona aleatoriamente um, vai para a sua posição na grade, faz a avaliação da função de vizinhança e decide probabilisticamente se carrega este padrão. Este processo de escolha de um padrão livre na grade é executado até que a formiga encontre um padrão que deva ser carregado.

Só então esta fase é reiniciada, escolhendo-se outra formiga até que um critério de parada seja satisfeito.

2.2. Recuperação do agrupamento

O processo inicia com cada padrão formando um grupo. Depois de calcular as distâncias entre todos os grupos, deve-se fundir (ligar) os dois grupos com menor distância. Os tipos de ligações mais comuns são: Ligação Simples, Ligação Completa, Ligação Médias e Método de Ward (JOHNSON; WICHERN, 1998). As distâncias entre os grupos são definidas em termos de sua distância na grade. Cada padrão é agora composto de apenas dois atributos, que o posicionam na grade bidimensional. A distância entre cada dois padrões é então a distância euclidiana entre dois pontos da grade. Este processo se repete até que um critério de parada seja satisfeito.

Quando padrões em torno das bordas dos grupos estão isolados, Handl, Knowles e Dorigo (2006) introduziram um peso que incentiva a fusão destes padrões com os grupos.

2.3. Avaliação do agrupamento

Na avaliação de grupos, diferentes aspectos podem ser observados: determinação da tendência de agrupamento de um conjunto de dados, comparação dos resultados de uma análise de grupos com resultados externamente conhecidos, avaliação de quão bem os resultados de uma análise de grupos se ajustam aos dados sem referência a informação externa, comparação dos resultados de dois diferentes conjuntos de análise de grupos para determinar qual deles é melhor ou, ainda, determinação do número correto de grupos (TAN; STEINBACH; KUMAR, 2005).

Boryczka (2009) e Handl, Knowles e Dorigo (2006) utilizaram em seus trabalhos dois índices internos (a Variância Intra-Grupos e o Índice de *Dunn*) e dois índices externos (a medida *F* e o Índice Aleatório). Os índices externos, descritos em Handl, Knowles e Dorigo (2006) também foram utilizados neste trabalho.

3. Materiais e Métodos

As bases de dados utilizadas neste trabalho foram: Iris, Wine e Pima Indians Diabetes, disponíveis em http://mlearn.ics.uci.edu/databases. O quadro 1 mostra o número de padrões, o número de atributos e a quantidade de grupos para cada uma destas bases de dados. Os dados foram padronizados antes da aplicação dos métodos para agrupamento. A padronização foi feita por dimensão.

Base de Dados	N° de Padrões	N° de atributos	N° de grupos
Íris	150	4	3
Wine	178	13	3
Pima Indians Diabetes	768	8	2

QUADRO 1 - Bases de dados utilizados para avaliação do algoritmo.

O Método Ward (JOHNSON; WICHERN, 1998) foi aplicado às três bases de dados com o auxílio do *software* computacional *MatLab*2008. Nestas bases de dados é conhecido o número correto de grupos, os quais foram fornecidos para que os agrupamentos fossem avaliados. A medida de dissimilaridade utilizada foi a distância euclidiana por ser a mais conhecida entre as medidas de dissimilaridade e por ter sido empregada em trabalhos anteriores para todos os métodos aqui utilizados.

3.1. Agrupamento através dos Mapas de Kohonen Unidimensionais

O agrupamento por Mapas de Kohonen Unidimensionais (*SOM- Self Organizing Maps*) aplicado às bases de dados foi implementado no *software* computacional *MatLab*2008 e foi executado dez vezes para cada base de dados.

Neste método é necessária a definição do número de neurônios, neste trabalho foi definido que a quantidade de neurônios deve ser igual ao número de grupos (*k*). O número de grupos é conhecido para as bases de dados utilizadas.

O primeiro passo na execução do algoritmo dos Mapas de Kohonen Unidimensionais é a inicialização (FAUSETT, 1994; KOHONEN, 1995). Nesta implementação, foram definidos: a taxa de aprendizagem inicial igual a 0,5; a taxa de aprendizagem mínima igual a 0,05; o raio de vizinhança inicial igual ao valor máximo entre "1" e "¼ *k*"; os pesos sinápticos iniciais dos neurônios igual a valores aleatórios pertencentes ao intervalo [0, 1] e o número máximo de iterações $N = 500$.

O segundo passo na execução do algoritmo é a definição do critério de parada que, neste trabalho, foi definido como o número máximo de iterações. No algoritmo implementado, foram definidas duas fases (inicial e final) nas quais os ajustes de parâmetros são modificados. A fase inicial foi definida como $t_{inicial} = 0{,}2\ N$. Na fase final, o raio de vizinhança inicial é igual ao raio de vizinhança ao final da primeira fase.

O terceiro passo é o treinamento, que envolve as fases competitiva, cooperativa e adaptativa, onde cada padrão deve ser apresentado à rede. Nesta implementação, a ordem de entrada dos padrões foi definida para ser aleatória, ou seja, a cada iteração todos os padrões são apresentados à rede de forma aleatória.

Na fase competitiva, calculam-se as distâncias do padrão a todos os neurônios e verifica-se qual é o neurônio vencedor. Nesta implementação foi utilizada a distância euclidiana.

Na fase cooperativa, localizam-se os vizinhos do neurônio vencedor e na fase adaptativa, atualizam-se os pesos sinápticos dos neurônios vizinhos ao neurônio vencedor. A atualização dos pesos sinápticos foi feita segundo a equação (6) com função de vizinhança definida pela equação (7). Esta atualização leva em consideração a distância do vizinho até o neurônio vencedor e a taxa de aprendizagem.

$$\underline{w}_j(n+1) = \underline{w}_j(n) + \eta\,(n)\ .\ h_{j,i(x)}(n)\ .\ (\underline{x}-\underline{w}_j(n)) \qquad (6)$$

$$h_{j,i(x)}(n) = e^{\left(-\frac{d_{j,i}^2}{2\sigma^2(n)}\right)} \qquad (7)$$

No quarto passo, a taxa de aprendizagem e o raio de vizinhança devem ser atualizados, e isto foi feito segundo as equações (8) e (9), respectivamente.

$$\eta(n) = \eta_0 e^{-\frac{n}{\tau 2}} \qquad (8)$$

$$\sigma(n) = \sigma_0 e^{-\frac{n}{\tau 1}} \qquad (9)$$

Nestas equações, $\tau_1 = \frac{N}{\log(\sigma_0)}$ e $\tau_2 = N$, onde N é o número máximo de iterações e σ_0 é o raio de vizinhança inicial. Estes valores foram definidos baseando-se nos valores utilizados por Haykin (2001), $\tau_1 = \frac{1000}{\log(\sigma_0)}$ e $\tau_2 = 1000$.

3.2. Agrupamento através do Agrupamento baseado em Formigas

O Agrupamento baseado em Formigas proposto, baseado no algoritmo básico de Deneubourg *et al.*, apresentado na seção 2.1, foi implementado no *software* computacional *MatLab*2008. Nesse trabalho foram utilizados recursos da grade computacional do LCPAD: Laboratório Central de Processamento de Alto Desempenho/UFPR, parcialmente financiado pela FINEP projeto CT-INFRA/UFPR/Modelagem e Computação Científica.

Neste algoritmo básico de Deneubourg *et al.*, várias propostas quanto a implementação foram analisadas e são apresentadas a seguir, com o intuito melhorar o seu desempenho. Alguns procedimentos permaneceram os mesmos, os quais são igualmente

enfatizados. Além disso, três principais modificações foram propostas.

O algoritmo implementado utilizou como critério de parada o número de iterações e o algoritmo foi executado dez vezes. Sendo n é o número de padrões e m é o número de atributos, o número de iterações N_{max} foi definido como $N_{max} = 500.n.m$. No algoritmo implementado, foram definidas duas fases (inicial e final) na qual os ajustes de parâmetros são modificados. A fase inicial foi definida como $t_{inicial} = 0{,}2.N_{max}$.

Na definição do tamanho da grade, escolheu-se o número de células igual a 10 vezes o número de padrões e foram utilizadas 10 formigas (p=10), como em Handl, Knowles e Dorigo (2006). Observou-se que a alteração destes valores não é imprescindível no processo de agrupamento, por este motivo foram utilizados os mesmos valores. Foi utilizada vizinhança quadrada na busca dos padrões vizinhos.

Como em Handl, Knowles e Dorigo (2006), o raio de vizinhança inicial foi definido igual a "1", com a utilização de incremento deste valor durante a fase inicial. O aumento deste valor foi feito segundo a equação (10), onde t é a iteração atual da fase inicial. Durante a fase final, este valor decresce em 0,05 a cada 100 substituições do padrão carregado por uma formiga (modificação sugerida e que será descrita na seção 3.2). O valor do raio de vizinhança é sempre o valor inteiro menor ou igual ao definido em qualquer uma das fases.

$$\sigma = 4^{\frac{t}{t_{inicial}}} \qquad (10)$$

Na definição da vizinhança para o cálculo da probabilidade de descarregar um padrão em sua posição atual e para o cálculo da probabilidade de carregar um padrão, considerou-se o raio de vizinhança sempre igual a "1".

Na busca de uma nova posição, a direção do passo é aleatória. Definida a direção, calcula-se o tamanho máximo possível do passo. Um número aleatório pertencente ao intervalo [0, 1] foi utilizado para determinar este tamanho, multiplicando-se este número pelo tamanho máximo do passo.

As probabilidades de carregar (p_{pick}) e descarregar (p_{drop}) utilizadas são as descritas pelas equações (1) e (2) da seção 2, respectivamente, onde $k_p = 0{,}1$ e $k_d = 0{,}3$, como em Deneubourg *et al.*. Um padrão é carregado se a probabilidade p_p for maior que um valor mínimo para carregamento ($pick_{min}$). Um padrão é descarregado se a probabilidade p_d for maior que um valor mínimo para descarregamento ($drop_{min}$).

Os valores de $drop_{\min}$ e $pick_{\min}$ foram definidos como 0,13397 durante a fase inicial. Durante a fase final estes valores foram definidos aleatoriamente, com a restrição de serem maiores que 0,13397, a cada vez que todas as formigas executassem uma iteração. O valor 0,13397 foi definido fazendo a probabilidade de carregar (p_{pick}) igual a probabilidade de descarregar (p_{drop}).

No cálculo da função f, foi utilizada a função f* definida pela equação (5) da seção 2, substituindo-se o parâmetro escalar $\frac{1}{\sigma^2}$ por $\frac{1}{N_{occ}}$, onde N_{occ} é o número de células da grade ocupadas observadas dentro da vizinhança local.

O parâmetro α foi definido como 0,8. Este valor foi atualizado durante a fase inicial segundo a equação (11). Durante a fase final este valor decresce em 0,001 a cada 100 substituições do padrão carregado por uma formiga.

$$a = a_0 + \frac{2t}{p.t_{inicial}} - 0{,}01 \qquad (11)$$

Observa-se que qualquer alteração nos valores de k_p, k_d e α, influenciam diretamente o processo de agrupamento. Optou-se por manter os valores de k_p e k_d e utilizar somente uma adaptação para α. Se os valores de k_p e k_d forem alterados, a adaptação para α, bem como os valores $pick_{min}$ e $drop_{min}$ deverão ser revistos.

Quando um padrão é descarregado na grade, um novo padrão deverá ser carregado. A busca deste padrão é aleatória, porém, cada padrão livre é avaliado somente uma vez, até que todos sejam avaliados. Caso nenhum padrão apresente probabilidade p_{pick} maior que $pick_{\text{min}}$, o padrão que apresentar a maior probabilidade p_{pick} é carregado.

Quando um padrão não tem vizinhos, definiu-se a função f igual a zero. Isso faz com que a probabilidade p_{d} seja igual a "0", ou seja, o padrão não deve ser descarregado naquela posição e a probabilidade p_{p} foi igual a "1", ou seja, o dado deverá ser carregado e futuramente deixar esta posição.

A medida de dissimilaridade utilizada foi a distância euclidiana. A matriz de distâncias foi calculada segundo a equação (12) e depois foi padronizada. Nesta equação (12), o *peso* se refere ao atributo e é calculado dividindo-se o desvio-padrão pela média, calculado para cada atributo da matriz dos dados já padronizada (Q).

$$\tilde{d}(i,j) = \sum_{a=1}^{m} \left[(Q(a,i) - Q(a,j)) * peso(a,1) \right]^2 \qquad (12)$$

Na recuperação dos grupos foi utilizado o Método Ward e foi definido um número máximo de grupos. Em Villwock e Steiner (2008), outros métodos foram testados e o Método Ward apresentou melhores resultados.

Para avaliação dos resultados foram utilizados dois índices externos (Medida F e Índice Aleatório) e o percentual de classificação errada.

Além dos detalhes de implementação descritos e da inclusão de melhorias já propostas anteriormente, três principais modificações foram propostas neste trabalho e são descritas a seguir.

3.2.1. Modificações Propostas para o Agrupamento baseado em Formigas

Durante o estudo do Agrupamento baseado em Formigas, foi observado que muitas das mudanças de posição dos padrões ocorrem desnecessariamente. Considera-se uma mudança desnecessária quando um padrão está entre similares na grade e, neste caso, não há necessidade da mudança deste padrão para outra posição. Com o objetivo de evitar estas mudanças desnecessárias, uma comparação da probabilidade de descarregar um padrão na posição escolhida aleatoriamente com a probabilidade de descarregar este padrão em sua posição atual foi introduzida. A decisão de descarregar um padrão na posição escolhida aleatoriamente só ocorre se esta probabilidade for maior que a probabilidade de descarregar este padrão em sua posição atual.

Também foi observada a ocorrência de fusão de grupos próximos na grade. Quando a decisão de descarregar um padrão for positiva e a célula em que o padrão deveria ser descarregado está ocupada, busca-se aleatoriamente uma posição vizinha a esta, que esteja livre. Porém, esta nova posição pode estar próxima também a outro grupo de padrões na grade. Este pode ser um motivo para a fusão de grupos próximos. Como uma alternativa para

evitar a fusão de grupos próximos na grade, foi proposta neste trabalho uma avaliação da probabilidade para a nova posição. O padrão só é descarregado na célula vizinha se a probabilidade de descarregar o padrão nesta posição for maior que a probabilidade de descarregar este padrão em sua posição atual. Todas as posições vizinhas livres são avaliadas. Se em nenhuma posição vizinha livre a probabilidade de descarregar o padrão for maior que a probabilidade de descarregar este padrão em sua posição atual, o padrão não é descarregado e o processo se reinicia escolhendo-se outra formiga.

Outra questão observada no Agrupamento baseado em Formigas é que uma formiga pode carregar um padrão que está entre similares na grade. Uma formiga só carrega um padrão quando este não está entre similares na grade, porém, desde que a formiga carregue um padrão até ela ser sorteada para tentar descarregar o padrão, mudanças ocorrem na vizinhança deste, podendo deixá-lo então entre similares. Sendo assim, esta formiga fica inativa, pois a operação de descarregar o padrão não é executada. Neste caso, foi proposta a substituição do padrão carregado por uma formiga, caso este padrão não seja descarregado em 100 iterações consecutivas. O novo padrão é escolhido por sorteio, mas ele só é carregado pela formiga se a probabilidade de carregar este padrão for maior que 0,13397. Caso não exista nenhum padrão com probabilidade de carregar maior que 0,13397, o último padrão sorteado é carregado pela formiga. Este também poderia ser um critério de parada.

4. Resultados e Discussões

O algoritmo de Agrupamento baseado em Formigas proposto foi aplicado às bases de dados apresentadas no início da seção 3 (conhecido o grupo a que cada padrão pertence). Por não se tratar de um método exato, ou seja, há variação nos resultados se aplicado por diversas vezes, este método foi aplicado a cada base de dados por 10 vezes.

Para avaliação dos resultados foram utilizadas as seguintes medidas de avaliação do agrupamento: Similaridade (*Sim*), Índice Aleatório *R* (quanto maior, melhor), Medida *F* (quanto maior, melhor) e percentual de classificação errada. Resultados preliminares para as bases de dados Íris e Wine foram publicados em Villwock e Steiner (2009a; 2009b).

O quadro 2, a seguir, apresenta a média e o desvio-padrão das medidas de avaliação para as bases de dados. Este quadro também apresenta as medidas de avaliação do agrupamento para o melhor resultado.

Resultados		R	F	Classificação errada (%)
Íris	**Média**	0,871	0,877	11,9
	Desvio-padrão	0,039	0,050	4,6
	Melhor resultado	0,927	0,940	6,0
Wine	**Média**	0,843	0,871	12,7
	Desvio-padrão	0,019	0,021	1,9
	Melhor resultado	0,871	0,899	10,1
Pima	**Média**	0,510	0,583	43,6
	Desvio-padrão	0,010	0,022	4,0
	Melhor resultado	0,531	0,623	37,5

QUADRO 2 - Resultados da aplicação do algoritmo de Agrupamento baseado em Formigas proposto, médias da execução de 10 vezes, para as bases de dados (IRIS, WINE e PIMA).

A figura 1 apresenta a grade para o melhor resultado (cujas medidas de avaliação foram apresentadas no quadro 2) para a base de dado Íris. Nesta base de dados os padrões em

vermelho pertencem ao grupo 1, os padrões em preto pertencem ao grupo 2 e os padrões em azul pertencem ao grupo 3.

O quadro 3 (matriz de confusão) mostra a distribuição dos padrões para a base de dados Íris, onde pode-se observar os padrões atribuídos aos grupos corretamente e os padrões atribuídos aos grupos erroneamente. Nesta base de dados são apenas nove padrões em grupos errados de um total de 150 padrões. O grupo 1 contém todos os padrões atribuídos a ele.

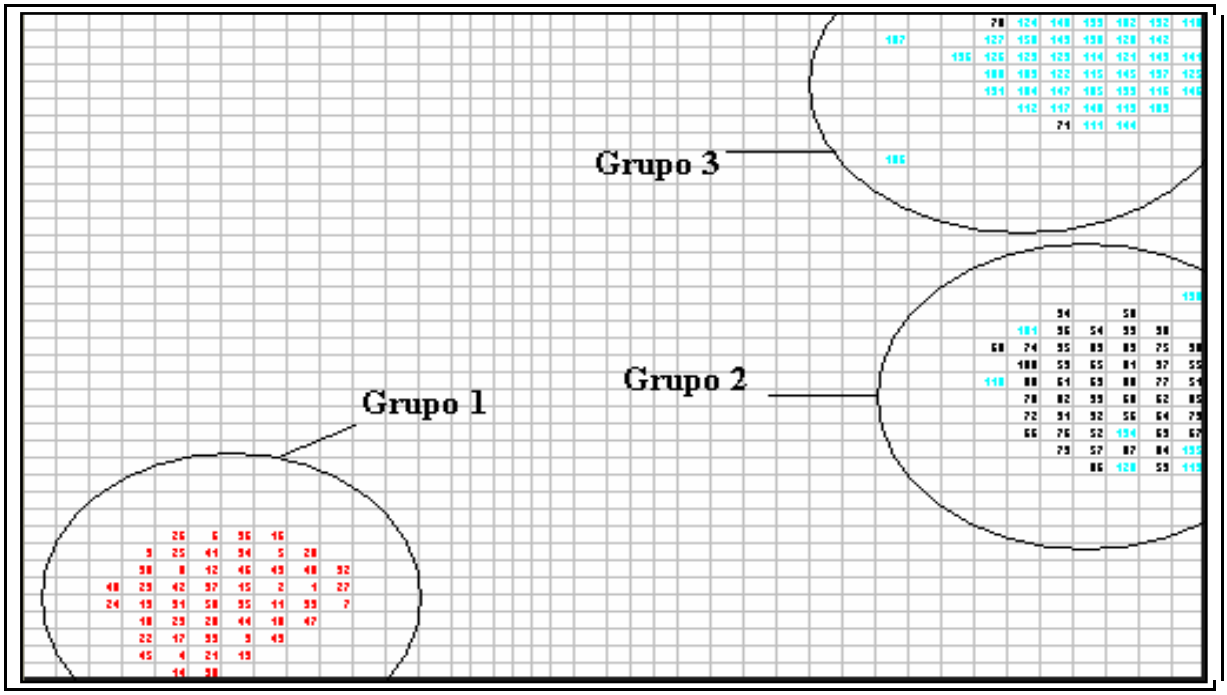

FIGURA 1 – Resultado do algoritmo de agrupamento baseado em Formigas proposto para a base de dados Iris – melhor resultado.

IRIS	Solução Gerada		
Agrupamento Correto	**Grupo 1**	**Grupo 2**	**Grupo 3**
Classe 1	50	0	0
Classe 2	0	48	2
Classe 3	0	7	43

QUADRO 3 - Distribuição dos Padrões – IRIS – melhor resultado.

No quadro 4 são apresentadas as comparações das medidas médias de avaliação para os três métodos, para as bases de dados Íris, Wine e Pima. O melhor resultado encontra-se em negrito.

Os resultados não mostram superioridade de algum método. Na base de dados Íris, o Método Ward foi melhor (cerca de 3% de erros); na base de dados Wine, o Algoritmo baseado em Formigas proposto foi melhor (cerca de 12% de erros) e na base de dados Pima, os Mapas de Kohonen Unidimensionais foram melhores (cerca de 34% de erros). Handl, Knowles e Dorigo (2006) também afirmam que nenhum algoritmo domina os outros sempre.

O quadro 5 apresenta a comparação das médias das medidas de avaliação do agrupamento para o algoritmo proposto e para o algoritmo *ACAM* (*Ant-based Clustering Algorithm Modified*) proposto por Boryczka (2009). O melhor resultado encontra-se em

negrito. Os resultados mostram que o algoritmo proposto é melhor que o *ACAM* para duas das três bases de dados.

Base de Dados	Medidas de Avaliação	Ward	1D-SOM	Formigas
Íris	**R** (quanto maior melhor)	**0,957**	0,863	0,871
	F (quanto maior melhor)	**0,967**	0,865	0,877
	Classificação errada (%) (quanto menor melhor)	**3,333**	12,8	11,9
Wine	**R** (quanto maior melhor)	0,819	0,764	**0,843**
	F (quanto maior melhor)	0,845	0,761	**0,871**
	Classificação errada (%) (quanto menor melhor)	15,169	22,416	**12,7**
Pima	**R** (quanto maior melhor)	0,531	**0,549**	0,510
	F (quanto maior melhor)	0,624	**0,655**	0,583
	Classificação errada (%) (quanto menor melhor)	37,370	**34,570**	43,6

QUADRO 4 - Comparação dos resultados médios da aplicação dos métodos de agrupamento Ward, Mapas de Kohonen Unidimensionais e Agrupamento baseado em Formigas proposto para as bases de dados Íris, Wine e Pima.

BASES	Medidas de Avaliação	*ACAM*	Algoritmo Proposto
Íris	**R**	0,819	**0,871**
	F	0,810	**0,877**
	Classificação errada (%)	18,7	**11,9**
Wine	**R**	**0,849**	0,843
	F	0,868	**0,871**
	Classificação errada (%)	13,9	**12,7**
Pima	**R**	**0,522**	0,510
	F	0,574	**0,583**
	Classificação errada (%)	**33,7**	43,6

QUADRO 5 - Comparação dos resultados médios da aplicação do algoritmo proposto com resultados disponíveis em Boryczka (2009) para as bases de dados.

5. Considerações Finais

O algoritmo de Agrupamento proposto baseado em Formigas foi aplicado a três bases de dados e, para avaliação do seu desempenho, foi comparado com o Método Ward e aos Mapas de Kohonen Unidimensionais.

Quando comparadas as médias das medidas de avaliação (quadro 4), na aplicação dos métodos de agrupamento Ward, Mapas de Kohonen Unidimensionais e Agrupamento proposto baseado em Formigas, para as bases de dados, os resultados não mostraram superioridade de algum dos métodos. Handl, Knowles e Dorigo (2006) também afirmam que nenhum algoritmo domina os outros sempre.

Já na comparação das médias das medidas de avaliação (quadro 5) do agrupamento para o algoritmo proposto e para o algoritmo *ACAM* (*Ant-based Clustering Algorithm Modified*), proposto por Boryczka (2009), os resultados mostram que o algoritmo proposto apresentou um desempenho melhor do que o *ACAM* para duas das três bases de dados.

6. Agradecimentos

À FINEP, pelo apoio financeiro ao projeto de pesquisa CT – INFRA / UFPR / Modelagem e Computação Científica.

Referências

BORYCZKA, U. Finding groups in data: Cluster analysis with ants. **Applied Soft Computing**, v. 9, p. 61-70, 2009.

DORIGO, M.; STÜTZLE, T. **Ant colony optimization**. Cambridge: MIT Press, 2004.

FAUSETT, L. **Fundamentals of Neural Networks – Architectures, Algorithms, and Applications**. New Jersey: Prentice Hall, 1994.

HANDL, J.; KNOWLES, J.; DORIGO, M. Ant-Based Clustering and Topographic Mapping. **Artificial Life**, v. 12, n. 1, p. 35-61, 2006.

HAYKIN, S. **Redes neurais: princípios e prática.** Tradução: Paulo Martins Engel. Porto Alegre: Bookman, 2001.

JOHNSON, R.A.; WICHERN, D.W. **Applied Multivariate Statistical Analysis.** Fourth Edition. New Jersey: Prentice Hall, 1998.

KOHONEN, T. **Self-Organizing Map**. Berlin: Springer-Verlag, 1995.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Inc. Boston, MA, USA: Addison-Wesley Longman Publishing Co., 2005.

VILLWOCK, R.; STEINER, M. T. A. Agrupamento baseado em Colônia de Formigas: Estudo Comparativo de Algoritmos para Recuperação dos Grupos. In: XII Encontro Regional de Matemática Aplicada e Computacional, Foz do Iguaçú, 2008. **XII Encontro Regional de Matemática Aplicada e Computacional**. Foz do Iguaçú: 2008. CD-ROM.

VILLWOCK, R.; STEINER, M. T. A. Análise do Desempenho do Algoritmo de Agrupamento Baseado em Colônia de Formigas Modificado. In: XXXII Congresso Nacional de Matemática Aplicada e Computacional, Cuiabá, 2009. **XXXII Congresso Nacional de Matemática Aplicada e Computacional**. Cuiabá: SBMAC, 2009a, CD-ROM.

VILLWOCK, R.; STEINER, M. T. A. Análise do Desempenho de um Algoritmo de Agrupamento Modificado Baseado em Colônia de Formigas. In: XLI Simpósio Brasileiro de Pesquisa Operacional, Porto Seguro, 2009. **XLI Simpósio Brasileiro de Pesquisa Operacional**. Porto Seguro: SOBRAPO, 2009b, CD-ROM.

VIZINE, A. L.; DE CASTRO, L. N.; HRUSCHKA, E. R.; GUDWIN, R. R. Towards improving clustering ants: an adaptive ant clustering algorithm. **Informatica**, v. *29*, p. 143–154, 2005.