Capítulo 1

Descoberta Automática de Conhecimento por meio de Redes de Regras de Associação Filtradas

Matheus William Gomes dos Santos, Andreiver Mateus Ferreira Silva e Dario Brito Calçada

Abstract

Association Rules Mining is an excellent technique for finding patterns between elements within a data set. There is a considerable variation in applications of this knowledge extraction technique in several areas, ranging from direct applications in the scientific field or even in the financial market. However, the process for viewing patterns identified by using Association Rules mining should be improved, as the volume of data is very high. Within this context, the filtered Association Rules Networks (Filtered-ARNs) appear, which take into account mathematical factors to prove the influence between the elements of a rule, to facilitate the identification of related patterns of interest to a particular study.

Resumo

A Mineração de Regras de Associação é uma excelente técnica para encontrar padrões entre elementos dentro de um determinado conjunto de dados. Existe uma variação considerável de aplicações desta técnica de extração de conhecimento em várias áreas, estendendo-se desde aplicações diretas no campo científico ou até no mercado financeiro. No entanto, o processo para visualização dis padrões identificados pela uso da mineração de Regras de Associação deve ser melhorado, já que o volume de dados é muito elevado. Dentro desse contexto, aparecem as Redes de Regras de Associação Filtradas (Filtered-ARNs), que levam em consideração fatores matemáticos para comprovação da influência entre os elementos de uma regra, a fim de viabilizar a identificação de padrões de interesse relacionados a um determinado estudo.

1.1. Considerações Iniciais

A mineração de dados por meio da descoberta de Regras de Associação é uma tarefa que visa a identificação de padrões em bases de dados e, posteriormente, alcançar conhecimento acerca do tema e do problema em questão [Le and Vo 2016]. Além disso, esse tipo

de mineração pode ser utilizado para descobrir hipóteses em determinado domínio de conhecimento, possibilitando avanços em processos de pesquisa em conjunto com métodos estatísticos [Vinaya and Shah 2016].

A Mineração de Regras de Associação inicia-se a partir de observações dos eventos para formar uma estrutura na qual o processo que está gerando os eventos possa ser evidenciado. Uma Regra de Associação [Agrawal and Shafer 1996, Agrawal and Srikant. 1994, Agrawal et al. 1994] possui uma forma padrão A => B, na qual A e B podem ser atributos, itens ou "objetos de dados". A Mineração de Regras de Associação corresponde a uma considerável gama de áreas de pesquisa. Tornando-se necessário limitar o escopo de cada trabalho que tiver a sua utilização.

Neste capítulo são apresentados os conceitos fundamentais relacionados a Redes de Regras de Associação Filtradas. Na seção 2, são descritos os conceitos e a definição de Regras de Associação e sobre o seu processo de mineração. Na seção 3, são mostrados a definição e os conceitos a cerca de Redes de Regras de Associação bem como conceitos básicos sobre Redes. Na seção 4 são introduzidas as Redes de Regras de Associação Filtradas bem como as métricas utilizadas no processamento de análise dessas regras. E, por fim , na seção 5 é discutida a geração automática de hipóteses pela descoberta de padrões de interesse em *datasets*.

1.2. Mineração de Regras de Associação

Nos últimos anos, cada vez mais empresas, organizações e usuários manipulam e armazenam quantidades cada vez maiores de dados. Este fato, tornou a compreensão dos dados uma atividade de relevância considerável no suporte à tomada de decisões [Zaki and Meira 2013].

Como também descrito na seção anterior, a mineração de dados tem como definição a descoberta de padrões em bases de dados [Aggarwal 2015]. Além disso, a mineração aparece como uma alternativa na solução à dificuldade em resumir e encontrar padrões não óbvios em quantidades de dados cada vez maiores. Em relação à Mineração de Regras de Associação, pode-se dividi-la em 3 etapas:

- **Pré-processamento:** preparação da base e domínio dos dados para a etapa de extração de padrões, podendo ocorrer a remoção de itens não interessantes.
- **Extração de padrões:** são efetuados os cálculos de medidas, construção dos *itemsets* frequentes e formulação das Regras de Associação.
- **Pós-processamento:** nessa etapa ocorre a remoção de regras não interessantes, causando uma redução do número de regras a serem exploradas pelo usuário.

Em decorrência de quantidades cada vez maiores de dados, o número de Regras de Associação descobertas pode tornar-se mais um problema para a sua interpretação, gerando assim, uma nova necessidade de mineração. Portanto, é necessária a utilização de melhores maneiras de interpretar o conhecimento trazido pelas Regras de Associação, como o uso de Redes [Pandey et al. 2009].

A fim de elucidar o conceito e objetivo das Regras de Associação, pode-se citar duas definições:

- **Definição 1:** seja $I = \{i_1; i_2; ...; i_n\}$ um conjunto de objetos denominados itens que podem assumir valores binários 0 ou 1 , que representam a presença ou não de um objeto em particular. Seja T um conjunto de transações, em que cada transação D corresponde a um conjunto de itens tal que $D \subseteq I$. Considera-se ainda que um conjunto de itens A está contido numa transação D, se todos os itens do conjunto tiverem valor "verdadeiro" na transação, ou seja, fizerem parte dessa mesma transação. Uma Regra de Associação R pode ser representada por uma expressão no formato: $A \Rightarrow B$, com $A \subseteq I; B \subseteq I$ e $A \cap B = \emptyset$. É ainda possível tratar as variáveis quantitativas ou qualitativas, criando intervalos de valores e utilizando-as, posteriormente, como variáveis binárias. A é denominado de antecedente (LHS – Left Hand Side) da regra e B o consequente (RHS - Right Hand Side).

- **Definição 2:** para cada regra $(LHS \Rightarrow RHS)$, extraída de um conjunto de transações T, é calculado um valor de suporte (sup), apresentado na Equação 1, que verifica a força de associação entre LHS e RHS (probabilidade de ocorrência da transação $LHS \cup RHS$); e um valor de confiança (conf), Equação 2, que mede a força da implicação lógica da regra (probabilidade condicional de RHS dado LHS) [AGRAWAL, IMIELINSKI and SWAMI 1994].

$$sup(LHS \Rightarrow RHS) = P(LHS \cup RHS) \tag{1}$$

$$conf(LHS \Rightarrow RHS) = P(RHS|LHS) \tag{2}$$

O suporte de uma regra tem como definição a probabilidade de que uma transação qualquer satisfaça LHS e RHS simultaneamente, enquanto que a confiança é a chance de uma transação satisfazer RHS, dado que ela satisfaz LHS.

1.2.1. Extração das Regras de Associação

O problema de se descobrir todas as Regras de Associação de um problema divide-se em duas partes [de Vasconcelos and de Carvalho 2004]:

- Encontrar todos os conjuntos de itens que possuam um suporte de transações acima de um limite mínimo informado.

- Selecionar, dentre as regras geradas, apenas as que possuam o grau de confiança mínimo, correspondente à confiança mínima estabelecida.

Dado um conjunto de transações, o problema de mineração de Regras de Associação está em gerar todas as regras que contenham o suporte e confiança iguais ou maiores do que os valores mínimos determinados pelo usuário, referenciados, respectivamente, como suporte mínimo (minsup) e confiança mínima (minconf). Por exemplo, considerando a base de dados "Compra" apresentada na Tabela 1, a qual possui dados referentes

a compras diárias de um indivíduo dentre um período de 10 (dez) dias. Neste caso, define-se que minsup é 0,3 (30%) e que minconf é de 0,8 (80%).

Considerando-se que LHS = CAFÉ e RHS = PÃO, pode-se calcular o suporte e a confiança para a regra (CAFÉ $\Rightarrow$ PÃO) tendo como resultados sup(CAFÉ $\cup$ PÃO) = 0,3 ou 30% e conf(CAFÉ $\Rightarrow$ PÃO) = 1 ou 100%. Este resultado implica duas afirmações:

- "em 30% das compras (em 10 dias) desse indivíduo ele comprou café e pão"
- "sempre que esse indivíduo comprou café, ele comprou pão"

Essas afirmações podem servir como base para a elaboração de hipóteses que podem conduzir estudos futuros sobre o comportamento padrão de compras do sujeito deste caso.

1.2.1.1. Algoritmo Apriori

O *Apriori* é considerado o algoritmo mais conhecido para extração de Regras de Associação [Agrawal and Srikant. 1994]. Por ele, é possível obter todos os conjuntos de *itemsets* frequentes (L_k). Em seu funcionamento, o algoritmo gera conjuntos de itens candidatos (padrões) por meio da busca em profundidade. Posteriormente, os padrões não frequentes são eliminados. Todos os conjuntos de *itemsets* frequentes são obtidos a partir dos conjuntos de itens candidatos.

O funcionamento do algoritmo principal (Figura 1.1) pode ser explicado pelo uso de duas funções: *Apriori_gen*, para gerar os itens candidatos e eliminar os não frequentes, e a *Gen_rules* para extração das Regras de Associação.

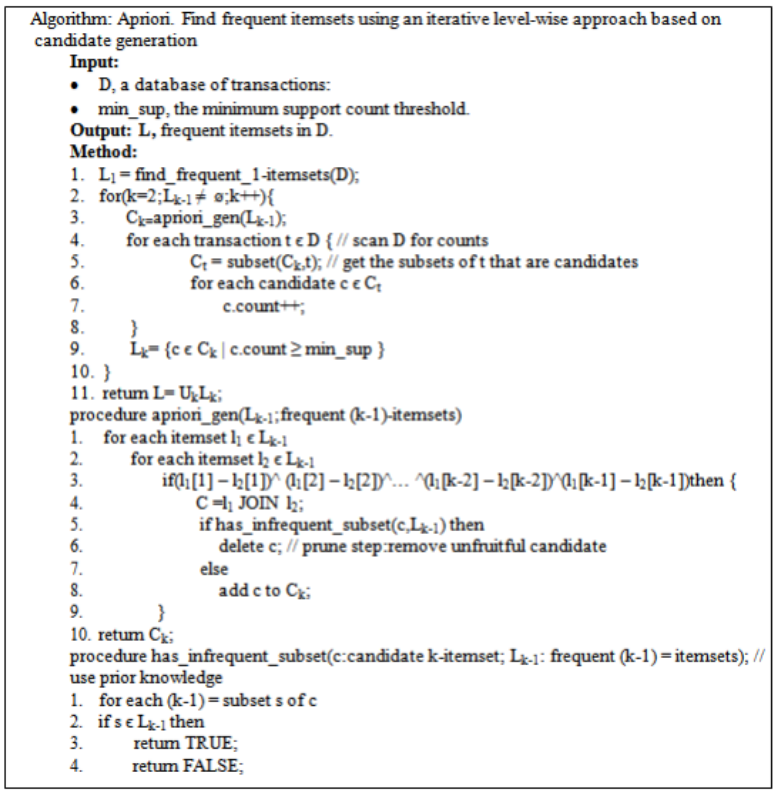
A descoberta de Regras de Associação pelo *Apriori* pode ser dividida em:

- Descoberta de todos os conjuntos de *itemsets* frequentes e;
- Geração de Regras de Associação (a partir dos conjuntos de *itemsets* frequentes descobertos).

O primeiro passo, dentre os dois citados anteriormente, é considerado o de maior custo computacional, causando maior atenção nos estudos e pesquisas em mineração de dados. Tal custo motivou o surgimento de diferentes algoritmos com objetivo de maximizar a eficiência computacional do *Apriori*.

As maiores desvantagens do *Apriori* decorrem do fato de não possuir suporte para uma quantidade elevada de dados brutos. Essas desvantagens são significativas devido ao fato de a quantidade de dados produzidas nos últimos anos ter aumentado consideravelmente. Assim, é necessário o uso de técnicas eficientes que analisem e gerenciem quantidades massivas de dados. Mesmo com essa necessidade, os algoritmos da família do Apriori (gerados a partir do original) são o mais utilizados para Mineração de Regras de Associação.

Além do *Apriori*, existem outros algoritmos com finalidade de geração de Regras de Associação. Esses algoritmos diferem-se do *Apriori* pela atenção na correção da



Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on candidate generation

Input:

- D, a database of transactions:
- min_sup, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

```
1.  L1 = find_frequent_1-itemsets(D);
2.  for(k=2;Lk-1 ≠ ø;k++){
3.      Ck=apriori_gen(Lk-1);
4.      for each transaction t ϵ D { // scan D for counts
5.              Ct = subset(Ck,t); // get the subsets of t that are candidates
6.              for each candidate c ϵ Ct
7.                  c.count++;
8.      }
9.      Lk= {c ϵ Ck | c.count ≥ min_sup }
10. }
11. return L= UkLk;
procedure apriori_gen(Lk-1;frequent (k-1)-itemsets)
1.   for each itemset l1 ϵ Lk-1
2.       for each itemset l2 ϵ Lk-1
3.           if(l1[1] – l2[1])^ (l1[2] – l2[2])^... ^(l1[k-2] – l2[k-2])^(l1[k-1] – l2[k-1])then {
4.               C =l1 JOIN l2;
5.               if has_infrequent_subset(c,Lk-1) then
6.                   delete c; // prune step:remove unfruitful candidate
7.               else
8.                   add c to Ck;
9.           }
10. return Ck;
procedure has_infrequent_subset(c:candidate k-itemset; Lk-1: frequent (k-1) = itemsets); //
use prior knowledge
1.  for each (k-1) = subset s of c
2.  if s ϵ Lk-1 then
3.       return TRUE;
4.       return FALSE;
```

Figura 1.1. Algoritmo Apriori [Han and Kamber 2006]

necessidade de percorrer os conjuntos de elementos várias vezes, consequentemente, possuem melhor desempenho computacional com o mesmo resultado. Dentre esses outros algoritmos, pode-se citar, *FP-Growth* e o *Apriori-TID*.

1.2.2. Critérios para Seleção e Classificação das Regras de Associação

Após a descoberta das Regras de Associação de determinado problema, é necessário medir o nível de interesse dos padrões descobertos [Usman and Usman 2016]. No total, nove critérios específicos são usados mensurar se determinado padrão é ou não interessante, sendo eles concisão, cobertura, confiabilidade, peculiaridade, diversidade, novidade, surpreendimento, utilidade e capacidade de ação.

Além disso, esses critérios podem ser divididos em três classificações: medidas objetivas, medidas subjetivas e medidas baseadas em semântica. Que são definidas da seguinte forma:

- **Medidas objetivas:** baseiam-se apenas nos dados originais. Nenhum conhecimento sobre o usuário ou aplicativo é necessário. Em geral, baseiam-se em probabilidade ou teoria da informação. Algumas dessas medidas consideram que uma Regra de Associação é interessante apenas quando o valor de suporte dessa regra é maior que o valor de suporte esperado. Outra função das medidas objetivas é demonstrar como os itens influenciam uns aos outros, podendo esta influência ser de modo direto, quando os itens variam de modo proporcional, ou inverso, quando o

aumento da incidência de um item leva à diminuição de outro.

- **Medidas de interesse subjetivas:** Consideram principalmente a opinião de um analista para determinar a força da regra [Gonçalves 2005].
- **Medidas baseadas em semântica:** Utilizam as estruturas semânticas das regras para estabelecimento de sua importância [Han et al. 2012].

Considerando esses critérios, pode-se utilizar três métodos para determinar o interesse de um padrão. Primeiro, pode-se classificar cada padrão em interessante ou não. Em segundo lugar, pode-se determinar uma relação de preferência para representar padrões mais interessantes que outros. E por fim, pode-se também elencar os tipos de padrões. Assim, o uso de medidas de interesse facilita a identificação automática de padrões interessantes em Regras de Associação.

Além das medidas supracitadas de suporte e confiança, outras medidas para as regras podem ser calculadas. A seguir segue a definição de duas medidas assimétricas, *Added Value* e *Gain*, que são utilizadas nas Redes de Regras de Associação Filtradas, que serão explicitadas nas Seções seguintes.

- ***Added Value* [-1..0..1]:** a medida *Added Value* (AV) (Equação 3) indica o quanto a frequência do consequente aumenta na presença do antecedente, ou seja, mede o ganho de *RHS* na presença de *LHS* [Sahar 2003]. Se AV for positivo, então a frequência de RHS aumenta na presença de LHS. Sendo AV negativo, a frequência de RHS diminui na presença de LHS. Se AV possuir valor nulo (zero) , tem-se uma coincidência aleatória, ou seja, a frequência de LHS não altera em nada a frequência de RHS.

$$AV = Conf(LHS \Rightarrow RHS) - sup(RHS) \tag{3}$$

- ***Gain* [0..1]:** é uma medida proposta por Fukuda et al. (1996) que dá um trade-off entre suporte e confiança (Equação 4), auxiliando na seleção das regras de acordo com a frequências da mesma em relação à confiança mínima.

$$Gain = sup(LHS \cap RHS) - minconf.sup(LHS) \tag{4}$$

Considerando-se a Equação 2, pode-se inferir que:

$$sup(LHS \cap RHS) = sup(LHS).conf(LHS \Rightarrow RHS) \tag{5}$$

portanto, fazendo a substituição na Equação 4, e colocando-se em evidência o fator sup(LHS) obtém-se a Equação 6.

$$Gain = [conf(LHS \Rightarrow RHS) - minconf].sup(LHS) \tag{6}$$

Por meio da Equação 6, percebe-se que a medida objetiva *Gain* funciona como uma normalização da medida de Confiança. Quando o valor de *Gain* = 0 a confiança da

regra é igual a confiança mínima ($conf(LHS \Rightarrow RHS) = minconf$), e a partir daí todos os outros valores são calculados, sendo que *Gain* = 1 evidencia um valor de dependência estatística total, pois seria possível apenas quando o valor de minconf é insignificante (minconf = 0, i.e. sem limitação de confiança), a $conf(LHS \Rightarrow RHS)$ = 1 e o sup(LHS) = 1, ou seja, em todos os momentos que ocorresse LHS, também ocorreria RHS, com LHS em todas as instâncias da base de dados analisada.

A vantagem desta medida sobre a confiança é que pode-se calcular com mais exatidão a influência do elemento antecedente sobre o consequente, provocando assim uma maior credibilidade à medida, podendo também, ser utilizada para selecionar regras.

1.2.3. Pós-processamento das Regras de Associação

Na etapa de Pós-Processamento, o conhecimento, obtido por meio das Regras de Associação, pode ser simplificado, avaliado, visualizado ou simplesmente documentado [Piri et al. 2018]. A etapa de extração de Regras pode ser considerada simples, mas a compreensão de um conjunto significativo de regras torna a etapa de pós-processamento um desafio. Pode-se dividir o Pós-Processamento em três etapas [Namaki et al. 2017, Simard et al. 2016, Hendrickx et al. 2015, Zhao et al. 2009], sendo elas:

- **Avaliação:** etapa na qual é realizada a avaliação do conhecimento extraído do conjunto de dados por meio de critérios, tais como confiabilidade, aplicabilidade.
- **Interpretação e explanação:** etapa na qual é realizada a documentação, visualização, modificação e/ou comparação (ao conhecimento pré-existente) do conhecimento extraído do conjunto de dados de forma a torná-lo compreensível.
- **Filtragem:** etapa na qual é realizada a filtragem do conhecimento que foi extraído do conjunto de dados, podendo ser realizada por vários mecanismos que variam de acordo com a técnica utilizada.

Na busca por facilitar a interpretação de um número elevado de Regras de Associação, foram realizados diversos estudos, nos quais foram propostas diferentes abordagens tais como: Avaliação por consulta, Técnicas de visualização, etc.

- **Avaliação por consulta:** permite que o usuário explore o conjunto de regras por meio do uso de uma linguagem de consulta (tipicamente inspirada no SQL - Structured Query Language) [Meo et al. 1996].
- **Poda de regras:** Tem o objetivo de eliminar (excluir) regras redundantes ou que não são interessantes para o usuário [Toivonen 1996].
- **Técnicas de visualização:** utilizam de recursos gráficos, como Redes, para uma estruturação visual do conhecimento, possibilitando o uso de técnicas gráficas, bem como a observação direta das relações obtidas no processo de mineração de dados [Klemettinen et al. 1994] e, consequentemente, auxiliando na interpretação de Regras de Associação.

1.3. Redes de Regras de Associação

Para compreender o funcionamento e a estrutura das Redes de Regras de Associação, é importante ter o conhecimento de alguns conceitos básicos de Redes. Portanto, é imprescindível o esboço de alguns elementos associados às Redes para se poder introduzir as Redes de Regras de Associação

1.3.1. Redes

Uma rede é uma representação simplificada de uma estrutura de padrões e conexões, ou elementos e suas relações, que podem ser demonstradas por um Grafo. O termo "Rede" é usado para referenciar todo sistema que possa ser retratado por um Grafo, que é um termo usado para caracterizar um conjunto de ferramentas conceituais específicas com o objetivo de descrever a Rede.

Uma Rede, em seu modelo mais simples, tem como componentes um conjunto de pontos unidos em pares por linhas. Os pontos também podem ser referidos como "nós" ou "vértices", e as linhas chamadas de "arestas" (Figura 1.2).

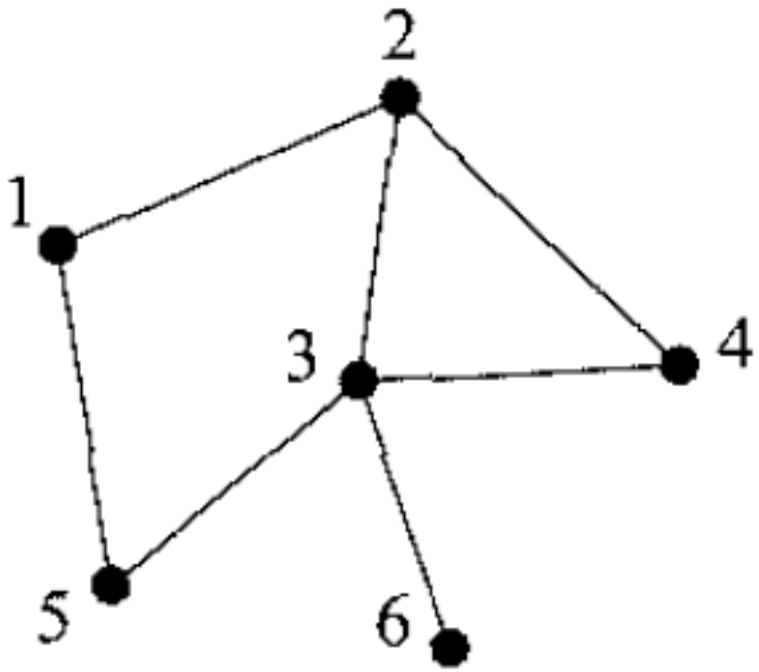


Figura 1.2. Representação simplificada de uma Rede

Normalmente, uma Rede R é representada com $R = (V, E)$, em que V é o conjunto de vértices e E é um conjunto de arestas, que podem estar ligados a alguns pares de vértices em V. Estatisticamente, um Grafo pode ser caracterizado por valores derivados, tais como o grau médio dos nós e o comprimento médio (caminho) entre os nós. Características adicionais como: diâmetro da rede, número de triângulos, número de isomorfismos e o coeficiente de agrupamento também podem ser analisados [Nettleton 2013].

Em uma Rede R = (V, E), vários links e auto-conexões não são permitidas de-

pendendo do tipo de Rede que está sendo implementada. Se G é uma Rede Dirigida, considera-se o conjunto universal, denotado por U, contendo todas as $|V| * (|V| - 1)$ potenciais ligações dirigidas entre um par de nós em V, o qual |V| denota o número de elementos em V. Se R é uma Rede Sem Direção, o conjunto universal U contém $|V| * (|V| - 1)$ links. Deste modo, a representação da Rede está relacionada diretamente ao tipo de dado que ela representa.

O primeiro passo na análise da estrutura de uma Rede é muitas vezes a visualização da mesma, existem diversas ferramentas que podem gerar a imagem que representa uma Rede e suas características. Esta etapa é extremamente útil para análise de dados de Rede, pois permite que se possa identificar informações importantes que de outra forma seria difícil com dados brutos. O olho humano é extremamente talentoso em escolher padrões, e visualizações permitem ao ser humano colocar esta cognição para trabalhar em problemas de Rede [Newman 2010].

O estudo de um grande volume de Redes do mundo real, revelou que algumas dessas estruturas, estão presentes em muitos sistemas naturais e artificias, constituindo a base de classificação de sistemas reais. Uma Rede, também chamada de Grafo em literatura matemática, é uma coleção de vértices unidos por aros. Vértices e aros são também chamados de nós e arestas em ciência da computação, sites e títulos em física e atores e os laços em sociologia [Newman 2010].

As Redes, em sua maioria, possuem no máximo uma única aresta entre qualquer par de vértices. Em casos raros, em que está condição não é atendida, e um par de vértice está ligado por mais de uma aresta, essas são referenciadas como multiarestas. Outra ocasião não muito comum na maioria das Redes, são as auto-arestas ou *auto-loops*. Estas estruturas são chamadas dessa forma pois conectam um vértice a se mesmo. Uma Rede que não tem nem auto-arestas nem multiarestas é chamada de Rede Simples ou Grafo Simples. Uma Rede com multiarestas e/ou auto-arestas é chamada um Multigrafo Figura 1.3.

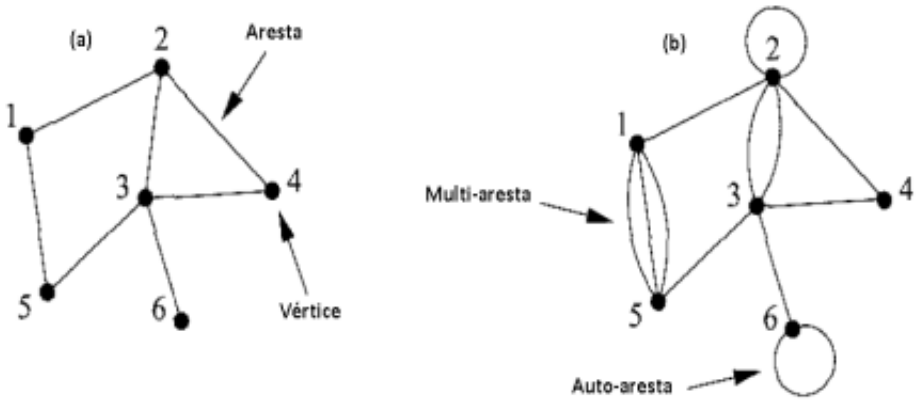


Figura 1.3. Duas pequenas Redes. (a) Rede Simples (b) Rede Multiarestas com 1 (uma) multiaresta e 2 (duas) auto-arestas

Existem diferentes maneiras de representar matematicamente uma Rede, uma delas, que se apresenta com bons resultados, é a Matriz de Adjacência que representa as

ligações existente entre os vértices de uma Rede.

As Matrizes de Adjacência das Redes indicam a quantidade de arestas entre cada um dos vértices que são identificados pela linha i e coluna j da matriz.

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 0 & 1 & 0 & 0 & 3 & 0 \\ 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{pmatrix}$$

Figura 1.4. Matrizes de Adjacência

A matriz da Figura 1.4 por exemplo, representa os grafos da Figura 1.3.

Segundo Newman (2010), os valores em uma Matriz podem representar o peso, ou seja, o grau de aproximação entre os vértices de uma Rede formando uma Rede Ponderada. As matrizes também podem indicar algum tipo de direcionamento da Rede formando, nesse caso, uma Rede Direcionada que pode adotar, por exemplo, o valor "1"se for de i para j e o valor "0"se for de j para i.

Juntar dois vértices de uma só vez pode ser fazer necessário em alguns tipos de ligações. Uma Rede que represente relações familiares por exemplo, na qual uma família possui mais de um membro, deve usar uma hiper-aresta, que é um tipo generalizado de aresta que junta mais de dois vértices, a fim de demonstrar da melhor maneira uma conexão entre seus membros. Os Grafos em que as hiper-arestas aparecem são chamados de Hipergrafos Figura 1.5.

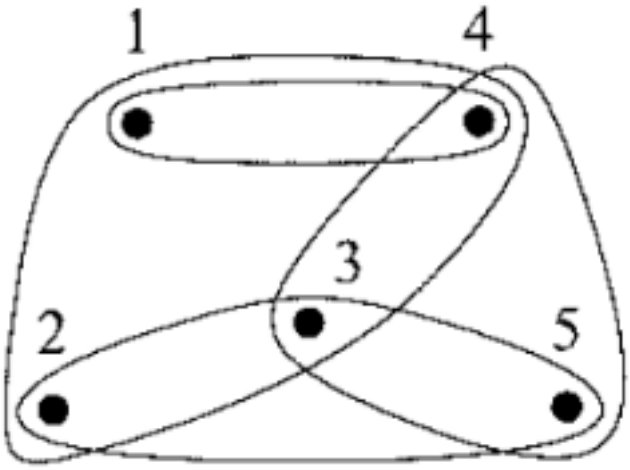


Figura 1.5. Hipergrafo - em hipergrafos as ligações são simbolizadas por loops que circulam os vértices

Uma associação de nós representado por um Hipergrafo, pode ser representado por uma Rede Bipartida.

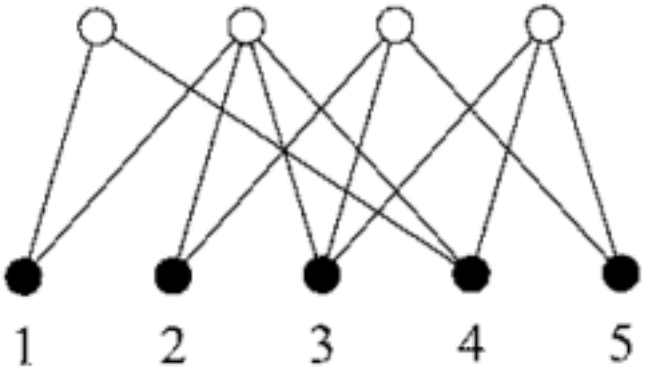


Figura 1.6. Rede Bipartida

Em uma Rede Bipartida (Figura 1.6), existe dois tipos de vértices: Um representando os vértices originais e o outro representando os grupos a que pertencem.

Uma outra forma de representar uma Rede é por meio de árvores. Esta estrutura é uma Rede Conectada, não direcionado e que não contém circuitos fechados. Os nós na Rede são acessíveis por todos os outros através de algum caminho. Uma Rede pode também consistir em duas ou mais partes desconexas uma da outra, e também é chamado de árvore um nó individual sem conexão. A Rede é chamada de floresta quando todos os seus componentes são árvores. A representação gráfica de uma Árvore consiste de um nó raiz no topo da estrutura, e ramificações abaixo, nas quais os vértices na parte inferior que só possuem uma ligação são chamados de folha Figura 1.7.

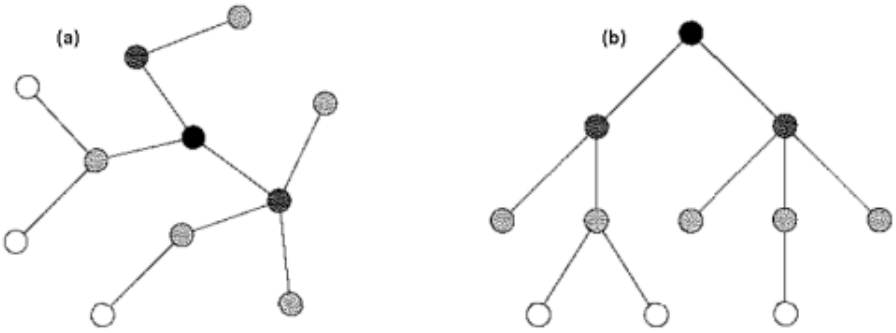


Figura 1.7. Dois esquemas de uma mesma árvore. Na árvore (a) os vértices estão posicionados conforme uma conveniência e na árvore (b) segue-se a estrutura com nó raiz

1.3.1.1. Medidas de Centralidade

Se a estrutura de uma Rede é conhecida, pode-se calcular, a partir de uma variedade de medidas quantitativas quais são aquelas características que conseguem se adaptar melhor a cada tipo de Rede. Algumas métricas são oportunas e selecionadas de acordo com o conhecimento que se deseja extrair [Newman 2010].

Estas medidas são importantes métricas que podem ser aplicadas em diversos tipos de situações. Redes Sociais, Redes de transportes e mercado financeiro, são alguns exemplos. Ela podem ser utilizadas para medir o quanto um vértice de uma Rede é mais ou menos importante do que os demais.

A Centralidade de grau ou informação, é uma medida que diz respeito ao número de ligações diretas que um certo vértice possui, ou seja, a centralidade de grau é uma contagem das adjacências de um vértice.

Outra medida de centralidade é a Centralidade de Auto-vetor, que define a importância de um nó verificando se este está ligado a outros vértices que se encontram em uma posição central na Rede. Se for esse for o caso, este nó tem uma ta Centralidade de Auto-vetor.

A Centralidade de Intermediação, também chamada de *Betweenness*, avalia a importância de um vértice considerando a quantidade de menores caminhos que passa por tal vértice.O vértice com maior centralidade de intermediação é aquele que participa de maneira mais ativa em um processo de interação, no qual os caminhos mais curtos são percorridos [Silva 2010].

O *Page Rank*, considera a importância de um *Website* levando em conta a quantidade de links que cada página possui. Na internet, os sites e páginas estão ligados por links, formando uma Rede de Informação.

Pode-se citar mais uma série de outras métricas importantes no estudo das Redes. Com os valores da centralidade de todos os vértices, ou de todas as arestas, define-se o Comprimento Médio do Caminho [Brandes 2001] e o Diâmetro da Rede [Coleman and Moré 1983].

1.3.2. Construindo uma Rede de Regras de Associação

Redes de Regras de Associação (ARN, do inglês Association Rules Network) consistem de uma estrutura para sumarizar, podar, e analisar um conjunto de Regras de Associação extraídas, para a concepção de hipóteses. Pode-se utilizar as Redes com o intuito de facilitar o processo e alcançar uma estrutura capaz de obter Regras automaticamente.

De uma perspectiva de descoberta de conhecimento, a ARN facilita o entendimento das relações e da utilidade das características dos dados. Matematicamente, as ARNs são Hipergrafos Direcionados.

A ideia central das Redes de Regras de Associação é que as Regras descobertas no processo de mineração passem por uma espécie de poda, afim de identificar e informações que atendam o objetivo específico da pesquisa. Se a pesquisa tiver uma variável de interesse, esta se torna o "alvo" ou "objetivo" e uma Rede formada por variáveis que estão relacionadas ao objetivo pode ser formada. Em seguida, uma estrutura pode ser elaborada e testada por meio de métodos estatísticos.

Normalmente a tarefa de Mineração de Dados envolve centena de variáveis que podem apresentar inconsistências. A ARN pode ser utilizada, para podar essas informações. Resumindo as ARNs, oferecem os seguintes recursos: poda de informações, estrutura de rede e geração e avaliação de hipóteses.

Para a criação da ARN, são realizadas quatro etapas:

- **Passo A:** Tendo uma base de dados e o um suporte e confiança mínima, todas as Regras de Associação devem ser extraídas utilizando algum algoritmo padrão como *Apriori*, *Apriori-TID* ou *FP-Growth*.
- **Passo B:** Escolher um alvo. Uma variável de interesse, que será representada no grafo como o nó objetivo.
- **Passo C:** Realizar a poda. Retirando informações inconsistentes e não relevantes. O resultado disso é uma ARN.
- **Passo D:** Identificar caminhos mais curtos entre o nó objetivo e os demais na ARN. O conjunto destes caminhos representa a Rede exploratória para o nó objetivo.

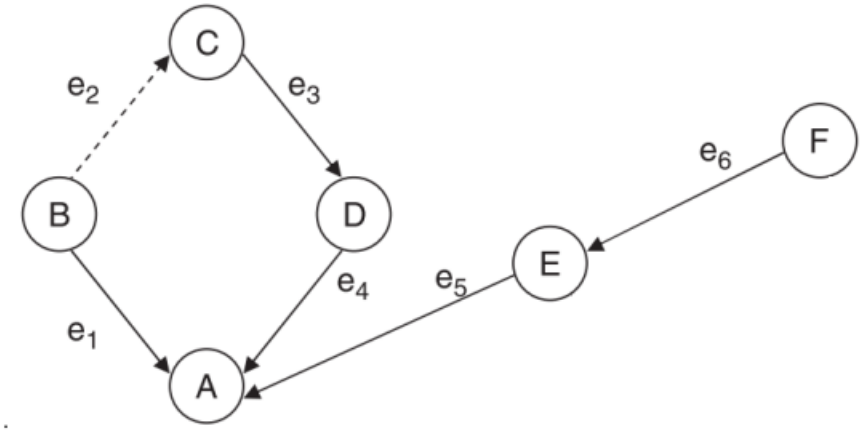


Figura 1.8. Exemplo de ARN

Na Figura 1.8, a ARN tem como alvo o vértice "A". Logo, todas as Regras extraídas que possuem A como consequência são selecionadas. Neste caso apenas as regras $(B \Rightarrow A)$, $(D \Rightarrow A)$ e $(E \Rightarrow A)$ serão selecionadas.

A ARN oferece o benefício de organização de regras em um contexto, de tal forma que um raciocínio objetivo pode ser explicado usando as regras mais relevantes no conjunto. A poda local é uma outra vantagem da ARN, já que uma regra redundante para um nó objetivo em particular pode tornar-se relevante para outro alvo, tornando essa abordagem mais flexível do que a poda com base em medidas estatísticas.

As Redes de Regras de Associação fornecem um mecanismo para sintetizar as Regras de Associação de maneira estruturada. Elas implementam uma metodologia que integra a pesquisa de mineração de dados com métodos estatísticos. A pesquisa em mineração está relacionada ao aperfeiçoamento de uma abordagem teórica existente, já os métodos estatísticos estão ligados a validação das teorias sobre os dados da pesquisa. Estas características demonstram a amplitude do uso da ARN e as diferentes áreas em que podem ser aplicada.

1.4. Redes de Regras de Associação Filtradas

Com o aumento dos conjuntos de dados, enumerar todas as possíveis combinações dos elementos, e depois verificar sua a correlação se torna uma tarefa computacional inviável. A utilização de Mineração de Regras se faz importante pois oferece mecanismos que podem ser utilizadas na validação de hipóteses.

Por meio da Mineração de dados é possível filtrar grandes quantidades de dados e gerar padrões que são interessantes, e até surpreendentes em certos casos. A desvantagem dessa metodologia é grande quantidade de padrões e hipóteses que são geradas a partir de um conjunto de dados, tantas que se torna difícil decidir quais são confiáveis e quais valem a pena serem analisadas.

Algoritmos utilizados na descoberta de Regras usam medidas capazes de avaliar a qualidade de uma regra. O supor mínimo e a confiança se destacam, embora *Lift*, *Gain*, *Certainty Factor*, *Added Value* ou *Leverage* também sejam medidas que fornecem informações sobre a regra. Os algoritmos, normalmente, extraem todas as Regras de Associação de um conjunto de dados de acordo com o suporte mínimo e valor mínimo de confiança, fazendo com que o número de Regras extraídas seja alto, dificultando a capacidade de exploração do usuário.

A Rede de Regras de Associação Filtrada (*Filtered*-ARN) alia medidas objetivas com estrutura de Rede para modelar e auxiliar a visualização e análise das regras extraídas de um *dataset*. A estrutura da *Filtered*-ARN é semelhante a da ARN, o que muda é a maneira como as regras são selecionadas, já que considera influência estatística comprovada para promover a identificação de hipóteses com maior probabilidade de serem verdadeiras.

O resultado desse processo são regras que possuam maior chance de serem interessantes, e o objetivo central das Redes de Regras de Associação Filtradas é apresentar um Grafo com estas regras. As hipóteses geradas podem ser validadas por diversos métodos, um deles é a a avaliação realizada por um especialista na área que está sendo pesquisada. Na *Filtered*-ARN o usuário pode visualizar um conjunto de itens que têm influência estatística em vez de elementos que apenas se relacionam com o item objetivo.

1.4.1. Modelo Proposto para Geração de Hipóteses de Maior Confiabilidade

A Tabela 1.1 contém o *dataset Lenses*. Nela é possível identificar algumas informações sobre um paciente e que tipo de lente é recomendado para ele.

Usando o *dataset Lenses*, se o usuário construir uma ARN com nó alvo o atributo "[lenses]=hard contact lense", o atributo "[prescription]=myope" aparece diretamente conectado ao nó objetivo.

A hipótese "um paciente com miopia tem maior probabilidade de usar uma lente rígida" pode ser formulada a partir do conhecimento extraído da Rede. No entanto a ARN, pode apresentar relações que não possuem influência entre os elementos da regra.

Ao calcular a medida *Added Value* da regra "[prescription]=myope $\Rightarrow$[lenses]=hard contact lense", encontra-se AV = 0, o que indica que a variável "prescription]=myope" não apresenta relevância para a ocorrência da variável "[lenses]=hard contact lense". Portanto

Tabela 1.1. Trecho do *Dataset Lenses*

Age	Prescription	Astigmatic	Tear	Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft contact lenses
young	myope	yes	reduced	none
young	myope	yes	normal	hard contact lense
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft contact lenses
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard contact lense
Pre-presbyopic	myope	no	reduced	none
Pre-presbyopic	myope	no	normal	soft contact lenses
Pre-presbyopic	myope	yes	reduced	none
Pre-presbyopic	myope	yes	normal	hard contact lense
Pre-presbyopic	hypermetrope	no	reduced	none
Pre-presbyopic	hypermetrope	no	normal	soft contact lenses
Pre-presbyopic	hypermetrope	yes	reduced	none
Pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard contact lense
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft contact lenses
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

a hipótese levantada possui caráter equivocado, pois os dados não possuem influência direta.

A *Filtered*-ARN permite a exploração de um alvo com análise de dependência entre os elementos das regras utilizando filtros de medidas objetivas na etapa da seleção das regras.

O Algoritmo para geração da *Filtered*-ARN pode ser descrito em 3 passos:

- **Passo A:** Extração das Regras com corte por suporte e confiança mínimos.
- **Passo B:** Efetuar o calculo das medidas *Added Value* e *Gain*, cortando todas as regras que tiverem AV = 0 e valore inferiores ao ganho mínimo.
- **Passo C:** Escolher uma variável de interesse, ou seja, um nó alvo.

Em relação a etapa de Mineração de Regras, a única restrição adicionada se comparada a uma Mineração de Regras de Associação convencional é que as regras devem possuir conjuntos unitários no antecedente e consequente. Esse formato facilita a modelagem da *Filtered*-ARN.

O segundo passo, utiliza medidas objetivas para filtrar as regras. Para a seleção, considera-se apenas as regras que possuem elementos com dependência estatística e definição do ganho mínimo de influência.

O ultimo, passo o usuário deve decidir o item que deseja entender do conjunto de dados. Escolhido um alvo, é efetuada a construção da *Filtered*-ARN. Primeiro, o item alvo é colocado no gráfico (Nível 0). Então todas as regras em que seu *LHS* ainda não consta no gráfico e que possuem o item alvo como *RHS* são modelados na rede (Nível 1). O processo é repetido tendo agora como alvo os itens do Nível 1, em seguida o Nível 2 e assim por diante, e só termina quando não há mais regras a serem modeladas.

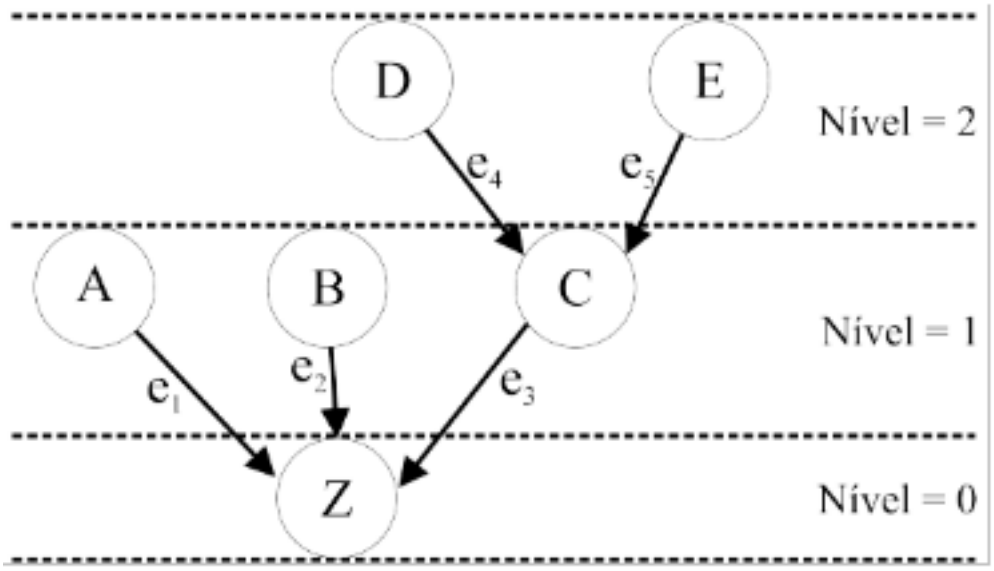


Figura 1.9. Níveis dos nós de uma *Filtered*-ARN

Para avaliar a *Filtered*-ARN dois métodos podem ser utilizados: i) faz-se uma comparação com uma ARN convencional, analisando as diferenças, as vantagens e desvantagens de usar cada uma dessas abordagens; ii) faz-se uma comparação com uma árvore de decisão com o objetivo de confrontar os resultados gráficos e decidir qual a melhor estrutura para análise do *dataset*.

1.4.2. Procedimento Metodológico

Para a aplicação do minicurso foram explanados os conceitos básicos descritos nas Seções anteriores e realizada a demonstração de experimentos práticos com alguns datasets. Foi selecionado a base de dados "*Lenses*" para exemplificação dos resultados.

Os objetivos desejados com a aplicação desse minicurso foram:

- Apresentar fundamentos sobre Redes de Regras de Associação Filtradas e suas características.
- Abordar conceitos que se fazem necessários para a construção de uma *Filtred*-ARN, como por exemplo, Regras de Associação, Redes e Medidas Objetivas.
- Realizar, por meio de um experimento, a construção de uma *Filtreddataset*.
- Comparar a *Filtered*-ARN com outras estruturas, para visualizar as diferença na geração de regras ou saídas, e demonstrar como o levantamento de hipóteses pode ser impactado.

1.4.2.1. *Dataset*

Para a abordagem prática da *Filtered*-ARN, o *dataset Lenses* foi utilizado. Neste conjunto de dados, cada linha representa os atributos de um paciente e a lente de contato que foi prescrita para ele. Por meio dessas informações é possível descrever quais características influenciam na prescrição de cada tipo de lente de contato. Vale a pena destacar que este *dataset* é uma simplificação do problema, logo as hipóteses levantadas podem não corresponder ao cenário real.

Por possuir um pequeno número de atributos, todos os valores foram considerados (suporte mínimo definido como 0), já a confiança mínima foi definida para 0,25, dessa forma classes que ocorreram pelo menos 1 em 4 vezes podem serem estudadas. O tamanho da classe foi definido com 2, ou seja, um item no LHS e um item no RHS.

Foram extraídas 99 Regras de Associação desta configuração. Após a etapa de filtragem 60 regras restaram.

1.4.2.2. *Filtred*-ARN e comparações

A Rede Filtrada foi gerada tendo como um alvo o item "[lenses] = hard". A Rede possui 4 níveis, onde no nível 0 se encontra o item objetivo. Apenas 2 itens estão ligados ao alvo: "[tear]=normal" e "[astigmatic]=yes". Essas regras podem ser capazes de implicar no item "[lenses] = hard", tornando interessante o estudo, já que são os únicos parâmetros que geram influencia no item objetivo. Por este motivo a geração de hipóteses verdadeiras apresenta um alto grau de probabilidade.

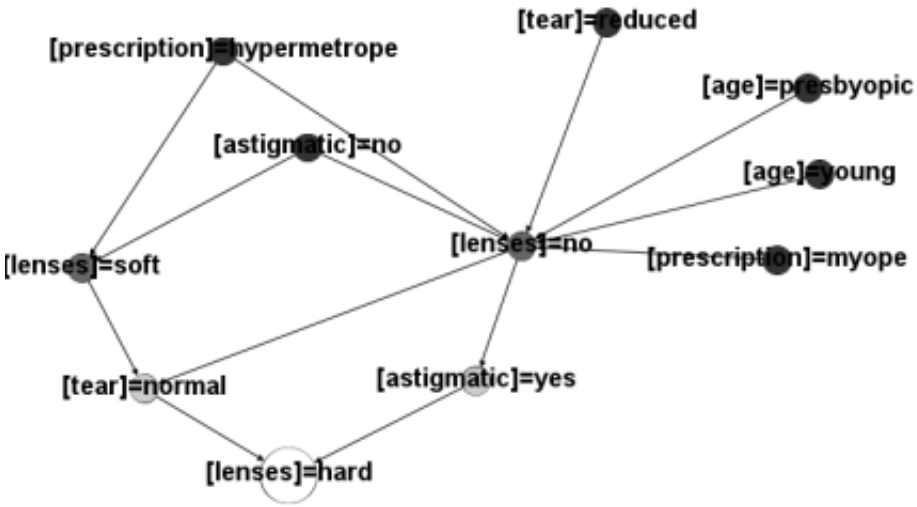


Figura 1.10. *Filtered*-ARN com "[lenses]=hard" como item alvo

Na Figura 1.10 é possível observar que os itens do nível 2 são as outras classes que indicam o tipo de lente "[lenses]=soft" e "[lenses]=no" e que estas implicam na ocorrência dos itens no nível 1 ("[astigmatic]=yes" e "[tear]=normal").

Nota-se no nível 3 a existência de diferentes regras que possuem conexão apenas

com a classe “[lenses]=no”. O que gera hipóteses para a construção de uma nova *Filtered*-ARN com “[lenses]=no” como objetivo, provocando um novo direcionamento na exploração do conhecimento.

A Figura 1.11 apresenta uma Rede de Regra de Associação comum. O item objetivo é o mesmo (“[lenses]=hard), no entanto esta estrutura possui 3 níveis, e é completamente diferente da Rede Filtrada.

Nota-se, no nível 1, que diversos itens implicam no nó alvo. O motivo desta ocorrência está no fato de que essa rede considera regras sem influencia comprovada (*Added Value* = 0), como “[prescription]=myope”=>“[lenses]=hard” e “[age]=young”=>“[lenses] =hard”, o que leva a geração de hipóteses equivocadas a respeito do uso de lentes rígidas.

Vale salientar outra diferença notória: O aumento de número de nós no nível 3, sem a distinção de dependência entre os mesmos. Isto dificulta a realização de estudos, pois a estrutura demonstra que todos os itens de um nível possuem o mesmo grau de importância.

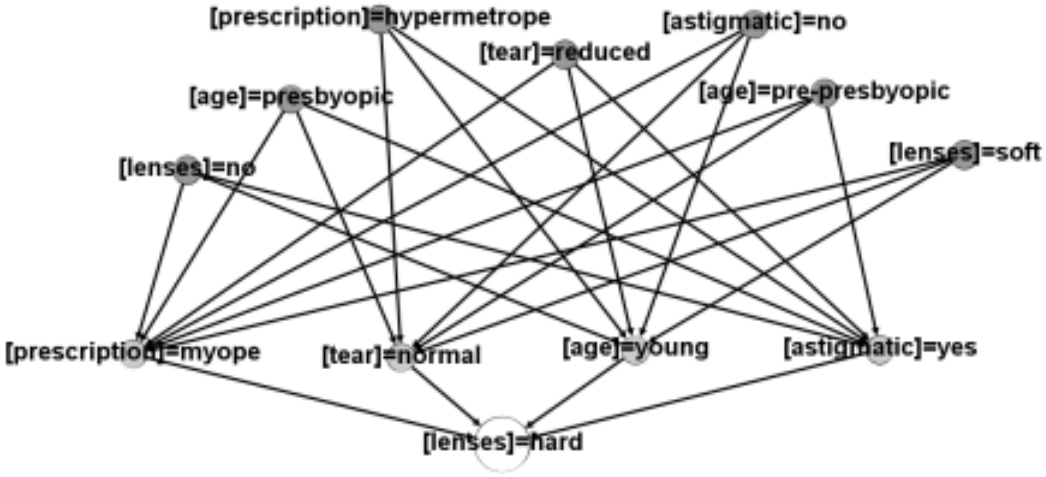


Figura 1.11. ARN com “[lenses]=hard” como item alvo

Ao comparar a *Filtered*-ARN com uma estrutura da árvore (Figura 1.12) de decisão, é possível identificar diferenças na explicação dos itens objetivos. Nas duas estruturas os itens “[tear]=reduced” e “[prescription]=hypermetrope” estão conectados diretamente a “[lenses]=no”, porém na Rede Filtrada, outras regras foram geradas.

A árvore indica que “[prescription]=myope” se liga diretamente à “[lenses]=hard”, porém, levando em consideração que esta condição não tem influência (AV = 0), esta saída pode ocasionar na geração de hipóteses equivocadas. Para evitar este tipo de equívoco se utiliza a *Filtred*-ARN.

1.5. Considerações Finais

As Regras de Associação possuem, como principal, tarefa de auxiliar a encontrar elementos que implicam na presença de outros em uma mesma transação. A partir do encontro

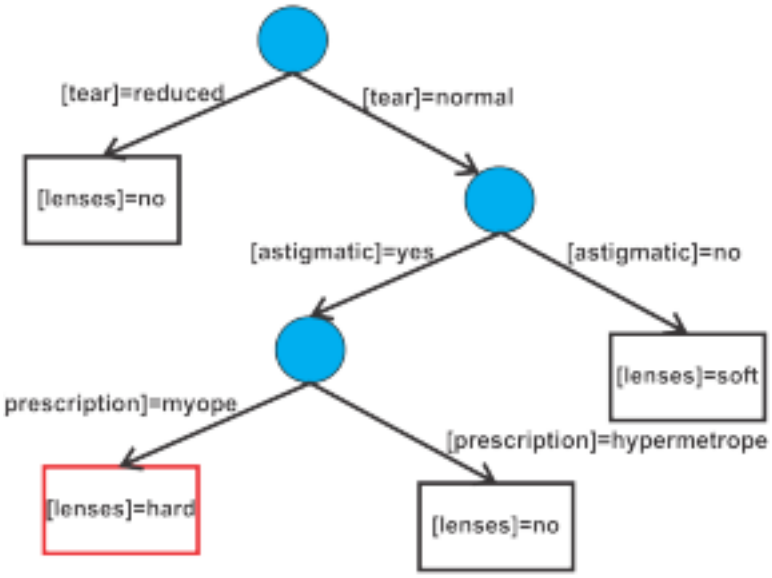


Figura 1.12. Árvore de Decisão para o *dataset Lenses*

desses elementos, é possível identificar relacionamentos e padrões frequentes em conjuntos de dados.

A utilização da descoberta de Regras de Associação possui grande aplicabilidade em diferentes contextos, permitindo a obtenção de conhecimento e resolução de problemas. Porém, há a necessidade de adequações nas regras, principalmente na fase de pós-processamento, a fim de evitar problemas relacionados às suas análises. A aplicabilidade das Regras se estende para várias tarefas da mineração de dados, sendo necessária a adequação do processo de extração das regras levando em conta fatores como custo computacional, tempo de resposta, entre outros.

Nesse capítulo foram esclarecidos temas relacionados à obtenção de Regras de Associação, bem como conceitos básicos relacionados à Redes voltadas a Mineração de Regras de Associação para extração de conhecimento. E, finalmente, foi explicado o funcionamento das Redes de Regras de Associação Filtradas (*Filtered*-ARNs), que foram o maior enfoque deste trabalho.

Por fim, acredita-se que o uso das *Filtered*-ARNs, podem auxiliar na extração de padrões de interesse ao usuário, e também, em diversas fases da Mineração de Regras de Associação, conforme demonstrado na avaliação experimental.

Referências

[Aggarwal 2015] Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer, New York, USA, 1st edition.

[Agrawal et al. 1994] Agrawal, R., Imielinski, T., and Swami, A. (1994). Mining Association Rules between Sets of Items in Large Databases. *Special Interest Group on Management of Data*, 22(2):207–216.

[Agrawal and Shafer 1996] Agrawal, R. and Shafer, J. C. (1996). Parallel mining of asso-

ciation rules. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):962–969.

[Agrawal and Srikant. 1994] Agrawal, R. and Srikant., R. (1994). Fast algorithms for mining association rules. *Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proceedings of Twentieth International Conference on Very Large Data Bases (VLDB)*, pages 487–499.

[Brandes 2001] Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177.

[Coleman and Moré 1983] Coleman, T. F. and Moré, J. J. (1983). Estimation of Sparse Jacobian Matrices and Graph Coloring Blems. *SIAM Journal on Numerical Analysis*, 20(1):187–209.

[de Vasconcelos and de Carvalho 2004] de Vasconcelos, L. M. R. and de Carvalho, C. L. (2004). Aplicação de Regras de Associação para Mineração de Dados na Web. *Instituto de Informática da Universidade Federal de Goiás*, page 20.

[Gonçalves 2005] Gonçalves, E. C. (2005). Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas Objective and Subjective Measures for Association Rules. *INFOCOMP Journal of Computer Science*, 4(1):26–35.

[Han and Kamber 2006] Han, J. and Kamber, M. (2006). *Data Mining Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA, second edition.

[Han et al. 2012] Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Waltham, MA, USA, 3rd edition.

[Hendrickx et al. 2015] Hendrickx, T., Cule, B., Meysman, P., Naulaerts, S., Laukens, K., and Goethals, B. (2015). Mining association rules in graphs based on frequent cohesive itemsets. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9078, pages 637–648.

[Klemettinen et al. 1994] Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. (1994). Finding interesting rules from large sets of discovered association rules". In Nabil, R., editor, *Proceedings of 3rd International Conference on Information and Knowledge Management*, pages 401–407.

[Le and Vo 2016] Le, T. and Vo, B. (2016). The lattice-based approaches for mining association rules: a review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(4):140–151.

[Meo et al. 1996] Meo, R., Psaila, G., and Ceri, S. (1996). A new SQL-like operator for mining association rules. In Vijayaraman, T., editor, *Proceedings of the 22nd International Conference on very Large Data Bases*, pages 122–123.

[Namaki et al. 2017] Namaki, M. H., Wu, Y., Song, Q., Lin, P., and Ge, T. (2017). Discovering Graph Temporal Association Rules. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, pages 1697–1706.

[Nettleton 2013] Nettleton, D. F. (2013). Data mining of social networks represented as graphs. *Computer Science Review*, 7(1):1–34.

[Newman 2010] Newman, M. (2010). *Networks: An introduction*, volume 55. Oxford University Press, New York, USA, 1st edition.

[Pandey et al. 2009] Pandey, G., Chawla, S., Poon, S., Arunasalam, B., and Davis, J. G. (2009). Association Rules Network: Definition and Applications. *Statistical Analysis and Data Mining*, 1(4):260–179.

[Piri et al. 2018] Piri, S., Delen, D., Liu, T., and Paiva, W. (2018). Development of a new metric to identify rare patterns in association analysis: The case of analyzing diabetes complications. *Expert Systems with Applications*, 94:112–125.

[Sahar 2003] Sahar, S. (2003). *What Is Interesting: Studies on Interestingness in Knowledge Discovery*. PhD thesis, Tel-Aviv University.

[Silva 2010] Silva, T. S. A. (2010). *Um Estudo de Medidas de Centralidade e Confiabilidade em Redes*. PhD thesis, Centro Federal de Educação Tecnológica do Rio de Janeiro.

[Simard et al. 2016] Simard, F., St-Pierre, J., and Biskri, I. (2016). Mining and visualizing robust maximal association rules on highly variable textual data in entrepreneurship. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems - MEDES*, pages 215–222.

[Toivonen 1996] Toivonen, H. (1996). Sampling large databases for association rules. *The VLDB Journal*, pages 134–145.

[Usman and Usman 2016] Usman, M. and Usman, M. (2016). Multi-level mining and visualization of informative association rules. *Journal of Information Science and Engineering*, 32(4):1061–1078.

[Vinaya and Shah 2016] Vinaya, M. and Shah, K. (2016). Performance Evaluation of Distributed Association Rule Mining Algorithms. *Procedia - Procedia Computer Science*, 79:127–134.

[Zaki and Meira 2013] Zaki, M. J. and Meira, M. J. (2013). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

[Zhao et al. 2009] Zhao, Y., Zhang, C., and Cao, L. (2009). *Post-Mining of Association Rules : Techniques for Effective Knowledge Extraction*.