Agrupamento Hierárquico de Arestas em Redes de Associação

Alessandro M. Silva
alessandro.mattos@aluno.ufms.br

Aline M. Madoenho
aline.mazzuchelli@aluno.ufms.br

Ricardo M. Marcacini
ricardo.marcacini@ufms.br

Grupo de Estudo e Pesquisa em Inteligência Computacional
http://gepic.ufms.br/

INTRODUÇÃO. Um dos principais desafios em redes de associação é lidar com a grande quantidade de vértices e arestas durante a análise dos relacionamentos existentes na rede [1]. Os objetos e suas relações podem ser organizados em grupos, de forma que vértices de um mesmo grupo possuem relacionamentos e características similares [2]. Os algoritmos tradicionais de agrupamento em redes induzem um modelo de agrupamento baseado na similaridade entre vértices, em que a similaridade entre dois vértices é calculada de acordo com a quantidade de vértices vizinhos compartilhados entre eles.

OBJETIVO DO TRABALHO.
Neste trabalho, é investigada a ideia de que é possível induzir melhores modelos de agrupamento ao empregar um critério mais robusto de similaridade, chamado de agrupamento de arestas. Neste caso, além da relação de vizinhança, também são explorados os atributos utilizados para computar a associação entre dois vértices. A abordagem proposta neste trabalho, chamada de **EHC (Edge Hierarchical Clustering)**, possui o diferencial de (i) construir redes de associação de forma automática por meio de algoritmos para extração de regras de associação, e (ii) obter um modelo de agrupamento hierárquico, o que permite explorar os resultados em diversos níveis de abstração por meio de grupos e subgrupos.

EHC (Edge Hierarchical Clustering)

A abordagem EHC possui três etapas, conforme ilustradas na Figura 1: (1) extração de regras de associação; (2) construção de redes de associação; e (3) agrupamento hierárquico de arestas.

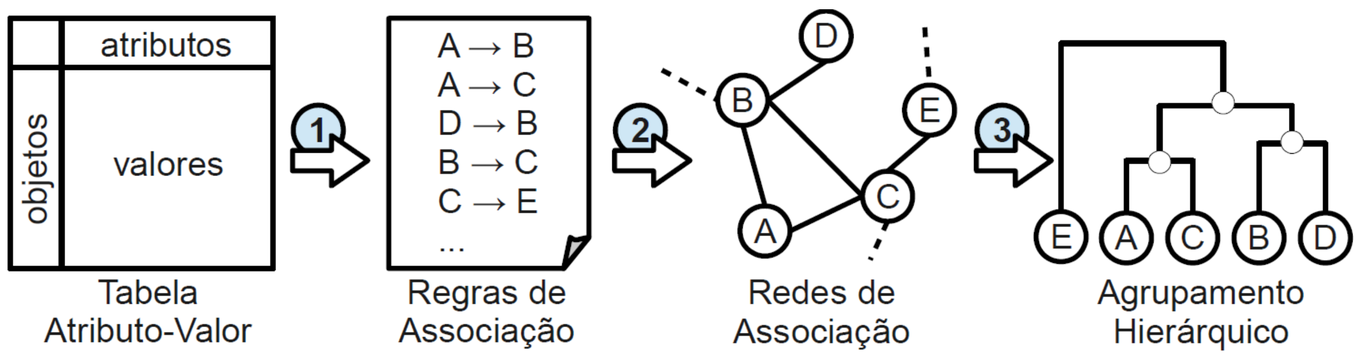

Figura 1. Esquema geral da abordagem EHC.

1 Extração de Regras de Associação

- Cada objeto define um conjunto de transações.
- O algoritmo Apriori [3] é aplicado para extrair regras de associação.
- Diferentes regras de associação podem ser obtidas de acordo com as definições dos parâmetros Suporte e Confiança.
- Na abordagem EHC são extraídas apenas regras de associação compostas por dois itens.

2 Construção de Redes de Associação

- Cada regra de associação define uma aresta na rede
- Dada uma regra A → B, uma aresta é criada para conectar os vértices A e B
- Na abordagem EHC cada aresta é mapeada em um "centroide" para representar a associação no formato atributo-valor

$$\vec{c}_{A \to B} = \frac{1}{m} \sum_{i=1}^{m} \vec{o_i} \qquad \forall \vec{o_i} \in \{A \to B\}$$

$\vec{c}_{A \to B}$: centroide da regra A → B

$\vec{o_i}$: objeto da tabela atributo-valor

3 Agrupamento Hierárquico de Arestas

- O agrupamento é realizado por meio da similaridade entre esses centroides da rede.
- Na abordagem EHC é usado o método *bisecting k-means* [4] para construção do agrupamento hierárquico:
  1. Todas as arestas são alocadas em um único grupo
  2. Cada grupo é dividido em k subgrupos por meio do k-means
  3. As divisões são repetidas até que todas as arestas sejam alocadas em seu único grupo.

Avaliação Experimental

Para avaliar a eficácia do EHC, foi realizada uma análise experimental em seis bases de dados (três textuais e três numéricas). O EHC foi comparado experimentalmente com um algoritmo tradicional de agrupamento (UPGMA) que utiliza apenas a similaridade entre vértices da rede. Uma análise estatística dos resultados indica que a abordagem EHC apresenta resultados superiores, sendo uma alternativa competitiva para agrupamento em redes de associação.

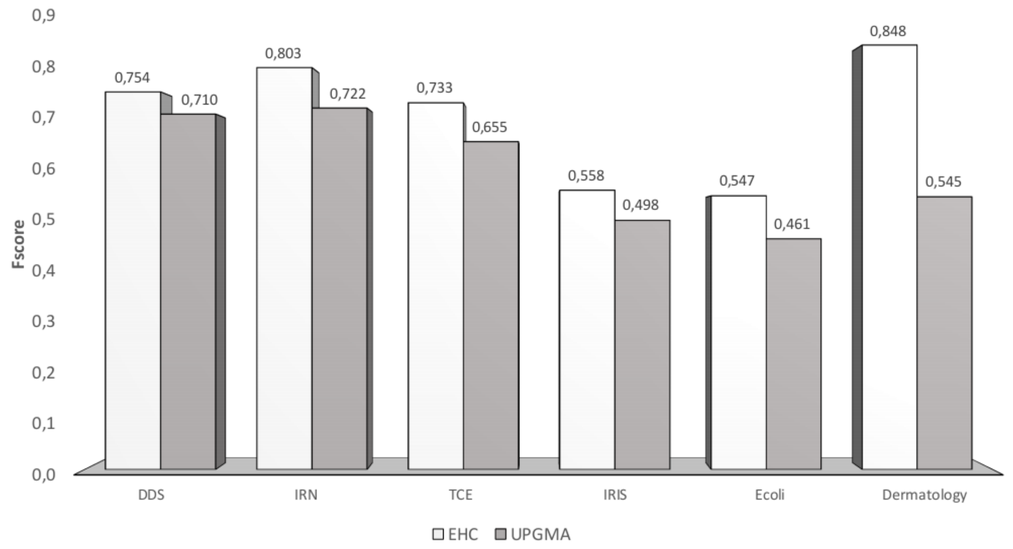

Figura 2. Comparação entre o algoritmo EHC e UPGMA para agrupamento hierárquico em redes de associação.

Código-fonte e Base de Dados disponíveis em:
http://gepic.ufms.br/ehc-eri2014/

Referências

[1] Kolaczyk, E. D. (2009). Statistical Analysis of Network Data: Methods and Models. Springer, 1st edition.

[2] Goldenberg, A., Zheng, A. X., and Fienberg, S. E. (2010). A survey of statistical network models. Foundations and Trends in Machine Learning, 2(2):129-233.

[3] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In 20th International Conference on Very Large Data Bases (VLDB), pages 487–499.

[4] Ye, N. (2013). Data Mining: Theories, Algorithms, and Examples. CRC Press.