Clusterização de Ocorrências Policiais utilizando k-means e um Mapa Auto-Organizável

Fabiano R. de Oliveira e Maria B. Zanusso

***Resumo*—A tarefa de clusterização em mineração de dados tem como função agrupar um conjunto selecionado de dados dentro de uma coleção de agrupamentos, assegurando a similaridade intraclasse e a heterogeneidade interclasse.**

O presente artigo apresenta os resultados obtidos na clusterização de ocorrências policiais utilizando um algoritmo clássico de clusterização, o k-means, e um mapa auto-organizável de Kohonen.

***Index Terms*—clusterização, mineração de dados, k-means, mapa auto-organizável.**

I. INTRODUÇÃO

AS duas últimas décadas acompanharam um aumento dramático na quantidade de informações que são armazenadas em formato eletrônico. Este acúmulo de dados aconteceu a uma taxa explosiva. Calcula-se que a quantidade de informação no mundo dobra a cada vinte meses e que o tamanho e número de bancos de dados estão aumentando ainda mais rapidamente.

O valor destes dados está diretamente ligado à capacidade de extrair informações de mais alto nível que neles se encontram subjacentes, i. é., "escondidas". Há a possibilidade de existir tendências ou padrões úteis e interessantes que, se descobertos, podem ser utilizados, por exemplo, para otimizar um processo de negócio em uma empresa, auxiliar no entendimento dos resultados de um experimento científico, ajudar médicos a entender efeitos de um tratamento, entre outros [13]. A mineração de dados consiste em um processo com várias fases e com várias tarefas, visando à automação da extração de conhecimentos úteis, ditos "interessantes" e que se encontram tipicamente em grandes massas de dados.

Nos últimos anos, o foco de interesse em mineração de dados tem se voltado para a tarefa de clusterização. Esta tarefa agrupa um conjunto selecionado de dados dentro de uma coleção de agrupamentos, assegurando a similaridade intraclasse e a heterogeneidade interclasse. Logo, se constitui como um processo de particionar um conjunto de dados em um conjunto de classes, chamadas *clusters* (agrupamentos), com os membros de cada cluster compartilhando algumas propriedades comuns [5].

Algoritmos de clusterização foram amplamente estudados, aplicados e comparados [2] [4] [11] [12]. Em [11], é utilizada Análise de Componentes Principais (ACP) para determinar a distribuição inicial dos dados sobre os clusters com a finalidade de aumentar o desempenho da tarefa de clusterização. Em [9] e [12], dois métodos de clusterização são associados a fim de explorar as melhores características de ambos. Em [10] um método interativo de clusterização é utilizado sobre dados genômicos. [4] mostra uma comparação entre métodos de clusterização.

No presente artigo, são apresentados os resultados obtidos na clusterização de ocorrências policiais, utilizando dois dos principais algoritmos de clusterização, o k-means e um Mapa Auto-Organizável (*Self-Organizing Map*, SOM) de Kohonen. Espera-se que os resultados obtidos possam ser utilizados para o auxílio na elaboração de políticas de segurança específicas no combate à criminalidade.

A Seção II discute o algoritmo k-means. A Seção III, expõe as diretrizes do SOM de Kohonen. O conjunto de dados de ocorrências policiais é brevemente apresentado na Seção IV. As Seções V e VI apresentam os resultados da clusterização de ocorrências policiais através do uso do k-means e de um SOM de Kohonen, respectivamente. Finalmente, a Seção VII apresenta algumas conclusões.

II. O ALGORITMO K-MEANS

Descrito em detalhes por Hartigan [1], o k-means é um dos algoritmos de clusterização mais importantes na literatura. Seus resultados estão bem próximos daqueles idealizados pelo algoritmo classificador de máxima verossimilhança [6].

Inicialmente, os vetores de características (e. g.: tuplas de uma relação) estão em um arquivo de dados e não se sabe qual a distribuição deles nos clusters existentes. Logo, estes vetores de características são distribuídos aleatoriamente entre os clusters. O algoritmo calcula, basicamente, os vetores médios para cada cluster e compara a distância euclidiana de cada vetor de características (amostras) com os vetores médios de cada cluster. Se o cluster que possuir o vetor médio mais próximo do vetor de características atual for diferente do cluster no qual este está inserido, então este vetor de características é excluído de seu cluster atual e inserido no cluster cujo vetor médio esteja mais próximo dele. Todo esse processo é executado dentro de um laço iterativo, cuja condição de parada é que a mudança nos vetores médios, de uma iteração para a outra imediatamente posterior, seja menor ou igual ao limite de similaridade previamente estabelecido, geralmente muito próximo ou igual a zero.

De acordo com [11], o desempenho deste método de partição depende diretamente da distribuição inicial destes vetores de características sobre os clusters.

Este algoritmo é um dos métodos de partição clássicos no campo da estatística. Os métodos de partição, em geral, procuram uma partição de p objetos em c clusters, de modo

F. R. de Oliveira - Departamento de Computação e Estatística, Universidade Federal de Mato Grosso do Sul Campo Grande, MS. Telefone: +55 67 385–9181, e-mail: fro@dct.ufms.br

M. B. Zanusso - Departamento de Computação e Estatística, Universidade Federal de Mato Grosso do Sul, Campo Grande, MS. Telefone: +55 67 345–7505, e-mail: mzanusso@dct.ufms.br

que satisfaçam a duas premissas básicas: coesão interna e isolamento dos grupos.

III. O SOM DE KOHONEN

O SOM de Kohonen [2] consiste de uma rede neural que combina uma camada de entrada com uma camada competitiva de unidades processadoras (camada oculta) capaz de encontrar a organização dos relacionamentos existentes entre padrões de entrada (vetores de características).

Esta rede utiliza o paradigma de aprendizado competitivo e implementa a inibição lateral existente nas redes de neurônios biológicos, i. é., os neurônios com maior nível de ativação tendem a inibir os níveis de ativação de seus vizinhos. Esta inibição é modelada sobre o esquema "o vencedor leva tudo" em que a unidade da camada competitiva com o maior valor da soma ponderada é designada como a vencedora. A esta unidade é então dado um novo valor de ativação, por exemplo 1, e a todas as outras unidades é dado um outro valor, por exemplo 0 [3].

Os padrões apresentados à rede são classificados correspondentemente pelas unidades que eles ativam na camada competitiva. O aprendizado competitivo realiza aprendizado não-supervisionado. No aprendizado não-supervisionado é apresentado um conjunto de padrões de treino à rede, mas não é dada nenhuma resposta desejada para cada padrão de entrada; por si mesma, a rede organiza os padrões de treino em um conjunto de classes. Isto estende a capacidade das redes neurais a aplicações no campo de reconhecimento de padrões, em que a classificação desejada não é conhecida, *a priori*, mas os dados podem, contudo, ser organizados em diferentes categorias [8]. Podemos sumarizar o algoritmo de Kohonen em alguns passos:

1. Inicialização: escolher valores aleatórios para os vetores de pesos iniciais. A única restrição aqui é que os valores dos pesos iniciais sejam diferentes para todos os neurônios da camada competitiva. É desejável manter a magnitude dos pesos pequena;
2. Amostragem: retirar uma amostra da distribuição das possíveis entradas (um exemplo de uma época) com uma certa probabilidade; o vetor de características assim obtido representa o sinal sensório ou o estímulo de entrada;.
3. Casamento (*Matching*) das similaridades: encontrar o neurônio que melhor se *case* (o neurônio vencedor) com a amostra apresentada no passo anterior, usando o critério da distância euclidiana;
4. Atualização: ajustar o vetor de pesos de todos os neurônios, usando a Equação 1:

$$w_j(n+1) = w_j(n) + etha[x(n) - w_j(n)] \qquad (1)$$

   Onde *etha* é o parâmetro taxa de aprendizagem, e j pertence à vizinhaça centralizada em torno do neurônio vencedor. Ambos, a taxa de aprendizado e a vizinhança são variados dinamicamente durante o aprendizado, com o objetivo de se obter melhores resultados. A vizinhança é decrementada linearmente e a taxa de aprendizado tem seu valor diminuído de uma porcentagem. Essas atualizações acontecem sucessivamente a cada período. O período é definido como sendo um número predeterminado de épocas;
5. Continuação: continuar com o passo 2 até que nenhuma mudança notável (ou significativa) no mapa de características seja observada.

IV. O CONJUNTO DE DADOS

O conjunto de dados a ser submetido à tarefa de clusterização é um subconjunto dos dados constantes das ocorrências policiais de lesão corporal em Campo Grande/MS registradas entre 01/01/2001 a 31/07/2001. Foram utilizados apenas os campos SEXO e IDADE das vítimas de cada ocorrência de lesão corporal registrada. Estes dados foram previamente pré-processados e selecionados para que pudessem ser utilizados pelos algoritmos de clusterização. A Tabela I mostra alguns registros deste conjunto.

TABLE I
AMOSTRA DO CONJUNTO DE DADOS DE OCORRÊNCIAS POLICIAIS

SEXO	IDADE
FEMININO	20
MASCULINO	57
FEMININO	35
...	...
FEMININO	21
MASCULINO	85

V. CLUSTERIZAÇÃO USANDO K-MEANS

Na Figura 1, é apresentado um gráfico correspondente à clusterização do conjunto de dados das ocorrências policiais produzida pelo algoritmo k-means. Em cada cluster mostrado na legenda da figura, o valor entre colchetes denota o intervalo de idades dentro do cluster e o valor entre parênteses, o número de idades distintas (amplitude) dentro do cluster. A saída é interpretada como: "De todas as vítimas de ocorrências de lesão corporal em Campo Grande/MS, registradas entre 01/01/2001 a 31/07/2001, $20,40\%$ são mulheres com idades entre 23 e 30 anos". Isto significa que em mais de um quinto de todas estas ocorrências, figuram como vítimas mulheres com apenas oito idades (23, 24, 25, 26, 27, 28, 29 e 30)! De fato, este padrão pareceu confirmar o que o setor social da Polícia já sabia: a maioria das vítimas de lesão corporal era formada por mulheres jovens que haviam se casado há pouco tempo.

VI. CLUSTERIZAÇÃO USANDO SOM DE KOHONEN

A arquitetura do SOM utilizada para a aplicação sobre a base de dados de ocorrências policiais é constituída de uma camada de entrada com duas unidades (para os atributos SEXO e IDADE) e uma camada competitiva, composta de dez neurônios (até dez clusters podem ser obtidos) e sem bias. A rede foi inicializada com os seguintes parâmetros: $etha = 0.7$, $vizinhanca =$

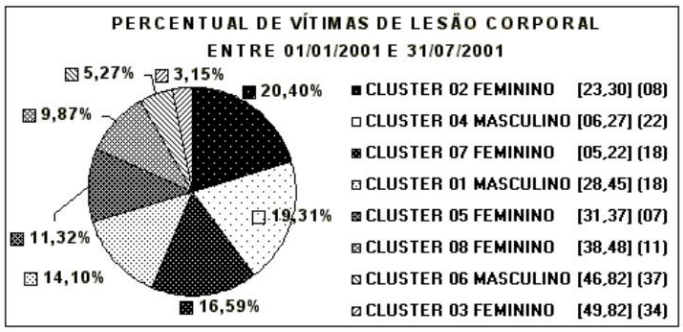

Figura 1. Resultado da clusterização k-means.

1, $periodo = 500$, $numeromaximodeepocas = 500$. Os parâmetros *vizinhanca* e *etha* são sucessivamente decrementados de 1 e 10%, respectivamente.

Na Figura 2, são mostrados os resultados produzidos pela rede, após a clusterização. A saída é interpretada como: "De todas as vítimas de ocorrências de lesão corporal em Campo Grande/MS, registradas entre 01/01/2001 a 31/07/2001, $38,68\%$ são homens com idades entre 6 e 82 anos". Este primeiro cluster (01) não parece ser bom, já que apresenta uma amplitude de 77 idades. No entanto, observando os clusters 08 e 07, nota-se algo mais interessante, já que, desta forma, tem-se um cluster com pessoas do sexo feminino respondendo por $25,79\%$ $(13,14\% + 12,65\%)$ e com apenas 13 $(6+7)$ idades de amplitude.

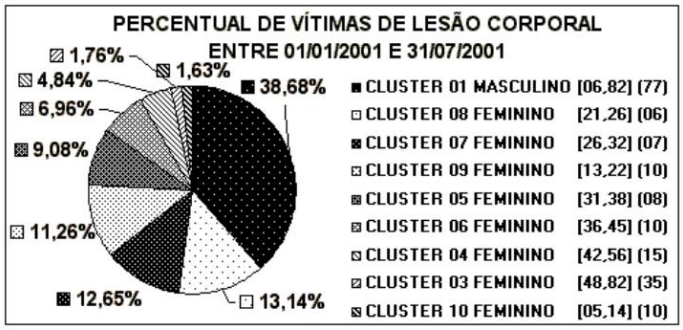

Figura 2. Resultado da clusterização por SOM de Kohonen.

Nota-se que este método de clusterização apresentou, neste caso de teste, uma pequena sobreposição nos clusters sobre o atributo faixa etária ($[21,26]$ e $[26,32]$). Entretanto, esta sobreposição não compromete os resultados de uma maneira significativa, já que, como observado no parágrafo anterior, pode-se considerar os dois clusters como um só ($[21,32]$).

VII. CONCLUSÕES

O principal foco de pesquisa nas tarefas de mineração de dados, e em particular na tarefa de clusterização, não está na eficiência, mas na eficácia. Busca-se, em um primeiro momento, fazer, para depois, em um segundo momento, fazer melhor.

A aplicação dos algoritmos de clusterização sobre o conjunto de ocorrências policiais de lesão corporal acabou por identificar o perfil dominante das vítimas deste delito. Estes resultados poderão, efetivamente, auxiliar na elaboração de programas sociais direcionados de combate à violência doméstica.

Uma comparação dos resultados obtidos pelos dois métodos não seria conveniente neste contexto. A comparação, neste caso, restringe-se ao comportamento dos algoritmos: o k-means é um algoritmo de execução mais rápida e produziu agrupamentos com limites mais definidos; enquanto o SOM de Kohonen mostrou-se um pouco mais lento, porém flexível e adaptável a um possível contexto de reconhecimento de padrões.

Há de se ressaltar que, apesar de já terem sido amplamente implementadas, as técnicas de clusterização são passíveis de melhorias [10] [11]. Hibridizações podem figurar como alternativas interessantes [9] e [12].

AGRADECIMENTOS

F.R.O. agradece à Polícia Civil do Estado de Mato Grosso do Sul pela disponibilização dos dados referentes às ocorrências policiais registradas em Campo Grande/MS.

Referências

[1] HARTIGAN, J. A., *Clustering Algorithms.*, New York, Wiley, 1975.

[2] KOHONEN, T., *Self-Organization and Associative Memory - Managing.*, H. K. V. Lostsch, Verlog Berlin Heidelberg, third edition, 1989.

[3] DAYHOFF, J. *Neural Network Architectures - An Introduction.*, Von Nostrand Reinhold, 1990.

[4] ULTSCH, A. *Self Organizing Neural Networks Perform Different from Statistical k-means Clustering.*, In: p.p. 433-443. M. van der Meer, R. Schmidt, G. Wolf, (Eds.):BMBF Statusseminar "Künstliche Intelligenz, Neuroinformatik und IntelligenteSysteme, 1996.

[5] HAN, J. et al., *DBMiner: A System for Data Mining in Relational Databases and Data Warehouses.*, In Proc. CASCON'97: Meeting of Minds, pp. 249-260, Toronto, Canadá, Novembro 1997.

[6] RAUBER, T. W., *Mini-curso: Reconhecimento de padrões.*, In: p.p. 311-353. JAI'97, 1997.

[7] CAZARINE, A.; BARBOSA, C.; VANCIN, F *et. al. Interactive clustering for exploration of genomic data.*, In: p.p. 753-758. Intelligent engineering systems through artificial neural networks, vol 12, eds. Dagli et al., New York, NY: ASME Press, 2002.

[8] HAYKIN, S. *Neural Networks: a Comprehensive Foundation.*, 2nd Ed. Upper Saddle River, NJ: Prentice Hall, 1999.

[9] LAERHOVEN, K. V. *Combining the Self-Organizing Map and K-Means Clustering for On-Line Classification of Sensor Data.*, In: p.p. 464-469. ICANN, 2001.

[10] WAN, X.; BRIDGES, S. M.; BOYLE, J. A.;BOYLE, A. P. *Interactive clustering for exploration of genomic data.*, In: p.p. 753-758. Intelligent engineering systems through artificial neural networks, vol 12, eds. Dagli et al., New York, NY: ASME Press, 2002.

[11] SU, T.;DY, J., *A Deterministic Method for Initializing K-Means Clustering.*, In: p.p. 784-786. ICTAI'2004, 2004.

[12] DUTRA, R. M.; SPERANDIO, M.; COELHO, J. *O Método Ward de Agrupamento de Dados e sua Aplicação em Associação com os Mapas Auto-Organizáveis de Kohonen.*, In: I WORKCOMP SUL - WORKSHOP DE CIÊNCIAS DA COMPUTAÇÃO E SISTEMAS DA INFORMAÇÃO DA REGIÃO SUL, 2004, Florianópolis. Anais do I WorkComp Sul. Florianópolis: UNISUL, 2004.

[13] KUMAR, P.; BAJCSY, P.; TCHENG, D.; CLUTTER, D.; MEHRA, V.; FENG, W-W; SINHA, P.; WHITE, A. *Using D2K Data Mining Platform for Understanding the Dynamic Evolution of Land-Surface Variables.*, In: The 2005 Earth-Sun System Technology Conference, University of Maryland, MD, June 28-30, 2005.