Agrupamento Hierárquico de Arestas em Redes de Associação

Alessandro M. Silva[1], **Aline M. Madoenho**[1], **Ricardo M. Marcacini**[1]

[1]Universidade Federal de Mato Grosso do Sul (UFMS)
Grupo de Estudo e Pesquisa em Inteligência Computacional (GEPIC) - UFMS/CPTL

{alessandro.mattos,aline.mazzuchelli}@aluno.ufms.br
ricardo.marcacini@ufms.br

1. Introdução

O uso de redes para representação de informação vêm sendo investigado de formas distintas em várias áreas da ciência. Por exemplo, Cientistas Sociais buscam interpretar o significado das relações geradas em redes sociais, como os padrões de interação entre pessoas e formação de comunidades. Já em Ciência da Computação, as redes são exploradas por meio de algoritmos utilizados em Aprendizado de Máquina e Mineração de Dados e Textos. Em geral, as redes são investigadas tanto para tarefas descritivas, que buscam extrair conhecimento das relações representadas na rede, quanto para tarefas preditivas, que visam classificar novas relações com base nas relações históricas da rede [Kolaczyk 2009].

De forma geral, qualquer conjunto de dados representado de forma proposicional, ou seja, por meio de uma tabela atributo-valor, pode ser convertido em uma rede a partir da associação de seus objetos e/ou atributos. Por este motivo, nos últimos anos diversos trabalhos têm nomeado tais representações como *redes de associação*, sendo apresentadas como uma proposta promissora para extração de conhecimento em base de dados e textos [Goldenberg et al. 2010]. Em particular, as redes de associação se tornaram populares por facilitar a identificação de relações significativas entre objetos e atributos, bem como uma exploração visual do conhecimento [Hadzic et al. 2011, Ye 2013].

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 [Hadzic et al. 2011]. Os objetos e suas relações podem ser organizados em grupos, de forma que vértices de um mesmo grupo possuam relacionamentos e características similares. 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. 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. 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 que utiliza apenas

a similaridade entre vértices da rede. Uma análise estatística dos resultados indica que a abordagem EHC apresenta resultados superiores, sendo um alternativa competitiva para agrupamento em redes de associação.

2. Desenvolvimento

A abordagem EHC possui três etapas principais: (1) extração de regras de associação; (2) construção de redes de associação; e (3) agrupamento hierárquico de arestas. Na Figura 1 é ilustrado um esquema geral da abordagem e suas etapas (identificadas pelas setas). Os detalhes de cada etapa são apresentados abaixo:

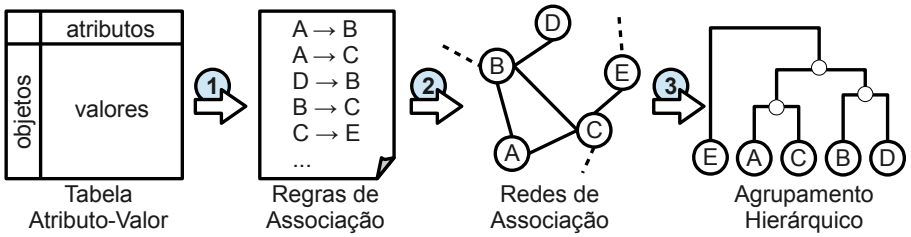

Figura 1. Esquema geral da abordagem EHC (*Edge Hierarchical Clustering*).

Etapa 1. Extração de Regras de Associação: A partir de um conjunto de dados estruturado por uma tabela atributo-valor, a extração de regras de associação é realizada usando o algoritmo Apriori [Agrawal and Srikant 1994]. Neste caso, cada objeto define um conjunto de transações, onde as transações são os atributos presentes no objeto. Em algumas situações, é necessário pré-processar o conjunto de dados para discretização dos atributos. Diferentes regras de associação podem ser obtidas de acordo com as definições dos parâmetros *suporte* e *confiança* do algoritmo Apriori. Na abordagem EHC são extraídas apenas regras de associação compostas por dois itens.

Etapa 2. Construção de Redes de Associação: Cada regra de associação obtida na etapa anterior define uma aresta na rede de associação. Desta forma, dada uma regra $A \rightarrow B$, 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. Formalmente, dada uma regra $A \rightarrow B$, o vetor centroide $\vec{c}_{A \rightarrow B}$ é computado por meio da Equação 1, em que $\vec{o_i}$ é um objeto da tabela atributo-valor que pertence ao conjunto de m objetos cobertos pela regra.

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

Enquanto a aresta enfatiza a associação direta entre os atributos A e B, o centroide tem a função de sumarizar os diferentes atributos indiretamente relacionados a esta associação por meio de um vetor médio dos objetos cobertos pela regra.

Etapa 3. Agrupamento Hierárquico de Arestas: Uma vez que cada aresta da rede de associação possui um determinado centroide mapeado, o agrupamento é realizado por meio da similaridade entre esses centroides. Na abordagem EHC é instanciado o método *bisecting k-means* para construção do agrupamento hierárquico. Inicialmente, todas as arestas são alocadas em um único grupo. Em seguida, cada grupo é dividido em k subgrupos por meio do *k-means*. As divisões são repetidas até que todas as arestas sejam alocadas em seu único grupo. No final, cada um dos dois vértices que constitui uma aresta se torna um grupo (nó folha), finalizando a indução do modelo de agrupamento.

Para avaliar a eficácia do método EHC foi realizada uma avaliação experimental em seis conjuntos de dados de *benchmark*. Os conjuntos DDS, IRN e TCE são bases textuais, enquanto os conjuntos Iris, Ecoli e Dermatology são bases numéricas. O índice F_{SCORE} [Ye 2013], baseado em precisão e revocação, foi utilizado para avaliar a acurácia do modelo. Neste índice, quanto mais próximo de 1, melhor a acurácia do modelo. Para a construção das redes de associação foi utilizado um suporte mínimo de 2 objetos por regra. Para o agrupamento, foi adotada a medida de similaridade cosseno. Os resultados do EHC foram comparados com o tradicional UPGMA (utilizando similaridade entre os vértices) [Ye 2013]. Na Figura 2 são apresentados os resultados experimentais obtidos. Uma análise estatística por meio do teste de Wilcoxon (com 95% de confiança) indica que o EHC obtém resultados superiores ao UPGMA. Os detalhes sobre a análise estatística, bem como os conjuntos de dados utilizados estão disponíveis em `http://gepic.ufms.br/ehc-eri2014`.

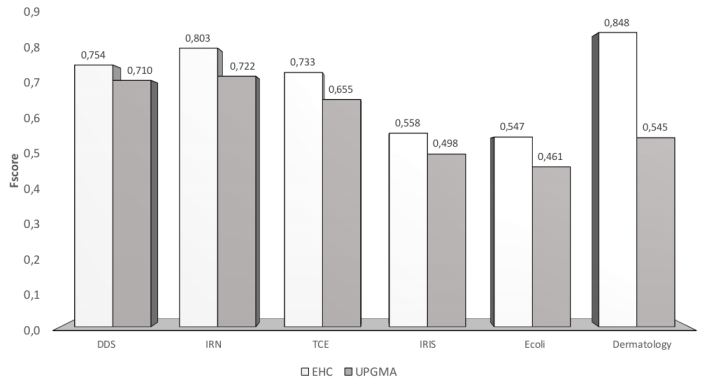

Figura 2. Comparação entre o algoritmo EHC e UPGMA para agrupamento hierárquico em redes de associação.

3. Considerações Finais

Neste trabalho foi proposta e avaliada a abordagem EHC (*Edge Hierarchical Clustering*), que explora similaridade entre arestas de redes de associação para induzir modelos de agrupamento mais robustos. A análise experimental realizada fornece evidências de que o EHC obtém estruturas de agrupamento mais significativas, quando comparado com agrupamento baseado apenas na informação de vizinhança dos vértices. As direções para trabalhos futuros envolvem avaliar o EHC em um conjunto maior de base de dados, bem como o desenvolvimento de ferramentas que permitam investigar a eficácia do EHC para exploração visual dos dados.

Referências

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.

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.

Hadzic, F., Dillon, T. S., and Tan, H. (2011). *Mining of Data with Complex Structures*, chapter Graph Mining, pages 287–298. Springer.

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

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