Usando Semi-supervisão para definir Representantes Auxiliares em Processos de Agrupamentos de Dados

W. J. Silva, M. C. N. Barioni, S. de Amo, H. L. Razente

Universidade Federal de Uberlândia, Brasil
waltersilva@mestrado.ufu.br, {camila.barioni, deamo, humberto.razente}@ufu.br

Resumo. A incorporação de semi-supervisão em processos de detecção de agrupamentos de dados tem se revelado especialmente útil quando se deseja obter uma alta consistência entre o particionamento dos dados e o conhecimento que se tem a respeito do domínio dos dados. Nos últimos anos, várias estratégias de detecção semi-supervisionada de agrupamentos foram propostas. As abordagens adotadas por essas estratégias buscam orientar o processo de detecção de agrupamentos utilizando restrições para: interferir na atribuição das instâncias aos grupos mais adequados a cada iteração do algoritmo; ou modificar a função objetivo usada pelo mesmo. Este artigo propõe uma nova abordagem para empregar as informações de semi-supervisão na definição de múltiplos representantes auxiliares para cada centróide definido pelo *k-means*. Os resultados experimentais iniciais com quatro conjuntos de dados sintéticos mostram o potencial da abordagem proposta para lidar com estruturas de dados mais complexas encontrando agrupamentos com formatos não necessariamente esféricos.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.5.3 [**Pattern Recognition**]: Clustering

Palavras-chave: análise de agrupamentos, aprendizado semi-supervisionado, mineração de dados

1. INTRODUÇÃO

A análise de agrupamentos em conjuntos de dados é uma tarefa de mineração de dados que emprega algoritmos e técnicas para agrupar instâncias de dados de acordo com a similaridade de suas características [Jain 2010]. Tradicionalmente, as técnicas de agrupamento de dados não utilizam rótulos categóricos em seus processamentos, ou seja, adotam uma abordagem puramente não-supervisionada. Nos últimos anos, a adoção de abordagens semi-supervisionadas têm sido uma tendencia na área de análise de agrupamentos [Barioni et al. 2014].

A abordagem de detecção semi-supervisionada de agrupamentos conta com alguma pequena quantidade de informação adicional sobre os dados que auxilia no processo de atribuição das instâncias aos grupos. Essa informação adicional pode ser representada na forma de rótulos para uma pequena quantidade de dados do conjunto, ou na forma de restrições entre pares de instâncias, como é o caso de restrições *must-link* e *cannot-link*, informando ao processo de agrupamento se duas instâncias devem estar juntas ou não podem estar juntas no mesmo grupo, respectivamente. Dentre os algoritmos mais relevantes que adotam essa abordagem está o algoritmo COP-*kmeans* [Wagstaff et al. 2001].

As estratégias de detecção semi-supervisionada de agrupamentos descritas na literatura da área utilizam restrições para orientar o processo de agrupamento de dados considerando duas abordagens principais: interferindo na atribuição de instâncias aos grupos mais adequados a cada iteraçao do algoritmo; ou modificando a função objetivo usada pelo mesmo [Basu et al. 2008]. Este artigo apresenta

Trabalho realizado com apoio financeiro da FAPEMIG (Processo 01290/12), CNPq (Universal 479930/2011-2 e 312224/2013-3), CAPES e PROPP/UFU.

uma nova abordagem que propõe utilizar as restrições para gerar múltiplos representantes auxiliares para cada partição gerada por algoritmos baseados em particionamento. Essa abordagem foi incorporada ao algoritmo *k-means* dando origem ao método denominado MRS-*kmeans* (*Multi-Representative Semi-supervised k-means*).

O uso de múltiplos representantes auxiliares no processo de agrupamento semi-supervisionado tem o intuito de utilizar as informações adicionais para permitir a detecção de formas mais complexas de agrupamentos de dados. Embora o MRS-*kmeans* ainda esteja em desenvolvimento, os resultados experimentais iniciais mostram que ele supera algoritmos de agrupamento semi-supervisionado descritos na literatura da área quando aplicado em conjuntos de dados sintéticos com grupos de diferentes tamanhos e formas. O restante do artigo está organizado como descrito a seguir. A Seção 2 descreve o método MRS-*kmeans*. A discussão a respeito dos resultados experimentais obtidos é apresentada na Seção 3. E, as conclusões e os trabalhos futuros são descritos na Seção 4.

2. MÉTODO DE AGRUPAMENTO SEMI-SUPERVISIONADO

A partir de um conjunto de dados $X = \{x_1, ..., x_n\}$, de um conjunto de restrições *must-link* R_{ml} e de um conjunto de restrições *cannot-link* R_{cl} o MRS-*kmeans* retorna uma partição dos dados em X que satisfaz todas as restrições informadas. Além de utilizar as restrições para guiar a etapa de atribuição de instâncias aos grupos, o MRS-*kmeans* também as utiliza para definir múltiplos representantes auxiliares para cada centróide do *k-means*.

Assim como no *k-means*, os centróides principais ($c_p^i \in C_p$) são definidos pelo valor médio das instâncias que compõem o grupo. Além dos centróides principais, há um outro tipo de representante para os grupos que é chamado de centróide auxiliar ($c_a^i \in C_a$). Os centróides auxiliares são computados pela média das instâncias conectadas por uma ou mais restrições *must-link*. Por exemplo, se $r_{ml}(x_i, x_j)$ e $r_{ml}(x_j, x_k)$, o centróide auxiliar que representa essas instâncias conectadas é dado por: $(x_i + x_j + x_k)/3$. A quantidade de instâncias conectadas por restrições *must-link* é denominada como a "população" de um determinado c_a^i. Essa população influencia a atribuição de peso w_a dada a um $c_a^i \in C_a$, significando que centróides auxiliares que representam mais instâncias devem ter maior influência na representação de um grupo.

Com base nas restrições *cannot-link* informadas previamente e nos centróides auxiliares já definidos, cria-se um conjunto de restrições *cannot-link-aux* R_{cla} entre centróides em C_a. Uma restrição entre um par de centróides em C_a é induzida da seguinte forma: seja uma restrição *cannot-link* $r_{cl}(x_i, x_j)$, observando que um centróide c_a^m está próximo de x_i e um outro centróide c_a^l está próximo de x_j, supõe-se que estes dois centróides não devem representar um mesmo grupo, logo cria-se uma restrição *cannot-link-aux* entre c_a^m e c_a^l.

Na etapa de atribuição de instâncias aos grupos, a avaliação de cada instância de X é realizada por meio da utilização de uma função de distância agregada conforme Equação 1. Nessa equação, Q é o conjunto de representantes de um grupo (principais e auxiliares), $d()$ é uma função de distância e w_j é um peso correspondente ao representante q_j.

$$d_g(Q, x_i) = min\left(d(q_j, x_i) \cdot 1/w_j\right), \quad \forall q_j \in Q \tag{1}$$

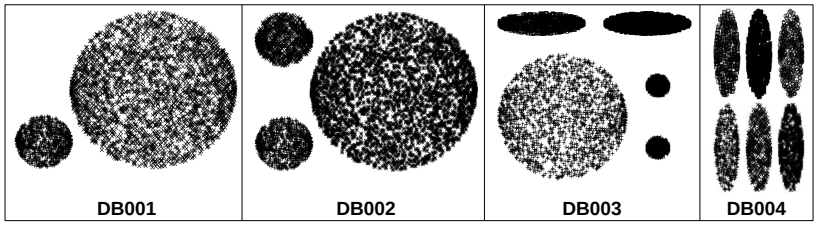
Resumidamente, os passos realizados pelo MRS-*kmeans* são apresentados abaixo:

(1) Selecionar aleatoriamente k instâncias como centróides $C_p = \left\{c_p^1, ..., c_p^k\right\}$ iniciais;
(2) Associar cada centróide auxiliar $c_a^i \in C_a$ ao grupo cujo centróide $c_p^j \in C_p$ está mais próximo, desde que não viole nenhuma restrição R_{cla};
(3) Atribuir cada instância $x_i \in X$ ao grupo $\pi_j \in \Pi$ com a menor distância agregada d_g em relação aos seus representantes, desde que não viole qualquer restrição *must-link* e *cannot-link*;
(4) Atualizar os centróides C_p com base nas instâncias por eles representadas;

(5) Retornar ao passo (2) até convergir.

3. AVALIAÇÃO EXPERIMENTAL

Os experimentos foram realizados utilizando quatro conjuntos de dados sintéticos, contendo grupos de diferentes formas e tamanhos. Esses conjuntos de dados foram gerados com base no trabalho apresentado em [Guha et al. 1998] com objetivo de avaliar a capacidade do algoritmo de encontrar formas diversas, em comparação a outros algoritmos da literatura. A Figura 1 apresenta os detalhes dos conjuntos de dados sintéticos.



(a)

	#Instâncias	#Atributos	#Grupos
DB001	4000	2	2
DB002	5000	2	3
DB003	5000	2	5
DB004	3000	2	6

(b)

Fig. 1. Conjuntos de dados utilizados nos experimentos. (a) Visualização. (b) Detalhes dos conjuntos.

Foram definidas três diferentes configurações para o MRS-*kmeans* (A, B e C). Primeiramente, o peso do centróide auxiliar foi fixado em $w_a = 1$ para o representante com maior população e os demais representantes w_a foram normalizados em no máximo 1 proporcionalmente ao tamanho da sua população. A variação ocorre de fato no peso do centróide principal, variando em: $w_p = 1$ (A); $w_p = 2$ (B); $w_p = 3$ (C). A função de distância $d()$ utilizada nos experimentos foi a distância Euclidiana.

Com base nas informações de rótulos, foram gerados conjuntos de restrições (*must-link* e *cannot-link*) aleatórios em 10% das instâncias de cada conjunto de dados. As variações do método proposto foram comparadas com o algoritmo COP-*kmeans* e com o algoritmo MLC-*kmeans* [Huang et al. 2008]. A escolha desses algoritmos deve-se a dois fatos: o COP-*kmeans* é um dos algoritmos mais empregados para comparação nos trabalhos na área de agrupamento semi-supervisionado; e o MLC-*kmeans* é o algoritmo que emprega a abordagem mais similar a apresentada neste artigo.

A metodologia de avaliação adotada para a realização dos experimentos descritos aqui seguiu o que foi apresentado em [Pourrajabi et al. 2014]. Essa metodologia, definida especificamente para avaliar processos de agrupamentos semi-supervisionados, propõe a divisão do conjunto de restrições em conjuntos de treinamento e teste, seguindo a estratégia *10-fold cross validation*. O algoritmo de detecção de agrupamentos é aplicado no conjunto de treinamento e o agrupamento resultante é utilizado para verificar a precisão e revocação do conjunto de teste. Isto é, para cada restrição *must-link* verifica-se se o par de instâncias foi agrupado no mesmo grupo e para cada restrição *cannot-link* verifica-se se o par de instâncias foi agrupado em grupos distintos. Os valores de precisão e revocação são usados para compor a medida F-*Measure*, sendo que, ao final tem-se a F-*Measure* média dos 10 *folds* gerados. Para a obtenção dos resultados descritos aqui os algoritmos foram executados 50 vezes para cada *fold*. Após realizar as execuções, os desempenhos dos algoritmos foram submetidos ao teste estatístico de Friedman e *post-hoc* Nemenyi [Demšar 2006], para constatar quais configurações do MRS-*kmeans* superam estatisticamente os algoritmos da literatura da área.

Os resultados de desempenho dos algoritmos são apresentados na Figura 2(a). Nota-se que o algoritmo MRS-*kmeans* superou os algoritmos COP-*kmeans* e MLC-*kmeans* para todos os conjuntos de dados testados, obtendo resultados expressivos para os conjuntos DB001 e DB002. Isso mostra a eficiência do novo método em encontrar agrupamentos de diferentes tamanhos e formas. Por fim, a Figura 2(b) apresenta o resultado do teste de significância estatística aplicado aos resultados de desempenho dos algoritmos. Observa-se que duas configurações do algoritmo MRS-*kmeas* superam estatisticamente em um nível de confiança de 90% e 95% os algoritmos COP-*kmeans* e MLC-*kmeans*.

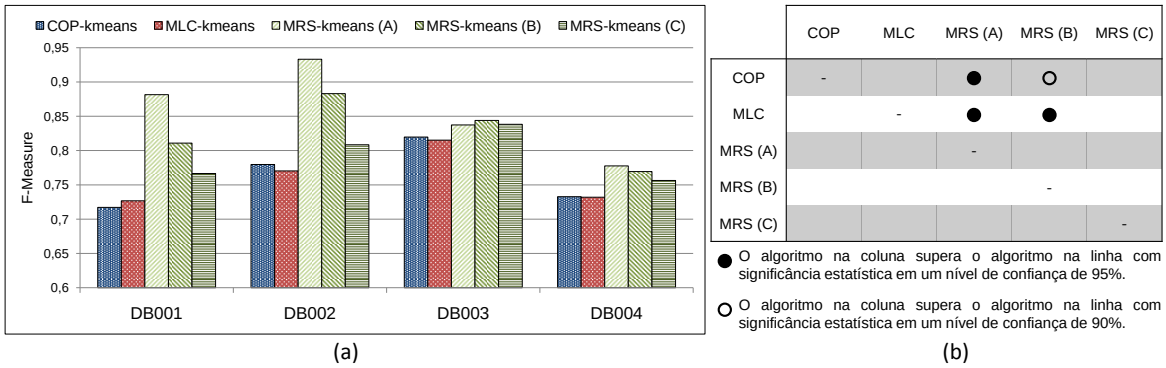


Fig. 2. Resultados dos experimentos. (a) Desempenho das variações do algoritmo MRS-*kmeans* em comparação aos algoritmos da literatura. (b) Teste de significância estatística de Friedman e *post-hoc* Nemenyi.

4. CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

O objetivo do método de agrupamento de dados proposto neste artigo é usar as informações adicionais, em forma de restrições, com o intuito de aumentar a qualidade (**i.e**., a interpretabilidade) dos agrupamentos resultantes. Para tal, o método extrai informações de um conjunto de restrições entre pares de instâncias e incorpora esse conhecimento ao processo de agrupamento, gerando múltiplos representantes que auxiliam na atribuição mais adequada das instâncias aos grupos.

Os resultados experimentais iniciais mostram que o método desenvolvido possui grande potencial para lidar com estruturas de agrupamento mais complexas. O MRS-*kmeans* supera estatisticamente duas abordagens propostas na literatura da área que extraem conhecimento de informações adicionais do conjunto de dados com o intuito de detectar melhores agrupamentos. Dentre os trabalhos futuros que se almeja realizar para dar continuidade ao desenvolvimento do método descrito aqui estão:

—Utilizar a diversidade na seleção de representantes auxiliares, com o intuito de obter um melhor aproveitamento do conhecimento fornecido por eles e diminuir o custo computacional que se paga ao adicionar mais representantes aos grupos quanto ao cálculo de distância;

—Definir estratégias para induzir restrições *must-link* também entre os centróides auxiliares com o objetivo de que esses representantes caracterizem a forma do grupo, contribuindo assim para melhorar a acurácia para grupos de formas diversas;

—Incluir o método proposto em um ambiente visual interativo, possibilitando a utilização de realimentação de relevância com o objetivo de detectar agrupamentos de acordo com a visão do usuário.

REFERÊNCIAS

Barioni, M. C. N., Razente, H. L., Marcelino, A. M. R., Traina, A. J. M., and Traina-Jr., C. Open issues for partitioning clustering methods: an overview. *Wiley Interdisc. Rew.: DMKD* 4 (3): 161–177, 2014.

Basu, S., Davidson, I., and Wagstaff, K. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, 2008.

Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* vol. 7, pp. 1–30, 2006.

Guha, S., Rastogi, R., and Shim, K. CURE: An Efficient Clustering Algorithm for Large Databases. *SIGMOD Rec.* 27 (2): 73–84, 1998.

Huang, H., Cheng, Y., and Zhao, R. A semi-supervised clustering algorithm based on must-link set. In *Int'l Conf. on Advanced Data Mining and Applications*. Berlin, Heidelberg, pp. 492–499, 2008.

Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.* 31 (8): 651–666, 2010.

Pourrajabi, M., Moulavi, D., Campello, R. J. G. B., Zimek, A., Sander, J., and Goebel, R. Model selection for semi-supervised clustering. In *Int'l Conf. on Extending Database Technology (EDBT)*. Atenas, pp. 331–342, 2014.

Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. Constrained k-means clustering with background knowledge. In *Int'l Conf. on Machine Learning (ICML)*. Williamstown, MA, pp. 577–584, 2001.