A ser submetido ao periódico *Plos One* - ISSN: 1932-6203
http://www.plosone.org/

Agrupamento Hierárquico de Arestas em Redes de Associação

Alessandro M. Silva (aluno), Ricardo M. Marcacini (orientador)

Grupo de Estudo e Pesquisa em Inteligência Computacional (GEPIC) - http://gepic.ufms.br
***E-mails**: alessandro.mattos@aluno.ufms.br, ricardo.marcacini@ufms.br

Resumo
Um desafio atualmente existente é a exploração de grandes bases de dados de forma inteligente visando extrair um conhecimento que pode ser valioso para obtenção de vantagem competitiva. Nesse sentido, métodos e algoritmos têm sido investigados e utilizados em tarefas de análise exploratória dos dados, em que o objetivo está em buscar padrões interessantes implícitos em bases de dados. Os algoritmos de agrupamento hierárquico são muito úteis para apoiar tarefas de análise exploratória, uma vez que permitem organizar, de forma não supervisionada, um conjunto de dados em grupos e subgrupos. Assim, é possível explorar grandes conjuntos de dados em diversos níveis de granularidade. Neste trabalho é investigado o uso de agrupamento hierárquico baseado em redes de associação, que são estruturas que permitem identificar padrões baseados em regras de associação extraídas nos dados. Uma abordagem denominada EHC (*Edge Hierarchical Clustering*) foi proposta e avaliada neste trabalho, que possui o diferencial de explorar tanto as relações entre os atributos do conjunto de dados (modelo relacional), quanto os valores de relevância desses atributos (modelo proposicional). A ideia geral do EHC é computar uma melhor medida da similaridade entre os objetos, permitindo a indução de modelos de agrupamento mais robustos. Uma análise estatísica dos resultados de avaliação experimental demonstrou que a abordagem EHC permite induzir modelos de agrupamentos de maior qualidade do que uma abordagem tradicional de agrupamento hierárquico baseada apenas no modelo relacional.

Palavras-chave
Agrupamento Hierárquico — Redes de Associação — Mineração de Dados

Sumário

1 Introdução **1**
2 Revisão da Literatura **4**
2.1 Pré-Processamento de Dados 4
Modelo Proposicional • Modelo Relacional
2.2 Extração de Padrões 7
Regras de Associação • Agrupamento de Dados
2.3 Pós-Processamento dos Dados 10
3 A abordagem EHC (Edge Hierarchical Clustering) **12**
4 Avaliação Experimental **13**
5 Prova de Conceito: Mineração de Dados do SISU **17**
6 Considerações Finais **20**
Referências **22**

1. Introdução

A popularização de plataformas online para publicar e armazenar conteúdo digital tem promovido um aumento significativo do volume de dados produzido pelas organizações [1]. Esses dados representam um importante repositório de conhecimento, uma vez que contém o histórico de diversas atividades, informações de pessoas, projetos, produtos, estratégias e o próprio conhecimento adquirido ao longo do tempo [2]. Nesse sentido, um desafio atualmente existente é a

exploração dessas grandes bases de dados de forma inteligente visando extrair um conhecimento que pode ser valioso para obtenção de vantagem competitiva [3]. Por outro lado, este imenso volume de dados inviabiliza a análise e exploração manual dos dados. Diante a este desafio, diversos métodos e algoritmos têm sido investigados em temas de pesquisa como Mineração de Dados e Textos, Aprendizado de Máquina e Big Data. Tais métodos e algoritmos geralmente são utilizados em tarefas de análise exploratória dos dados, em que o interesse está em buscar padrões interessantes implícitos em bases de dados [2].

Dentre métodos existentes para análise exploratória de bases de dados, os métodos para (i) regras de associação e (ii) agrupamento hierárquico têm recebido destaque na literatura. As **regras de associação** são muito populares para encontrar relacionamentos ou padrões frequentes em um conjunto de dados, facilitando analisar as estruturas implícitas nos dados. O formato geral de uma regra de associação é uma relação do tipo $A \Rightarrow B$, no qual A e B são conjuntos geralmente formados pelos items de uma base de dados — o que auxilia aspectos relacionados à interpretação dos padrões extraídos. Por exemplo, se a regra $\{leite, pão\} \Rightarrow \{manteiga\}$ é obtida a partir de uma base de dados com transações de compras, então é possível afirmar que clientes que compram leite e pão também tendem a comprar manteiga, dado um determinado nível de certeza. Embora existam vários algoritmos para extração de regras de associação, a análise de um conjunto de regras extraídas, bem como a qualidade da associação obtida, é uma limitação atual e um desafio de pesquisa em aberto. Além disso, dependendo dos parâmetros utilizados para a extração de regras, como o número de itens considerados na associação, o processo geralmente apresenta alto custo computacional [3].

Já os métodos de **agrupamento hierárquico** permitem organizar objetos de um conjunto de dados em grupos de acordo com uma medida de proximidade, em que objetos pertencentes a um mesmo grupo são altamente similares entre si [4]. Além disso, esta organização de agrupamento é baseada em uma estrutura hierárquica denominada dendrograma, na qual os grupos pais representam conhecimento mais generalizado, enquanto grupos filhos representam detalhamentos e conhecimento mais específico. A estrutura obtida por métodos de agrupamento hierárquico é muito útil para organizar, recuperar e visualizar informação em diferentes níveis de abstração e explorar grandes bases de dados. Por outro lado, uma conhecida limitação e também tema de pesquisa desses métodos é lidar com a dificuldade de interpretação do significado dos agrupamentos, principalmente, em tarefas de análise exploratória [2].

Neste trabalho é investigada e avaliada uma abordagem para agrupamento hierárquico baseado em **redes de associação** [5, 6, 7], visando lidar com os desafios e limitações comentados acima. Redes de associação são representações alternativas e promissoras para analisar regras de associação. Neste caso, as regras são organizadas em um grafo, onde cada vértice representa um item do conjunto de dados. Dois itens a e b são conectados por uma aresta se existe uma regra de associação $\{a\} \Rightarrow \{b\}$. Esta representação tem um menor custo computacional para ser obtida, uma vez que trabalha apenas com extração de regras compostas por dois itens, além de facilitar a análise visual das associações. Ainda, tais redes permitem identificar outros tipos mais complexos de associação, explorando-se as estruturas de

de agrupamento identificadas no grafo [6, 7].

Para ilustrar o uso de redes de associação e métodos de agrupamento hierárquico, considere a Figura 1, que apresenta uma rede de associação extraída a partir do conjunto sobre "Dados de Censo de Idosos". Este conjunto contém informações sobre a distribuição demográfica e renda dos idosos dos EUA. Na parte superior da figura, é apresentada a rede de associação composta por cinco regras de associação extraída dos dados sem considerar a implicação das regras (grafo não dirigido). Na parte inferior, o respectivo dendrograma foi obtido com um método de agrupamento hierárquico tradicional, em que a similaridade entre dois vértices é baseada na quantidade de vizinhos que estes vértices compartilham. Enquanto os nós folhas contém as regras de associação originalmente extraídas, os grupos e subgrupos representam novas estruturas formadas pela combinação dessas regras. O uso desta estrutura depende dos objetivos da aplicação, como a recomendação de consultas em banco de dados, tarefas de segmentação e visualização dos padrões extraídos.

A principal contribuição deste trabalho é a proposta da abordagem *EHC* (*Edge Hierarchical Clustering*), baseada no conceito de agrupamento de arestas [8], bem como a avaliação de seu uso no processo de Mineração de Dados (MD). Ao contrário dos métodos de agrupamento existentes que exploram uma medida de similaridade entre vértices da rede, a proposta do *EHC* é induzir melhores modelos de agrupamento ao empregar um critério mais robusto de similaridade, chamado de agrupamento de arestas [8]. Neste caso, além da relação de vizinhança da rede, também são explorados os objetos originalmente utilizados para computar a associação entre dois vértices. Cada

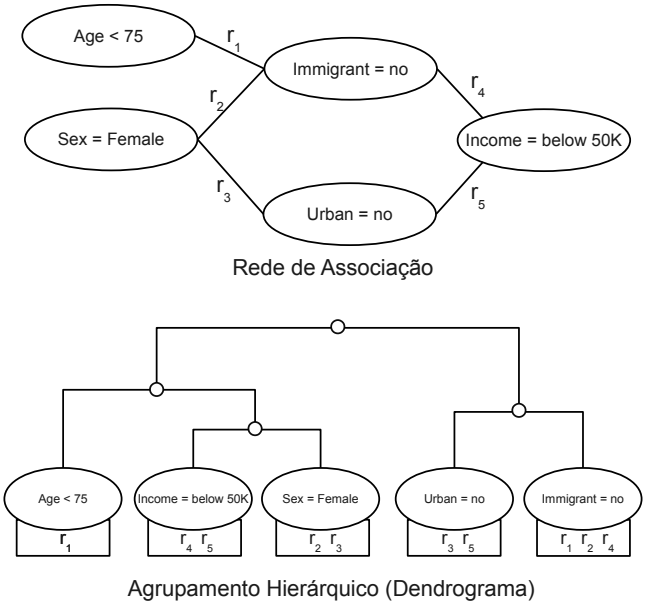


Figura 1. Exemplo de agrupamento hierárquico em redes de associação (adaptado de [6]).

aresta contém um vetor de características — extraído dos objetos relacionados à regra de associação — que é empregado para melhorar o cálculo da similaridade entre as arestas [8].

A abordagem proposta *EHC* foi avaliada experimentalmente utilizando-se seis conjuntos de dados de *benchmarking*, três de dados numéricos e três de dados textuais. A abordagem *EHC* foi comparada com um método de agrupamento hierárquico tradicional da literatura, denominado *UPGMA*, e uma análise estatística dos resultados experimentais demonstrou que o *EHC* apresenta resultados superiores em relação à qualidade do modelo de agrupamento. Para analisar a eficácia do *EHC* em um cenário prático, este trabalho colaborou com o projeto: *Mineração de Dados em Informações Acadêmicas da UFMS*[1], em que foi desenvolvido um ambiente web para apoiar a análise exploratória de dados provenientes do SISU/-MEC (Sistema de Seleção Unificada/Minis-

[1] http://sigproj1.mec.gov.br/apoiados.php?projeto_id=136714

tério da Educação)[2]. Neste ambiente, a abordagem *EHC* foi executada (como prova de conceito) em dados socioeconômicos de alunos matriculados na UFMS no ano de 2012. Os resultados preliminares indicam que o *EHC* pode ser empregado para recomendar consultas ao banco de dados, visando filtrar subconjuntos de dados que possam ser interessantes durante a análise exploratória.

O restante deste texto está organizado da seguinte forma: na Seção 2 é realizada uma revisão bibliográfica dos trabalhos e conceitos mais relevantes para a compreensão deste trabalho, como pré-processamento de dados, extração de padrões e pós-processamento de dados. Na Seção 3, é apresentada a abordagem proposta *EHC* (*Edge Hierarchical Clustering*), incluindo suas etapas e definições. Os resultados obtidos na avaliação experimental com dados de *benchmarking* são discutidos na Seção 4. Na Seção 5 é apresentada uma prova de conceito do uso da abordagem *EHC* em cenários práticos, com seu uso em mineração de dados acadêmicos do SISU. Finalmente, as considerações finais e direções para trabalhos futuros são discutidas na Seção 6.

2. Revisão da Literatura

Nesta seção é apresenta uma revisão da literatura, discutindo-se os principais trabalhos e os fundamentos teóricos relacionados ao desenvolvimento deste trabalho. Este trabalho engloba os conceitos investigados no tema de Mineração de Dados, que é definido como um conjunto de técnicas para descoberta, extração e organização de conhecimento em bases de dados [9]. Nesta seção, a discussão dos trabalhos relacionados está organizada em três subseções, que representam as grandes etapas do processo de Mineração de Dados.

Inicialmente, serão discutidos os trabalhos relacionados ao pré-processamento de dados, etapa necessária para preparar os dados em uma estrutura adequada para o processamento. Em seguida, são apresentados os algoritmos utilizados neste trabalho para a etapa de extração de padrões, como regras de associação e agrupamento de dados. Por fim, é realizada uma discussão sobre o pós-processamento de dados que lida com estratégias para avaliar os modelos obtidos na extração de padrões.

2.1 Pré-Processamento de Dados

Os trabalhos em Mineração de Dados, seja de dados ou textos, geralmente lidam com fontes de dados muito extensas e de diversos tipos (bancos de dados, logs de acesso, relatórios, etc). Dessa forma, faz-se necessário o uso de algumas técnicas para otimizar todo o processo, que fazem parte da etapa de pré-processamento dos dados. As principais tarefas para o pré-processamento de dados envolvem a limpeza, integração, seleção, redução e transformação desses dados [10]. Essas tarefas serão discutidas a seguir.

A partir de um conjunto de dados de interesse, efetua-se uma limpeza, que consiste na busca por valores que possam comprometer a qualidade dos dados, como valores em branco, valores errados, registros incompletos, entre outras. Esta etapa visa eliminar estes tipos de inconsistências de forma que elas não influam no resultado do processo de mineração [11].

Em geral, é comum obter-se dados de variadas fontes (data warehouses, arquivos texto, imagens, vídeos, entre outras), assim utiliza-se técnicas de integração dos dados, de forma que obtenha-se um repositório único e consistente. Tais técnicas buscam analizar redundâncias, dependências entre variáveis e valores conflitantes [12].

[2]SISU - Sistema de Seleção Unificada: `http://sisu.mec.gov.br`

Devido as restrições de espaço em memória ou tempo de processamento, a quantidade de exemplos e de atributos disponíveis para a análise pode tornar a aplicabilidade do algoritmo de extração de padrões inviável. Desta maneira, pode ser necessário aplicar, antes de iniciar o processo de extração, técnicas de seleção e redução de dados. A redução pode ser feita de três maneiras [10]:

1. Redução do número de exemplos, que consiste em gerar amostras representativas do conjunto de dados original;

2. Redução do número de atributos, objetiva selecionar um subconjunto de atributos que são potencialmente mais relevantes para o modelo final, no entanto, por não se saber quais atributos são importantes para se atingir os objetivos, remove-se somente àqueles que, com certeza, não têm relevância para o modelo final; e

3. Redução do número de valores de um atributo, feita por discretização, que consiste na substituição de um atributo contínuo (real ou inteiro) por um discreto, utilizando agrupamento de valores; ou então suavização dos valores de um atributo contínuo, que objetiva diminuir o número de valores do mesmo sem discretizá-lo.

A etapa de transformação dos dados é bastante específica e varia conforme os dados abordados e a aplicabilidade do algoritmo de extração de padrões que será adotado. Muitos desses algoritmos trabalham apenas com valores numéricos e outros com valores categóricos, assim a etapa de transformação consiste em adaptar os dados no formato apropriado através de operações como generalização (conversão de valores específicos em genéricos), normalização (definir escala geral entre variáveis) e criação de novos atributos.

É importante destacar que este processo de pré-processamento é iterativo e pode ser repetido diversas vezes até que os dados estejam adequados para serem utilizados na etapa de extração de padrões.

Por fim com o objetivo de representar os dados em uma estrutura que possa ser processada pelos algoritmos de extração de padrões, é comum o uso de dois tipos de modelos: o modelo proposicional e o relacional.

2.1.1 Modelo Proposicional

O modelo utilizado para representação dos dados é de vital importância para o processo de mineração, pois ele define como os dados serão submetidos aos algoritmos de extração de padrões. Um dos modelos mais utilizados é o modelo espaço-vetorial que consiste em representar cada objeto como um vetor em um espaço multi-dimensional [10], onde cada dimensão é um atributo presente no conjunto de dados. No modelo proposicional, os dados são representados por meio de uma tabela atributo-valor, como ilustrado na Tabela 1, onde o_i corresponde ao i-ésimo objeto, t_j representa o j-ésimo atributo e a_{ij} é um valor que relaciona o i-ésimo objeto com o j-ésimo atributo.

	T_1	T_2	$\ldots$	T_j
O_1	a_{11}	a_{12}	$\ldots$	a_{1j}
O_2	a_{21}	a_{22}	$\ldots$	a_{2j}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
O_i	a_{i1}	a_{i2}	$\ldots$	a_{ij}

Tabela 1. Esquema geral de uma tabela atributo-valor para representação de dados no modelo proposicional.

Através da Tabela 1, cada objeto é representado por um vetor $\vec{v_i} = (a_{i1}, a_{i2}, \ldots, a_{iM})$, no qual o valor de a_{ij} indica a relevância do atributo j para o objeto i. Algumas estratégias para computar tais valores são descritas a

seguir:

- Um valor que indica a ocorrência ou não de um atributo nos objetos, representado como um binário (1 se presente e 0 se ausente);
- Um valor nominal de um conjunto finito de valores distintos, geralmente utilizado para representar atributos categóricos (ex: cidade, estado, sexo, etc) ou atributos ordinais (ex: grau de instrução); e
- Um valor contínuo que indica a relevância ou grandeza escalar do atributo sobre o objeto (ex: salário, altura, peso, etc).

Também é importante observar que os atributos podem ser organizados em dois tipos [10]: qualitativos, representados pelos valores binários, nominais e ordinais; e quantitativos, representados pelos valores numéricos.

A representação dos dados através de uma tabela atributo-valor permite o emprego de inúmeros algoritmos de extração de padrões. Ainda, seu processo de concepção pode ser repetido nas demais etapas da mineração, pois pode-se detectar novos padrões que definam melhorias na tabela atributo-valor, como a ponderação ou a seleção dos atributos.

2.1.2 Modelo Relacional

Em contraste com o modelo proposicional, a representação dos dados por meio de um modelo relacional é baseada em uma estrutura de rede, na qual os atributos e/ou objetos são conectados por meio de uma aresta quando existe uma relação entre eles [13].

De forma simplista, uma rede nada mais é do que a interconexão entre objetos (ou atributos), assim, basicamente os estudos sobre esta área, objetivam a análise de medições, aspectos das relações, bem como a complexidade das dependências geradas [14]. Formalmente, uma rede é definida como um grafo G = (V, E, W), em que V é um conjunto de vértices (nós da rede), E é um conjunto de arestas $E = (e_1, e_2, \ldots, e_n)$ – em que cada aresta conecta dois vértices – e $W = (w_1, w_2, \ldots, w_n)$ é um conjunto de pesos de cada aresta que define o grau de relação entre dois vértices [15].

Considerando o dinamismo contextual em que podemos empregar o uso de redes, Newman [16] categoriza modelos relacionais em quatro tipos: redes tecnológicas, redes de informação, redes biológicas e redes sociais.

A análise de redes origina-se de diferentes problemas com diferentes interpretações e interesses. Cientistas sociais buscam interpretar o significado das relações geradas em uma rede social, para posteriormente identificar o tipo de estrutura que a rede segue. Já os Físicos procuram entender, por exemplo, os mecanismos de formação da rede, considerando sua distribuição. Biólogos buscam definir a afinidade para a ligação física entre proteínas, e também a formação de redes de regulação genética.

As Redes Tecnológicas são redes construídas pelo homem para suprir a necessidade de melhoria sob a distribuição de um recurso. Um exemplo são as redes de telecomunicações, com as linhas de transmissão de televisão extendendo-se por uma determinada região. Outras aplicações são as redes de telefonia, elétrica e de conexões entre computadores (Internet).

Uma Rede de Informação caracteriza-se por armazenar determinado tipo de informação ou conhecimento em seus vértices. Exemplos são as redes de citações entre artigos acadêmicos e a própria rede mundial de computadores (*World Wide Web*); que foi concebida a partir da proposta de hipertex-

tos – textos ligados através de links – em que as páginas representam os nós que são referenciados por outras páginas, estabelecendo relações entre as mesmas.

Redes Biológicas são naturalmente e comumente, a forma de representação de sistemas biológicos em diferentes escalas. Por exemplo, a descrição completa de interação entre organismos, estabelecendo as relações predador-presa, e tambem a propagação de uma doença em uma dada população.

As Redes Sociais modelam as ligações relacionais entre um conjunto de indivíduos – pessoas, empresas, entre outros - ou grupos de indivíduos – por exemplo estados-nações de um continente, departamentos dentro de uma organização, entre outros – determinados por um padrão de contatos ou interações. Um exemplo é a rede de coautoria, na qual os cientistas são representados nos vértices do grafo e as arestas relacionam a outros cientistas que colaboraram para a publicação de um dado artigo. Redes Socias têm sido alvo de inúmeras pesquisas recentes devido ao forte impacto de seu estudo na vida real, por exemplo: análise de conteúdo da Web para melhorias em negócios, análise social para fins governamentais, pesquisas envolvendo a recuperação de informação, entre outros.

Como vemos, a pesquisa em redes é multidisciplinar e engloba diversos sistemas naturais dinâmicos que são representados através de grafos. Mais recentemente, diversos estudos demonstram que qualquer conjunto de dados pode ser representado por meio do modelo relacional, mesmo que estejam representados originalmente pelo modelo proposicional [14, 15]. Nesse caso, é necessário buscar inicialmente por associações entre os objetos e, então, organizá-los por meio de um grafo. Por este motivo, este modelo relacional é definido como “Rede de Associação”; que é investigada neste trabalho na tarefa de Mineração de Dados.

2.2 Extração de Padrões

Esta etapa compreende a seleção da tarefa de Mineração de Dados que será empregada no processo de extração dos padrões. Essa fase deve ser cuidadosamente avaliada, pois o algoritmo escolhido deve se ajustar à solução do problema identificado de forma a obter uma melhor representação dos padrões encontrados. Portanto, esta etapa pode ser repetida, aplicando a extração de padrões mais de uma vez com o intuito de obter resultados adequados aos objetivos pré-estabelecidos [9].

A extração de padrões pode ser dividida em três grandes atividades, a *Escolha da função*, *Escolha do algoritmo* e a *Obtenção dos padrões* [9]. Para a escolha da função, há duas abordagens distintas que podem ser empregadas, as Preditivas que buscam a previsão de um determinado padrão dentro do conjunto dos dados, utilizando-se de experiências passadas e fornecendo respostas conhecidas; e as Descritivas que buscam a identificação de comportamentos siginificantes nos dados que possam prover um padrão nos dados.

Dessa forma, a escolha do algoritmo é realizada de acordo com a função de Mineração de Dados e deve considerar também a complexidade e interpretação do padrão extraído para os usuários. A seguir, serão apresentados dois métodos para extração de padrões usando abordagem Descritiva que são de interesse para este trabalho: Regras de Associação e Agrupamento de Dados.

2.2.1 Regras de Associação

Uma Regra de Associação caracteriza o quanto a presença de um conjunto de itens nos registros de uma base de dados implica na presença de algum outro conjunto distinto

de itens nos mesmos registros [17, 18]. Nesse caso, os itens são os atributos (não nulos) e respectivos valores de um objeto. Por exemplo, um objeto "pessoa" pode apresentar os itens $\{Sexo = F, Salario = Alto, GrauInstrucao = Superior\}$. É importante notar a necessidade de transformação dos atributos para um formato apropriado, uma vez que não é possível o uso de atributos contínuos em regras de associação [17].

Uma regra de associação busca encontrar tendências comportamentais nos dados que explicitem o entendimento dos mesmos para a exploração de padrões. O formato de uma Regra de Associação pode ser representado como uma implicação $LHS \Rightarrow RHS$, em que LHS e RHS são, respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra, definidos por conjuntos disjuntos de itens.

Para cada regra ($LHS \Rightarrow RHS$), extraída de um conjunto de transações T (modelo proposicional após a transformação dos dados), é dado um valor de suporte *(sup%)* que verifica a força de associação entre LHS e RHS; e um valor de confiança *(conf%)* que mede a força da implicação lógica da regra. O valor do suporte de uma regra $LHS \Rightarrow RHS$ é calculado como $sup(LHS \cup RHS)$, que mede a quantidade de transações em que os conjuntos LHS e RHS ocorrem juntos. Já o valor de confiança da regra é calculado pela relação $conf(LHS \Rightarrow RHS) = \frac{\sup(LHS \cup RHS)}{sup(LHS)}$ [17].

Ao conjunto de atributos ou itens ordenados lexicograficamente dá-se o nome de itemset [17]. Um itemset com k elementos costuma ser referenciado como k-itemset. Um itemset é considerado frequente quando seu suporte é maior ou igual ao suporte mínimo definido pelo usuário.

Para determinar k-itemsets frequentes utiliza-se o algoritmo conhecido como *Apriori*, que é um dos 10 algoritmos mais utilizados em tarefas de Mineração de Dados [19]. O funcionamento do algoritmo pode ser descrito da seguinte forma. Primeiramente é calculado a ocorrência dos itens, determinando o 1-itemsets frequentes que são armazenados em D_1. O passo seguinte, de forma iterativa, o algoritmo constrói um conjunto de k-itemsets candidatos (c_k) para encontrar o D_k. A seguir poda-se c_k, utilizando a propriedade de que se um itemset I não é frequente então para todo itemset A, $I \cup A$ não será frequente. Por fim, são formados os k-itemsets frequentes pertencentes a D_k.

Um itemset é considerado uma regra de associação quando seu valor de confiança é maior do que um valor definido pelo usuário [17].

2.2.2 Agrupamento de Dados

O objetivo dos algoritmos de agrupamento de dados é encontrar a formação de distintos grupos dentro do conjunto de dados [2]. Um grupo (ou cluster) consiste em uma coleção de objetos similares entre si, no entanto distintos dos outros objetos nos demais agrupamentos. O agrupamento é considerado uma tarefa não supervisionada, pois não depende de classes predefinidas (categorias) para classificar os objetos [4]. Os grupos são formados em função de uma medida de proximidade entre esses objetos e, consequentemente, dos atributos (variáveis) selecionados para representação desses objetos [20]. Desta forma, para a tarefa de agrupamento é necessário definir uma medida de proximidade entre os objetos.

As medidas de proximidade são utilizadas para medir o quão dois objetos são similares ou estão correlacionados em relação a seus atributos [20]. Para tal definição, depende-se do tipo de dados ao qual as medidas são aplicadas, sejam dados contínuos, discretos (binários, categóricos e ordinais) ou a mistura dos mesmos. É possível calcular a simi-

laridade como também a dissimilaridade (ou distância) e aplicar o inverso a partir de um cálculo já realizado. A seguir são apresentadas três medidas amplamente exploradas no cálculo de proximidade entre dois objetos [3]: similaridade baseada no Cosseno, Jaccard e Distância Euclidiana.

- Similaridade Cosseno: definida pelo ângulo cosseno formado entre os vetores de dois objetos o_i e o_j, conforme ilustrada na Equação 1:

$$cosseno(o_i, o_j) = \frac{o_i \cdot o_j}{\|o_i\|\|o_j\|} \quad (1)$$

  O valor resultante está no intervalo $[0,1]$ quando utilizada em dados textuais e $[-1,1]$ quando os atributos podem receber valores negativos.

- Similaridade de Jaccard: Utilizada para os casos em que os vetores expressam valores binários, indicando presença ou ausência de determinado atributo no objeto. Assim, dado dois objetos o_i e o_j, a similaridade Jaccard calcula a relação entre a quantidade de atributos que ocorrem nos dos dois objetos, dividida pela quantidade total de atributos, conforme ilustrada na Equação 2;

$$jaccard(o_i, o_j) = \frac{atributos(o_i) \cap atributos(o_j)}{atributos(o_i) \cup atributos(o_j)} \quad (2)$$

- Distância Euclidiana: Muitas vezes os m atributos de dois objetos o_i e o_j são contínuos e neste caso utilizamos a distância euclidiana como descrita na Equação 3:

$$euclidiana(o_i, o_j) = \sqrt{\sum_{i=1}^{m} (o_{im} - o_{jm})^2} \quad (3)$$

As estratégias de agrupamento podem utilizar diferentes aspectos quanto a sua classificação, gerando assim uma problemática em qual técnica adotar no processo de mineração. Comumente, esta escolha é baseada de acordo com a natureza dos dados analisados, bem como da validação da qualidade do agrupamento obtido [2].

Os métodos de agrupamento, são divididos em agrupamento particional e agrupamento hierárquico [20]. A diferença fundamental entre as duas formas de organização é que o agrupamento particional, divide a coleção de objetos em uma partição simples de **k** grupos, já o agrupamento hierárquico produz uma sequência de partições aninhadas, criando assim uma hierarquia de grupos e subgrupos [2].

O **Agrupamento Particional**, ou por otimização, tem como fundamento a divisão iterativa do conjunto de objetos em k distintos grupos, em que k é um número geralmente informado pelo usuário. O algoritmo mais representativo desta técnica é o **K-Means** [19]. Este algoritmo utiliza o conceito de centroide, que consiste em um vetor médio computado a partir dos demais vetores do grupo. O cálculo do centroide é obtido a partir da Equação 4, onde x representa um objeto pertencente ao grupo G, e $|G|$ engloba todos os objetos da coleção:

$$C = \frac{1}{\|G\|} \sum_{x \in G} x \quad (4)$$

Como visto, o centroide é muito representativo, prevalecendo as características centrais do grupo. Além disso, deve-se ressaltar que o mesmo somente é empregado para casos em que se pode calcular a média.

A ideia central do K-means é apresentada no pseudocódigo 1.

Como visto, o algoritmo depende de um parâmetro (k = número de grupos) definido pelo usuário, comumente gerando um pro-

Algoritmo 1: Pseudocódigo do k-means (adaptado de [8])

Entrada: Conjunto de dados e o número k desejado de grupos
Saída: Partição com k grupos

1 Inicializar aleatoriamente os centros dos k grupos;
2 Associar cada ponto ao grupo mais próximo;
3 Recalcular o centro de cada grupo;
4 Repetir os passos 2 e 3 até que não haja mais alterações entre os grupos;

blema em escolher um número k ideal para o agrupamento quando, normalmente não se conhece o conjunto de dados. A escolha de um número muito pequeno, por exemplo, poderia ocasionar a junção de dois grupos naturais, afetando diretamente na qualidade dos resultados. O critério de parada é definido quando não se há mais alterações no agrupamento, ou seja, a solução converge para uma determinada partição, ou também pode ser definido através de um número máximo de iterações realizadas.

O **Agrupamento Hierárquico** possui algoritmos que podem ser divisivos ou aglomerativos [2]. Neste último, os objetos são dispersos em grupos unitários (grupo de apenas um objeto) e a cada iteração, os pares de grupos mais próximos são unidos até se formar um único grupo. Ao contrário, o agrupamento hierárquico divisivo inicialmente contém todos os objetos em um único grupo, realizando-se as divisões até que cada objeto seja um grupo unitário.

Os métodos aglomerativos e divisivos plotam os resultados obtidos em uma árvore binária, conhecida como dendrograma, sendo esta uma forma de visualização e descrição da sequência do agrupamento. No dendrograma, cada nó representa um grupo de objetos, e a altura dos arcos que unem dois subgrupos indica o grau de compactação do grupo formado, quanto menor a altura, mais compactos são os grupos [8].

O algoritmo **UPGMA (Unweighted Pair Group Method with Arithmetic Mean)**[21, 2] é um exemplo de implementação do método de agrupamento aglomerativo, realizando a organização hierárquica dos dados em um dendrograma que reflete a estrutura apresentada em uma matriz de similaridade. O pseudosódigo para o UPGMA é apresentado pelo Algoritmo 2.

Um exemplo de agrupamento hierárquico divisivo é o algoritmo ***Bisecting k-Means*** [22, 4], que utiliza o agrupamento particional baseado no *k-means*. O pseudocódigo do *Bisecting K-means* é apresentado pelo Algoritmo 3.

O passo 2 deste algoritmo, possui diferentes tipos de abordagens, de forma simples e eficaz, elege-se o maior grupo (de acordo com o número de objetos) ainda não dividido em uma iteração anterior.

A complexidade de tempo e espaço dos algoritmos de agrupamento aglomerativos e divisivos é quadrática em relação ao número de objetos. Avaliações experimentais indicam o *Bisecting K-means* como melhor opção quanto a seus resultados, seguido do algoritmo UPGMA para agrupamento hierárquico aglomerativo [2].

2.3 Pós-Processamento dos Dados

Para a avaliação dos resultados, é realizada uma **análise qualitativa** dos padrões extraídos [9]. Neste momento, é considerado como parâmetro, o nível em que o especialista influencia os dados extraídos e a forma de representação dos mesmos, distinguindo entre o real conhecimento extraído e o de pertinência do especialista. Em outras palavras,

Algoritmo 2: Pseudocódigo do UPGMA (adaptado de [8])

Entrada: Conjunto de objetos S = (o_1, o_2, ..., o_n)
Saída: Lista de agrupamentos formados

1 Calcular matriz de (dis)similaridade;
2 Cada objeto o é um grupo g;
3 Enquanto existirem grupos, selecione o par de grupo mais similar;
4 Combinar os grupos selecionados em um único grupo;
5 Atualizar a matriz de (dis)similaridade, recalculando a similaridade do novo grupo formado e dos outros grupos;

Algoritmo 3: Pseudocódigo do Bisecting k-Means (adaptado de [8])

Entrada: Conjunto de objetos S = (o_1, o_2, ..., o_n)
Saída: Lista de agrupamentos formados

1 Inicializar com um grupo contendo todos os objetos de S;
2 Selecionar o próximo grupo (nó folha) a ser dividido;
3 Dividir este grupo em dois subgrupos (*K-means* com k=2);
4 Repetir os passos 2 e 3 até obter apenas grupos unitários;

busca-se analisar o quanto o padrão extraído está coerente em relação ao conhecimento do especialista. Por fim, também é importante analisar a interessabilidade do padrão extraído levando em conta a utilidade para o usuário [9].

Algumas métricas para **análise quantitativa** também são consideradas nesta etapa [3]. Por exemplo, para avaliar quantitativamente as regras de associação, é possível utilizar o valor de confiança da regra, bem como a facilidade de interpretação sendo mensurada através do número de regras encontradas (suporte) ou número de condições por regras encontradas e outras similares.

No caso de modelos de agrupamento, na existência de um conjunto de dados de referência (conjuntos de *benchmarking*) é possível utilizar critérios objetivos para avaliar a qualidade dos grupos. Um dos critérios mais conhecido é o F_{SCORE} [2][3], que calcula a precisão e a revocação do modelo de agrupamento baseado em um conjunto verdade. Para tal, considere:

- H um modelo de agrupamento (dendrograma);
- G_i um determinado grupo e seus respectivos objetos; e
- L_r uma determinada categoria (informação externa) de referência.

Considerando uma categoria L_r e um grupo G_i, o valor de precisão é calculado como $P(L_r, G_i) = \frac{|L_r \cap G_i|}{|G_i|}$ e a revocação como $R(L_r, G_i) = \frac{|L_r \cap G_i|}{|L_i|}$. A medida F é uma média harmônica entre precisão e revocação e é obtida através de $F(L_r, G_i) = \frac{2 * P(L_r, G_i) * R(L_r, G_i)}{P(L_r, G_i) + R(L_r, G_i)}$.

O cálculo da F_{SCORE} é baseada na medida F selecionada para uma determinada categoria L_r usando o maior valor obtido por algum grupo, conforme a Equação 5.

$$F(L_r) = \max_{G_i \in H} F(L_r, G_i) \tag{5}$$

Por fim, o valor F_{SCORE} global do modelo agrupamento envolvendo n objetos e com c categorias de referencia, é calculado conforme a Equação 6.

$$F_{SCORE} = \sum_{r=1}^{c} \frac{|L_r|}{n} F(L_r) \tag{6}$$

[3] O índice F_{SCORE} também é chamado de F-Measure por alguns autores.

Quando o modelo de agrupamento obtém alta concordância com o conjunto verdade, então o valor de F_{SCORE} se aproxima de 1, caso contrário se aproxima de 0.

3. A abordagem EHC (Edge Hierarchical Clustering)

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 [15].

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 [13]. 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 [14].

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 [14]. 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 uma alternativa competitiva para agrupamento em redes de associação.

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 2 é ilustrado um esquema geral da

abordagem e suas etapas (identificadas pelas setas). Os detalhes de cada etapa são apresentados a seguir:

- **Etapa 1. Extração de Regras de Associação:** A partir de um conjunto de dados estruturados por uma tabela atributo-valor, a extração de regras de associação é realizada usando o algoritmo Apriori [17]. 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 7, em que $\vec{o}_i$ é um objeto da tabela atributo-valor que pertence ao conjunto de m objetos cobertos pela regra.

$$\vec{c}_{A \Rightarrow B} = \frac{1}{m} \sum_{i=1}^{m} \vec{o}_i \qquad \forall \vec{o}_i \in \{A \Rightarrow B\} \qquad (7)$$

  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.

4. Avaliação Experimental

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. Os conjuntos de dados podem ser obtidos no *website* do *Machine Learning Repository* da Universidade da Califórnia (`http://archive.ics.uci.edu/ml/index.html`).

- **Iris** - Contém ao todo 150 exemplos, descrevendo três classes de flores de 50 objetos cada: iris-setosa, iris-versicolor e iris-virginica. Cada objeto desta base possui quatro atributos numéricos contínuos (comprimento da sépala, largura da sépala, comprimento da pétala, e largura da pétala) e um atributo nomi-

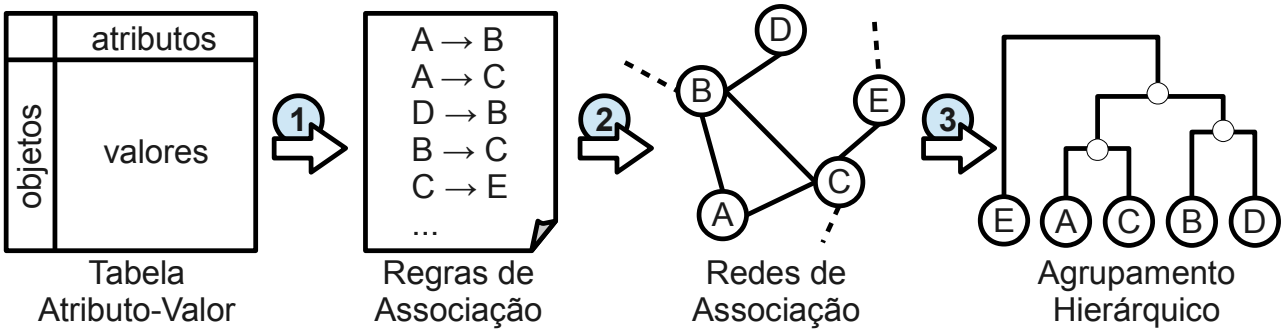


Figura 2. Esquema geral da abordagem EHC (*Edge Hierarchical Clustering*).

nal, o atributo *class*, responsável pela classificação do objeto.

- **Ecoli** - Esta base contém dados sobre a localização de proteínas em células, totalizando 336 exemplos. Consiste de sete atributos numéricos contínuos de medição de características das células, e um atributo nominal classificador de 8 classes distintas (cp - citoplasma; im - membrana interna sem sinal de sequência; pp - periplasma; imU - membrana interna com sinal de sequência não clivável; om - membrana externa; omL - membrana externa de lipoproteína; imL - membrana interna de lipoproteína; e imS - membrana interna com sinal de sequência clivável).

- **Dermatology** - Visa determinar o tipo de doença erimato-esquimatosa a partir de achados clínicos. Compreende 366 exemplos distribuídos em 6 classes distintas determinadas por um atributo nominal classificador, que baseia-se em 33 atributos sendo um numérico (*age*) e os demais categóricos.

- **DSS, IRN e TCE** - São três conjuntos de dados textuais compostos por artigos científicos coletados da Biblioteca Digital da ACM[4]. O primeiro (DSS) é formado por artigos de conferências sobre Sistemas Distribuídos e Simuladores, contendo 469 exemplos e 1964 atributos. O segundo (IRN) é formado por artigo sobre Recuperação de Informação e Redes, com 394 exemplos 1608 atributos. Por fim, o terceiro conjunto de dados (TCE) é formato por artigos sobre Teoria da Computação e Educação, contendo 471 exemplos e 1977 atributos. Todos os três conjuntos textuais são organizados em 5 classes de referências.

O índice F_{SCORE} [2], 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). Na Figura 3 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`.

[4]ACM Digital Liberary: `http://dl.acm.org/`

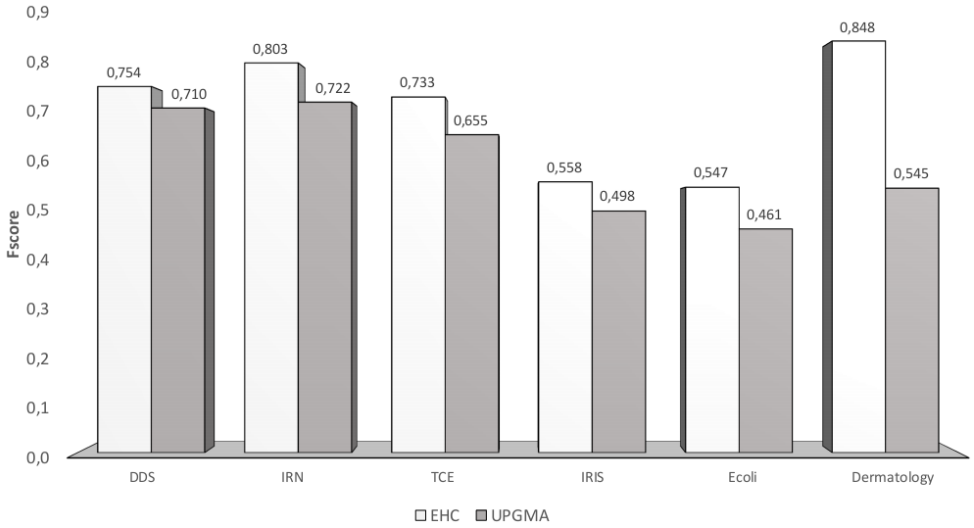
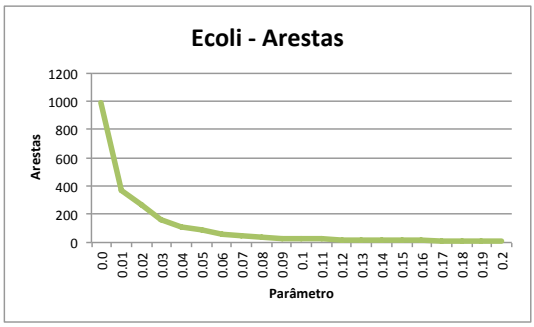
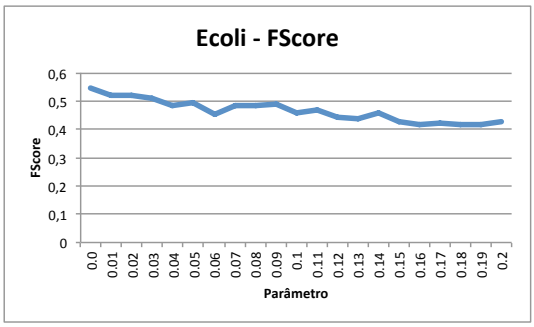


Figura 3. Comparação entre o algoritmo EHC e UPGMA para agrupamento hierárquico em redes de associação.

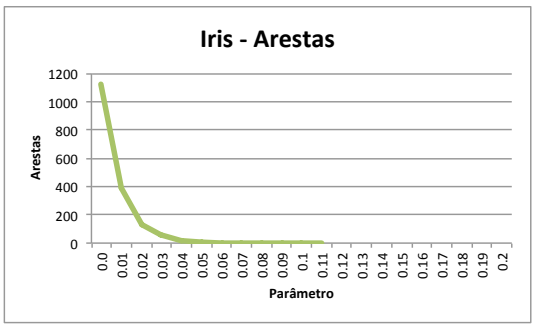


(a) Número de arestas formadas pela aplicação do algoritmo EHC sobre a base de dados Ecoli.

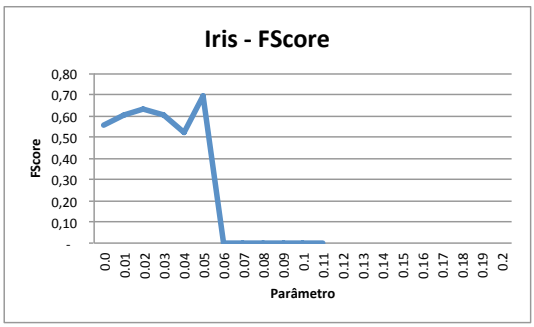


(b) Índice FScore sobre a base de dados Ecoli.

Figura 4. Análise de parâmetros para o conjunto de dados Ecoli.

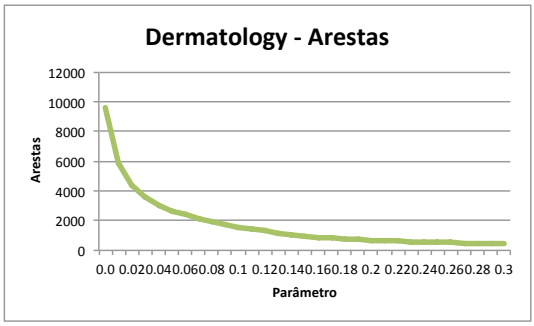


(a) Número de arestas formadas pela aplicação do algoritmo EHC sobre a base de dados Iris.

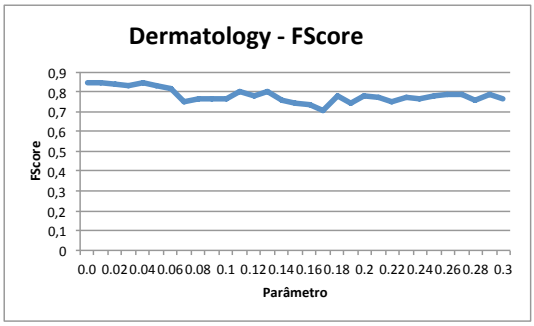


(b) Índice FScore sobre a base de dados Iris.

Figura 5. Análise de parâmetros para o conjunto de dados Iris.

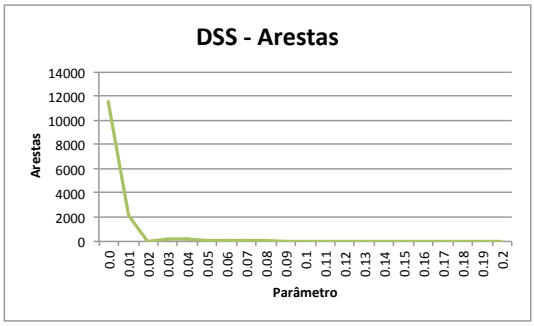


(a) Número de arestas formadas pela aplicação do algoritmo EHC sobre a base de dados Dermatology.

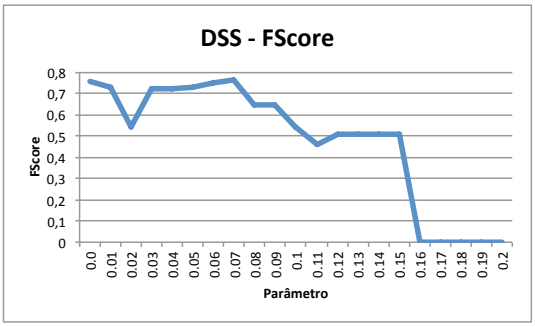


(b) Índice FScore sobre a base de dados Dermatology.

Figura 6. Análise de parâmetros para o conjunto de dados Dermatology.

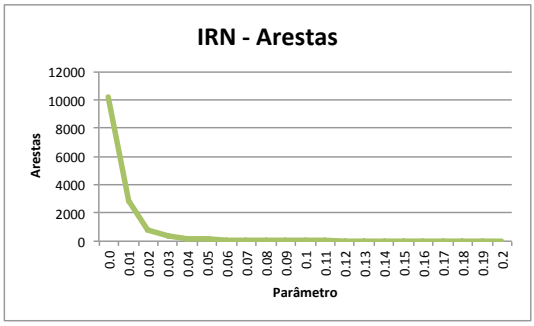


(a) Número de arestas formadas pela aplicação do algoritmo EHC sobre a base textual DSS.

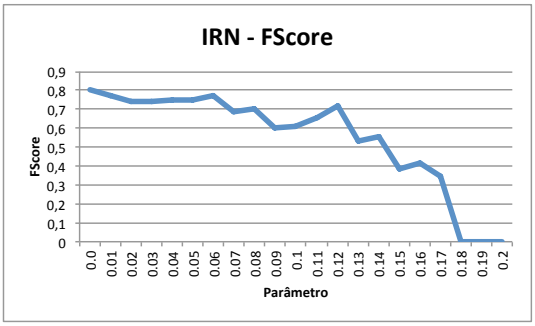


(b) Índice FScore sobre a base textual DSS.

Figura 7. Análise de parâmetros para o conjunto textual DSS.

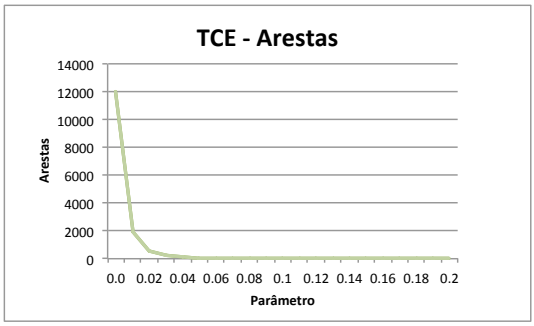


(a) Número de arestas formadas pela aplicação do algoritmo EHC sobre a base textual IRN.

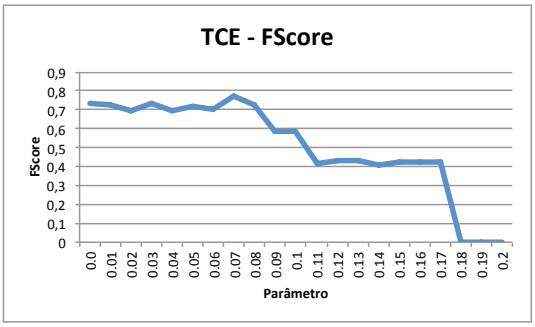


(b) Índice FScore sobre a base textual IRN.

Figura 8. Análise de parâmetros para o conjunto textual IRN.



(a) Número de arestas formadas pela aplicação do algoritmo EHC sobre a base textual TCE.



(b) Índice FScore sobre a base textual TCE.

Figura 9. Análise de parâmetros para o conjunto textual TCE.

5. Prova de Conceito: Mineração de Dados do SISU

A Mineração de Dados tem sido aplicada em diversas áreas do conhecimento, como por exemplo no setor de vendas, bioinformática e em grandes organizações do mercado financeiro. Recentemente com o advento da tecnologia e da grande procura e expansão de cursos a distância, pesquisadores têm demonstrado grande interesse na área de Inteligência Artificial aplicada à Educação, com a pretensão de responder questões científicas da área (Quais motivos levam a evasão escolar?, Quais fatores determinam a aprovação do aluno em dada disciplina?). Dentro deste contexto surgiu uma nova área de pesquisa conhecida como Mineração de Dados Educacionais (*Educational Data Mining*, EDM) que tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais [23].

Mesmo sendo uma área relativamente nova, o volume de trabalhos na área tem sido expressivo, tanto que levou a criação da *International Educational Data Mining Society*[5].

Exemplos de aplicação da EDM podem ser evidenciados no estudo de Estrada et al. [24] onde aplicou-se o processo de mineração de dados na base de dados acadêmica dos alunos de graduação do curso de Ciência da Computação da Universidade de Sinaloa, no México, com o objetivo de detectar fatores que determinam o sucesso acadêmico dos alunos no curso, variáveis que, por exemplo, podem ser utilizadas por programas sociais que visem a diminuição da evasão no curso.

A EDM tem em sua aplicabilidade uma extensão maior que a de dados transacionais utilizados e implementados em banco de dados. A análise dessas informações, por exemplo, pode contribuir positivamente para a administração institucional quanto à implantação de projetos sociais de manutenção de seus alunos e melhoria da educação com a criação de projetos específicos para atender uma determinada área.

Para apoiar este tipo de tarefa e realizar uma prova de conceito da abordagem *EHC*, neste trabalho foi desenvolvido um ambiente para análise exploratória de dados do SISU, sendo este, o sistema informatizado no qual as instituições públicas de ensino superior oferecem vagas para candidatos participantes do Exame Nacional do Ensino Médio (Enem), e portanto o principal meio de acesso ao ensino superior do país, demonstrando sua alta relevância para estudos sobre a temática. A análise exploratória tem como

[5] *International Educational Data Mining Society*: http://educationaldatamining.org

base o questionário socioeconônomico dos alunos ingressantes da UFMS, ano base de 2012. Este banco de dados possui 34.695 alunos e 69 de um total de 74 questões socioeconômicas, sendo 5 delas removidas na etapa de pré-processamento, evidenciadas pela Figura 10 que demonstra os atributos removidos por não oferecerem ganho de informação no processo de MD já que os mesmos indicam o mesmo valor de característica para todos os exemplos da base de dados.

A lista com os atributos é demonstrada pela Figura 11 que contempla a cardinalidade de cada atributo, ou seja, o número de alternativas à uma dada pergunta do questionário socioeconômico, como por exemplo, o atributo "*sexo*" de cardinalidade 2 ("Masculino" ou "Feminino"); a Figura 11 apresenta também o grau de abstinência das respostas dadas pelos alunos em cada pergunta, como por exemplo, o atributo "*opcao*" com 0% de abstinência, em contrapartida o atributo "*escola_regular*" apresenta 97% de abstinência, significando que 33.607 alunos deixaram de responder à questão quanto a este quesito. Devido ao grande volume de dados, uma análise manual destes exigiria um intenso esforço humano, oferecendo uma oportunidade para a abordagem investigada neste trabalho.

Na Figura 12 é demonstrado uma parte do modelo de rede de associação formado a partir da implementação do método *EHC* sobre a base de dados do SISU. A rede obtida, com suporte igual a 0.3, apresenta 2550 regras de associação e portanto há uma dificuldade em plotar visualmente a rede devido as inúmeras associações entre os atributos conforme exemplificado pelo modelo apresentado. Para ilustrar um possível cenário para análise da rede de associação formada, neste exemplo pode-se observar que atributos como "$nu\_idade\_comecou\_trabalhar = 13$" e "$curso\_vestibular = 0$" demonstram maior concentração na interação entre as regras, por outro lado atributos como "$custeio\_prouni = 1$" e "$foi\_matriculado = Nao$" indicam poucas interações com as demais regras. A concepção da rede de associação, em muitos casos, pode revelar ao especialista de domínio outras diferentes relações entre os itens do conjunto de dados. A validação desse cenário, no entanto, depende de critérios qualitativos, considerando a análise do especialista.

Já na Figura 13 é ilustrada a tela principal do ambiente desenvolvido para exploração visual dos resultados. Na parte *A* é apresentado um acesso à opção "Filtro", demonstrado pela Figura 14, que permite aos utilizadores configurarem suas próprias expressões de busca, onde para cada pergunta do questionário socioeconômico, de forma cumulativa, pode ser incluída possíveis respostas pertinentes à base de dados para a formação da expressão de busca. Na parte *B* é apresentada um subconjunto de dados resultante de uma determinada consulta realizada pelo explorador. Na parte *C* é ilustrado um gráfico que sumariza uma determinada questão, considerando o respectivo subconjunto de dados. Por fim, na parte *D* é apresentado o modelo de agrupamento hierárquico do *EHC*. Neste caso, foram selecionadas as arestas mais relevantes para representar cada grupo, ou seja, as arestas mais próximas ao centroide de cada grupo. Ao clicar em um grupo, o explorador utiliza aquela regra para refinar a expressão de busca, ou seja, o modelo de agrupamento está atuando como uma interface de recomendação de consultas. É importante observar que o usuário também tem acesso a uma interface tradicional para filtrar (consultar) o conjunto de dados, conforme ilustrado na Figura 14. Nesse caso, o usuário deve in-

Atributos	Pré-processamento	Observações
acao_afirmativa	Removido	Todas respostas "Ampla_Concorrencia"
concorreu_acao_afirmativa	Removido	Todas respostas "NÃO"
percentual_bonus	Removido	Todas respostas "0"
candidato_aprovado	Removido	Todas respostas "SIM"
utilizou_bonus	Removido	Todas respostas "NÃO"

Figura 10. Lista de atributos removidos da base de dados na etapa de pré-processamento

Atributos	Cardinalidade	Missing
in_casa_videiocassete_dvd	4	9904 (29%)
escola_ensino_medio	9	9904 (29%)
certificado_continua_estudo	6	33607 (97%)
tp_motivo_adquiri_experiencia	6	22920 (66%)
sexo	2	0 (0%)
part_bolsa_estudo	6	9904 (29%)
in_casa_freezer	4	9904 (29%)
tp_motivo_ser_independente	6	22920 (66%)
in_casa_telefone_celular	4	9904 (29%)
escola_regular	2	33607 (97%)
custeio_pro_uni	2	9904 (29%)
impedido_ensino_medio	5	9904 (29%)
custeio_bolsa_trabalho	2	9904 (29%)
in_trabalhou_durante_estudo	2	9904 (29%)
sua_renda	11	9904 (29%)
curso_vestibular	2	22920 (66%)
abandono_trabalho	6	33607 (97%)
tp_casa_localizada	4	9904 (29%)
in_casa_tv_assinatura	4	9904 (29%)
curso_outros	2	22920 (66%)
anos_ensino_fundamental	7	9904 (29%)
abandono_escola_perto	6	33607 (97%)
in_casa_maquina_lavar	4	9904 (29%)
in_casa_empregada_mensalista	4	9904 (29%)
foi_matriculado	2	0 (0%)
abandono_saude	6	33607 (97%)
in_casa_tv	4	9904 (29%)
escolaridade_pai	9	9904 (29%)
opcao	2	0 (0%)
nu_idade_comecou_trabalhar	13	22920 (66%)
in_casa_radio	4	9904 (29%)
tp_hora_trabalho	5	22920 (66%)
in_casa_banheiro	4	9904 (29%)
chamada	3	0 (0%)

Atributos	Cardinalidade	Missing
abandono_familiar	6	33607 (97%)
in_casa_telefone_fixo	4	9904 (29%)
testar_conhecimento	6	9904 (29%)
lista_espera	2	0 (0%)
prosseguir_estudo	6	9904 (29%)
certificado_emprego_melhor	6	33607 (97%)
renda_familiar	11	9904 (29%)
curso_superior	2	22920 (66%)
in_casa_microcomputador	4	9904 (29%)
anos_ensino_medio	6	9904 (29%)
abandono_pessoais	6	33607 (97%)
custeio_bolsa_instituicao	2	9904 (29%)
abandono_conseguir_vagas	6	33607 (97%)
num_pessoas	18	9904 (29%)
in_casa_acesso_internet	4	9904 (29%)
curso_informatica	2	22920 (66%)
tp_motivo_pagar_estudo	6	22920 (66%)
part_certificado	6	9904 (29%)
tp_motivo_ajudar_pais	6	22920 (66%)
tp_casa_onde_mora	4	9904 (29%)
certificado_emprego	6	33607 (97%)
in_casa_automovel	4	9904 (29%)
curso_concurso	2	22920 (66%)
uf	27	0 (0%)
idade	56	0 (0%)
in_preparatorio_trabalho	2	22920 (66%)
escolaridade_mae	9	9904 (29%)
tp_escola_fundamental	9	9904 (29%)
in_casa_geladeira	4	9904 (29%)
municipio	1602	0 (0%)
custeio_fies	2	9904 (29%)
anos_deixou_frequentar	7	33607 (97%)
in_curso_profissionalizante	2	22920 (66%)
tp_motivo_sustentar_familia	6	22920 (66%)
curso_lingua_estrangeira	2	22920 (66%)

Figura 11. Lista de atributos da base de dados do SISU após a tarefa de pré-processamento.

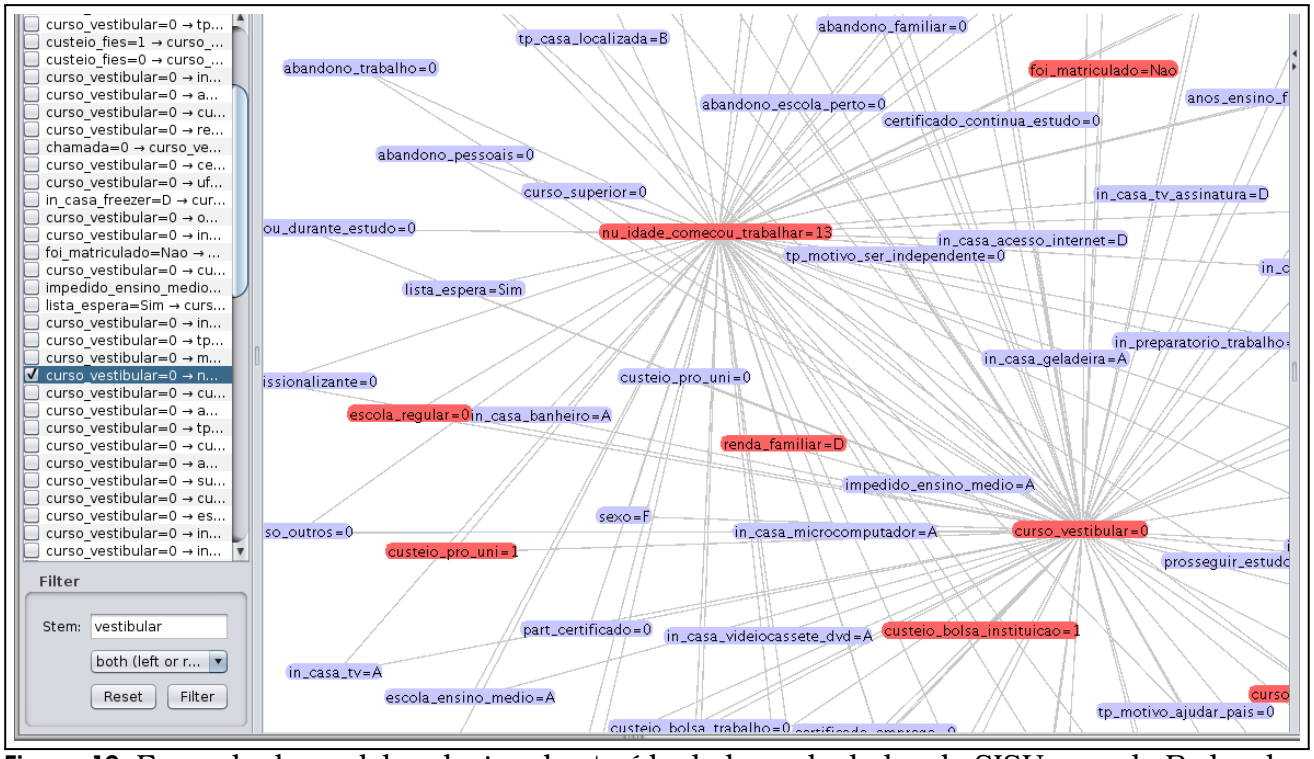


Figura 12. Exemplo de modelo relacional extraído da base de dados do SISU usando Redes de Associação.

formar manualmente seus atributos de interesse e respectivos valores, o que geralmente é uma tarefa difícil em análise exploratória de dados.

Os resultados preliminares obtidos com o explorador indicam que o *EHC* é promissor para tarefas de análise exploratória. No entanto, ainda é necessário uma análise do ambiente com especialistas de domínio da UFMS para analisar a qualidade do modelo de agrupamento obtido nesta aplicação.

6. Considerações Finais

A representação de dados por meio de redes de associação é uma alternativa promissora para apoiar muitas tarefas de extração de padrões. Enquanto modelos proposicionais permitem indicar, de forma explícita, a importância dos atributos em cada objeto, as redes de associação permitem capturar a noção de relação entre os atributos. Neste trabalho foi proposta e avaliada a abordagem *EHC* (*Edge Hierarchical Clustering*), que explora tanto as relações entre os atributos apresentadas no modelo relacional, quanto os valores de relevância desses atributos no modelo proposicional. Assim, é possível obter uma noção mais robusta da similaridade entre os objetos, permitindo a indução de modelos de agrupamento mais robustos.

A análise experimental realizada neste trabalho fornece evidências de que a abordagem *EHC* obtém estruturas de agrupamento mais significativas, quando comparado com agrupamento baseado apenas na informação de vizinhança dos vértices. Essa análise experimental foi realizada tanto em conjuntos de dados numéricos quanto conjuntos de dados textuais. Uma limitação existente na abordagem proposta é a necessidade de definir o parâmetro relacionado à construção da rede de associação. Um valor pequeno para este parâmetro induz a formação de redes de associação com um grande número de arestas, o que dificulta a análise visual do conjunto de dados. Por outro lado, os melho-

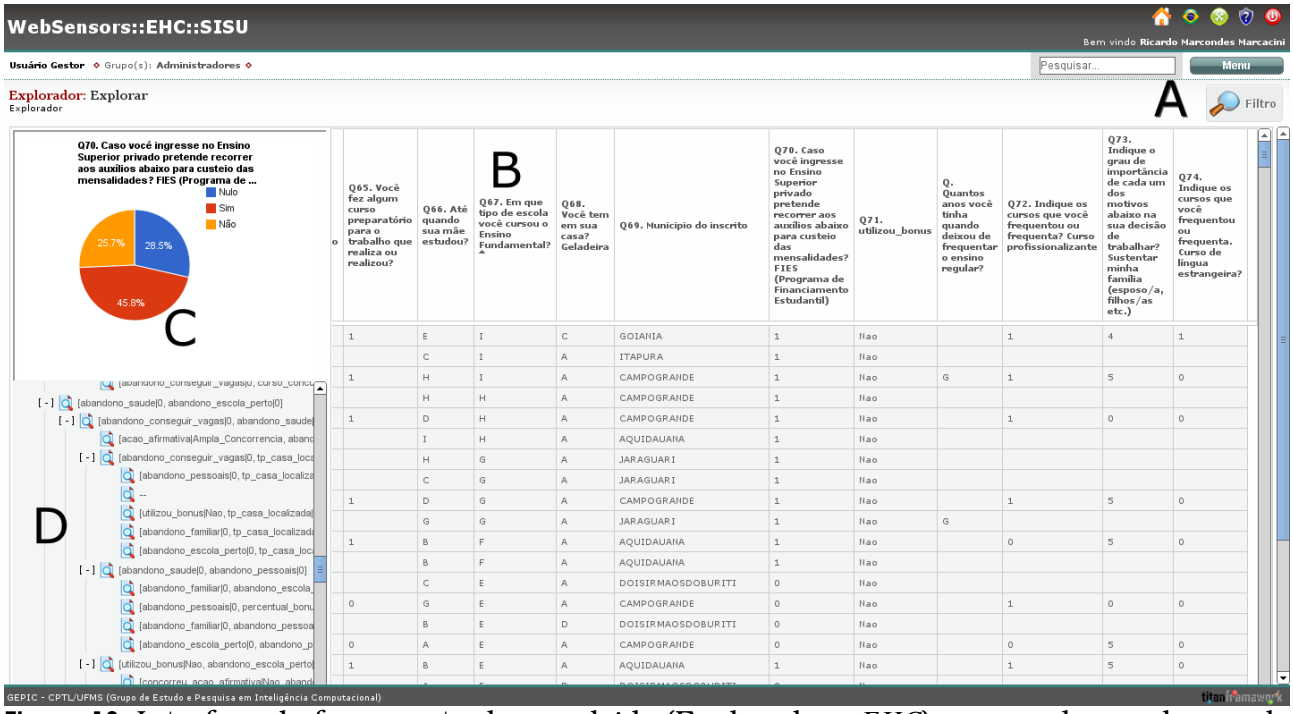


Figura 13. Interface da ferramenta desenvolvida (Explorador - *EHC*) para explorar o banco de dados com apoio de um modelo de agrupamento induzido pelo *EHC*.

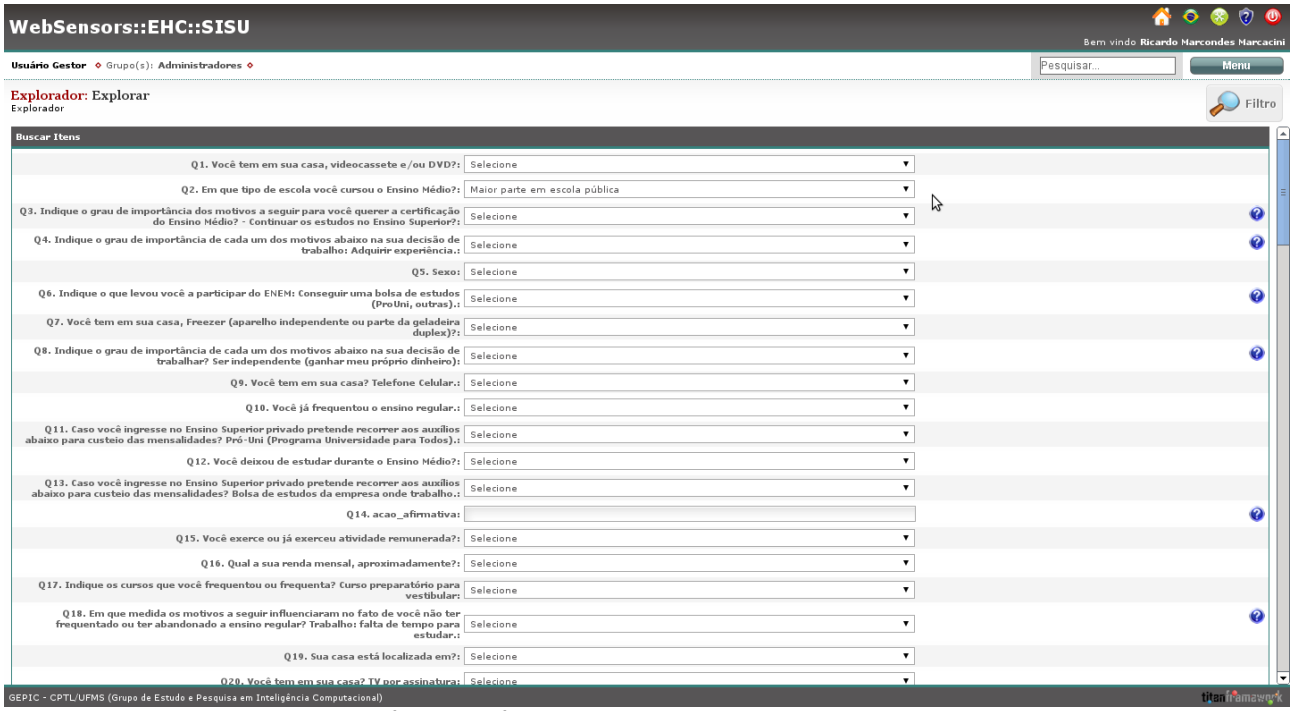


Figura 14. Interface da ferramenta desenvolvida (Explorador - *EHC* com um módulo tradicional para filtrar (consultar) o conjunto de dados.

res modelos de agrupamento foram obtidos quando as redes de associação continham um grande número de arestas.

Neste trabalho, o conjunto de dados do SISU também foi explorado em uma prova de conceito da abordagem *EHC* para um problema do mundo real. Os modelos de agrupamento obtidos foram utilizados para apoiar a construção de consultas no conjunto de dados, de forma a selecionar subconjuntos de dados que possam ser potencialmente interessantes. Nesse sentido, as direções para trabalhos futuros envolvem avaliar a abordagem *EHC* junto com especialistas de domínios, bem como o desenvolvimento de ferramentas que permitam uma análise visual mais eficaz das redes de associação.

Referências

[1] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadow s, and biggest grow th in the far east. Technical Report 1414, IDC iView, 2012. http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

[2] Charu C. Aggarwal and Chandan K. Reddy. *Data Clustering: Algorithms and Applications*. CRC Press, 2013. http://dl.acm.org/citation.cfm?id=2535015.

[3] Daniel T. Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014. http://dx.doi.org/10.1002/9781118874059.

[4] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[5] Sanjay Chawla. Feature selection, association rules network and theory building. In *The Fourth Workshop on Feature Selection in Data Mining*, 2010. http://jmlr.org/proceedings/papers/v10/chawla10a/chawla10a.pdf.

[6] Gaurav Pandey, Sanjay Chawla, Simon Poon, Bavani Arunasalam, and Joseph G. Davis. Association rules network: Definition and applications. *Statistical Analysis and Data Mining*, 1(4):260–279, 2009. http://dx.doi.org/10.1002/sam.10027.

[7] Melissa Rodrigues, João Gama, and Carlos Abreu Ferreira. Identifying relationships in transactional data. In *Advances in Artificial Intelligence – IBERAMIA 2012*, volume 7637 of *Lecture Notes in Computer Science*, pages 81–90. Springer Berlin Heidelberg, 2012. http://dx.doi.org/10.1007/978-3-642-34654-5_9.

[8] Ricardo M. Marcacini. Unsupervised learning of topic hierarchies from dynamic text collections. Master's thesis, Mathematical and Computer Sciences Institute. University of São Paulo, Brazil, (In portuguese), 2011. http://www.teses.usp.br/teses/disponiveis/55/55134/tde-28072011-163026/en.php.

[9] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of ACM*, 39(11):27–34, 1996. http://doi.acm.org/10.1145/240455.240464.

[10] Daniel T. Larose and Chantal D. Larose. *Data Preprocessing*, chapter 2, pages 16–50. John Wiley & Sons, 2014. http://dx.doi.org/10.1002/9781118874059.ch2.

[11] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull*, 23(4):3–13, 2000. http://www.ime.unicamp.br/~wanderson/Artigos/data_cleaning_approaches.pdf.

[12] Alon Halevy, Anand Rajaraman, and Joann Ordille. Data integration: The teenage years. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, VLDB '06, pages 9–16. VLDB Endowment, 2006. http://dl.acm.org/citation.cfm?id=1182635.1164130.

[13] Anna Goldenberg, Alice X. Zheng, and Stephen E. Fienberg. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, February 2010.

[14] Fedja Hadzic, Tharam S. Dillon, and Henry Tan. *Mining of Data with Complex Structures*, chapter Graph Mining, pages 287–298. Springer, 2011.

[15] Eric D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 1st edition, 2009.

[16] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. http://www.jstor.org/stable/25054401.

[17] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *20th International Conference on Very Large Data Bases (VLDB)*, pages 487–499, 1994. http://dl.acm.org/citation.cfm?id=645920.672836.

[18] Chengqi Zhang and Shichao Zhang. *Association Rule Mining: Models and Algorithms*. Springer, 2002. http://dl.acm.org/citation.cfm?id=1791549.

[19] Xindong Wu and Vipin Kumar. *The Top Ten Algorithms in Data Mining*. Chapman & Hall/CRC, 1st edition, 2009. http://dl.acm.org/citation.cfm?id=1554380.

[20] Daniel T. Larose and Chantal D. Larose. *Hierarchical and k-Means Clustering*, chapter 10, pages 209–227. John Wiley & Sons, 2014. http://dx.doi.org/10.1002/9781118874059.ch10.

[21] P. H. A SNEATH and R. R. SOKAL. Unweighted pair group method with arithmetic mean. In *Numerical Taxonomy: The principles and practice of numerical classification*, pages 230–234, 1973.

[22] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *Knowledge Discovery on Databases. Workshop on Text Mining*, pages 525–526, 2000. http://www-2.cs.cmu.edu/~dunja/KDDpapers/Steinbach_IR.pdf.

[23] Huseyin Guruler, Ayhan Istanbullu, and Mehmet Karahasan. A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education*, 55(1):247 – 254, 2010. http://dx.doi.org/10.1016/j.compedu.2010.01.010.

[24] R. Estrada, A. Zaldivar, L. Nava, J. Peraza, R. Mendoza, N. Zaragoza, E. Diaz, and C. Aguilar. Identification of variables associated with academic success of higher education students applying data mining. In *International Conference on Education and New Learning Technologies*, pages 4958–4963. IATED, 2011. http://library.iated.org/view/ESTRADA2011IDE.