UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

MEDIDAS OBJETIVAS ASSIMÉTRICAS APLICADAS NA CONSTRUÇÃO DE REDES DE REGRAS DE ASSOCIAÇÃO

DARIO BRITO CALÇADA
SOLANGE OLIVEIRA REZENDE

Nº 428

RELATÓRIOS TÉCNICOS

São Carlos – SP
Nov./2018

Medidas objetivas assimétricas aplicadas na construção de redes de regras de associação

Dario Brito Calçada[1]
Solange Oliveira Rezende[1]

[1]Universidade de São Paulo
Instituto de Ciêncicas Matemáticas e de Computação
Laboratório de Inteligência Computacional
São Carlos – SP – Brazil

Novembro/2018

Resumo

O número de dados gerados a cada dia aumenta potencialmente, consequentemente, a dificuldade em se obter conhecimento pela análise dos mesmos segue a mesma proporção. Compreender esses dados, ou seja, conhecer a informação e o conhecimento implícito nesses dados assume, cada vez mais, um papel relevante no apoio à tomada de decisão. Para análise dos dados e das informações e extração de conhecimento foram desenvolvidas técnicas de mineração de dados. Em geral, o conhecimento descoberto por intermédio de processos de mineração de dados, em um paradigma simbólico, é expresso na forma de padrões interpretáveis. A descoberta de regras de associação é uma técnica de mineração de dados, que procura identificar padrões de dados em datasets, permitindo, após a sua interpretação, usar o conhecimento específico acerca do problema em estudo. Para auxílio nos processos da mineração de regras de associação, em especial na geração automática de hipóteses, pode-se utilizar redes (grafos), que possibilitam uma melhor visualização das informações coletadas. Existem várias medidas de regras de associação que podem ajudar na seleção de padrões interessantes para a geração de hipóteses. A utilização de redes no auxílio da mineração de regras de associação, bem como a poda das regras para extração do conhecimento, são usadas as Redes de Regras de Associação (ARN - do inglês Association Rules Networks). Com o intuito de analisar o impacto de medidas objetivas assimétricas na construção das ARNs, foram realizados experimentos, os quais são descritos neste relatório técnico, bem como são apresentadas descrições das medidas e a relevância de cada uma delas no processo de seleção das regras e geração de hipóteses.

Abstract

The number of data generated each day potentially increases, therefore, the difficulty in obtaining knowledge by analyzing them follows the same proportion. Understanding these data, that is, knowing the information and the knowledge implied in these data, assumes, increasingly, a relevant role in the support to the decision making. Researchers developed data mining techniques for the analysis of data and information and extraction of knowledge. In general, knowledge discovered through data mining processes, in a symbolic paradigm, is expressed in the form of interpretable patterns. The discovery of association rules is a data mining technique, which seeks to identify patterns of data in datasets, allowing, after their interpretation, to use the specific knowledge about some problem. Networks (graphs) can be used to assist in the mining of association rules, especially in the automatic generation of hypotheses, which allow better visualization of the information collected. There are several measures of association rules that can help in selecting interesting patterns for hypothesis generation. Association Rules Networks (ARN) are used to aid the mining of association rules, as well as the pruning of rules for knowledge extraction. We performed experiments to analyze the impact of asymmetric objective measurements on the construction of the RNAs. We described this experiments in this technical report, as well as descriptions of the measures and the relevance of each of them in the process of selection of rules and generation of hypotheses.

LISTA DE ILUSTRAÇÕES

Figura 1 – Metodologia aplicada nos experimentos deste relatório técnico 19
Figura 2 – ARN com "lenses=soft" como item objetivo 21
Figura 3 – ARN com "lenses=hard" como item objetivo 21
Figura 4 – ARN com "lenses=no" como item objetivo 21
Figura 5 – ARN filtrada pela medida de Added Value com "lenses=soft" como item objetivo 23
Figura 6 – ARN filtrada pela medida de Added Value com "lenses=hard" como item objetivo 24
Figura 7 – ARN filtrada pela medida de Added Value com "lenses=no" como item objetivo 24
Figura 8 – ARN com minconf = 0 e filtrada pela medida de Added Value com "lenses=hard" como item objetivo 25
Figura 9 – ARN filtrada pela medida de CF com "lenses=soft" como item objetivo . . 26
Figura 10 – ARN filtrada pela medida de CF com "lenses=hard" como item objetivo . . 26
Figura 11 – ARN filtrada pela medida de CF com "lenses=no" como item objetivo . . . 26
Figura 12 – ARN com minconf = 0 e filtrada pela medida de Certainty Factor com "lenses=no" como item objetivo 27
Figura 13 – ARN filtrada pela medida de Convicção com "lenses=soft" como item objetivo 28
Figura 14 – ARN filtrada pela medida de Convicção com "lenses=hard" como item objetivo 28
Figura 15 – ARN filtrada pela medida de Convicção com "lenses=no" como item objetivo 29
Figura 16 – ARN com minconf = 0 e filtrada pela medida de Convicção com "lenses=soft" como item objetivo 29
Figura 17 – ARN filtrada pela medida de Gini Index com "lenses=soft" como item objetivo 30
Figura 18 – ARN filtrada pela medida de Gini Index com "lenses=hard" como item objetivo 30
Figura 19 – ARN filtrada pela medida de Gini Index com "lenses=no" como item objetivo 31
Figura 20 – ARN com minconf = 0 e filtrada pela medida de Gini Index com "lenses=hard" como item objetivo 31
Figura 21 – ARN filtrada pela medida de J-Measure com "lenses=soft" como item objetivo 32
Figura 22 – ARN filtrada pela medida de J-Measure com "lenses=hard" como item objetivo 32
Figura 23 – ARN filtrada pela medida de J-Measure com "lenses=no" como item objetivo 33
Figura 24 – ARN com minconf = 0 e filtrada pela medida de J-Measure com "lenses=no" como item objetivo 34
Figura 25 – ARN filtrada pela medida de Laplace com "lenses=soft" como item objetivo 34
Figura 26 – ARN filtrada pela medida de Laplace com "lenses=hard" como item objetivo 35

Figura 27 – ARN filtrada pela medida de Laplace com “lenses=no” como item objetivo 35
Figura 28 – ARN com minconf = 0 e filtrada pela medida de Laplace com “lenses=hard” como item objetivo 35
Figura 29 – ARN filtrada pela medida de Gain com “lenses=soft” como item objetivo . 36
Figura 30 – ARN filtrada pela medida de Gain com “lenses=hard” como item objetivo . 37
Figura 31 – ARN filtrada pela medida de Gain com “lenses=no” como item objetivo . . 37
Figura 32 – ARN com minconf = 0 e filtrada pela medida de Gain com “lenses=hard” como item objetivo 37

LISTA DE TABELAS

Tabela 1 – Filtros das Medidas Objetivas Assimétricas utilizadas nos experimentos . . 22

SUMÁRIO

1 INTRODUÇÃO 15

2 USO DE MEDIDAS OBJETIVAS NA CONSTRUÇÃO DAS ARNS: AVALIAÇÃO EXPERIMENTAL 19
2.1 Materiais e Métodos 19
2.1.1 Dataset 20
2.1.2 Pré-processamento 20
2.1.3 Extração de regras de associação 20
2.1.4 Construção da ARN 20
2.1.5 Seleção das Regras de Associação (Filtros das Medidas) 22
2.1.6 Construção das ARNs com regras filtradas 22
2.1.7 Experimento de Validação 22
2.2 Resultados e Discussão 23
2.2.1 ARN com filtro Added Value (AV-ARN) 23
2.2.2 ARN com filtro Certainty Factor (CF-ARN) 25
2.2.3 ARN com filtro Convicção (Conv-ARN) 27
2.2.4 ARN com filtro Gini Index (GI-ARN) 29
2.2.5 ARN com filtro J-Measure (J-ARN) 31
2.2.6 ARN com filtro Laplace (Laplace-ARN) 33
2.2.7 ARN com filtro Gain (Gain-ARN) 36
2.2.8 Síntese 38

3 CONSIDERAÇÕES FINAIS 39

REFERÊNCIAS 41

CAPÍTULO

1

INTRODUÇÃO

A evolução da tecnologia de computadores tem aumentado a capacidade humana de coleta, armazenamento e manipulação de dados, o que gerou uma nova necessidade de análise automática, classificação, e compreensão de todas as informações geradas. A mineração de dados, com sua abordagem de extração automática de conhecimento a partir de bases de dados, tem atraído cada vez mais atenção desde os anos 90. Na atual era do conhecimento, os dados gerados e armazenados por organizações modernas aumentam de modo extraordinário, e as tarefas de mineração de dados tornam-se uma tecnologia necessária e fundamental para a sustentabilidade e melhoria das mesmas (ZANIN et al., 2016).

Em geral, o conhecimento descoberto por intermédio de processos de mineração de dados, em um paradigma simbólico, é expresso na forma de padrões interpretáveis. A principal função das técnicas de mineração é a descoberta de quais padrões são mais frequentes e interessantes (WENG, 2016).

A descoberta de regras de associação é uma técnica de mineração de dados, que procura identificar determinados padrões de dados em bases, permitindo, após a sua interpretação, adquirir conhecimento específico acerca do problema em análise (AGRAWAL; IMIELINSKI; SWAMI, 1994).

O formato de uma regra de associação pode ser representado como uma implicação LHS⇒ RHS, em que LHS e RHS são, respectivamente, o lado esquerdo (do inglês - Left Hand Side) e o lado direito (do inglês - Right Hand Side) da regra, definidos por conjuntos disjuntos de itens.

Para cada regra (LHS ⇒ 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 (AGRAWAL; IMIELINSKI; SWAMI, 1994).

As regras de associação representam combinações de itens que ocorrem com determinada frequência em uma base de dados. Uma de suas aplicações típicas é a análise de transações de compra. A partir de uma base de dados que armazena produtos comprados por clientes, por exemplo, de um supermercado ou uma loja de departamentos, uma estratégia para a mineração de regras de associação (ARM - do inglês Association Rules Mining) poderia gerar o seguinte exemplo: $\{feijão, couve\} \Rightarrow \{linguiça\}$. Esta regra é utilizada para indicar que os clientes que compram os produtos feijão e couve, tendem também a comprar linguiça. O exemplo ilustra umas das características mais atrativas das regras de associação: elas são expressas em uma forma muito fácil de ser compreendida individualmente (WENG, 2016). Com o aumento do número de regras e do tamanho das mesmas a interpretabilidade fica prejudicada.

Conjuntos de itemsets frequentes são formados pelos itens que aparecem em um dataset com maior frequência. A frequência mínima desejada dos itemsets é definida pelo usuário. Encontrar itemsets frequentes é um dos desafios de mineração de regras de associação. O processo básico de extração das regras de associação foi apresentado pela primeira vez por Agrawal, Imielinski e Swami (1994). Em um trabalho posterior de Agrawal e Srikant (1994), foi elaborada uma abordagem mais explícita do processo de extração das regras de associação, sendo considerada como uma das mais importantes contribuições para o assunto.

O principal algoritmo apresentado, Apriori, não só influenciou a comunidade de mineração de regras de associação como afetou outros campos da mineração de dados, tornando-se uma das áreas amplamente pesquisadas e, portanto, outros algoritmos mais rápidos e eficientes foram apresentados. Muitos dos algoritmos foram elaborados baseados no Apriori ou em modificações do Apriori. O algoritmo Apriori também serviu de inspiração para abordagens em outros campos da mineração de dados, como nas regras de associação em sistemas distribuídos, incluindo Cloud Computing (KAKKAD; ALUVALU, 2013).

Para auxílio nos processos da mineração de regras de associação, pode-se utilizar redes em alguma fase do processo de mineração como no pré-processamento (LIU; ZHAI; PEDRYCZ, 2012), extração (NGUYEN et al., 2013) e pós-processamento (BARALIS et al., 2013), sendo que nesta última etapa, as redes também auxiliam na extração do conhecimento por meio de uma melhor visualização das informações coletadas. A análise dos dados com o uso de redes para o auxílio nas tarefas de mineração tem sido uma temática muito abordada nas pesquisas, devido a obtenção de resultados promissores com bancos de dados de larga escala (ARIDHI; NGUIFO, 2016).

Diversos sistemas no mundo real podem ser representados por meio de redes, por exemplo, os sistemas comerciais que produzem as relações entre clientes e produtos adquiridos. Desse modo, redes ou grafos são uma forma natural de representar matematicamente esses sistemas. A análise de Redes é a área do conhecimento que investiga a estrutura de uma rede a fim de obter conhecimento importante sobre seus elementos e suas interações (NEWMAN, 2010).

Realizando um estudo sobre mineração de regras de associação, Pandey et al. (2009) elaboraram uma abordagem que efetua a construção de uma rede utilizando as regras de associação extraídas de uma base de dados, a Rede de Regras de Associação (ARN - do inglês Association Rules Network). As ARNs possuem como objetivo principal, encontrar relações entre um item alvo dos dados (objetivo de estudo), a fim de se estabelecer as relações entre este e os demais itens existentes no dataset.

Embora as ARNs possibilitem um amplo estudo dos dados, elas permitem apenas conjuntos LHS e RHS unitários, isto significa que somente antecedentes e consequentes simples das regras de associação são aceitos, impossibilitando estabelecer maiores correlações dos itens com mais de um elemento como objetivo.

O processo de seleção das regras de associação realiza a filtragem por meio de valores mínimos para as medidas de suporte (minsup) e confiança (minconf)

Outras medidas de interesse objetivas, foram incluídas nas análises, e este relatório técnico visa demonstrar o impacto que algumas delas produzem na construção das ARNs e, por consequência, a relevância do estudo das mesmas na extração do conhecimento.

Assim, há indícios que o uso de medidas objetivas para a seleção de regras de associação podem ser utilizadas na construção de ARNs com menor grau de densidade possibilitando uma melhor visualização das informações.

Para comprovar essa hipótese são avaliados filtros de medidas objetivas para selecionar regras de associação que serão utilizadas na construção das ARNs. A validação foi feita comparando-se as redes desenvolvidas após a filtragem com aquelas que utilizam apenas suporte e confiança mínimos para seleção das regras de associação. Assim, os objetivos são: construir filtros de regras de associação com as medidas selecionadas para estudo e validar ARNs construídas com os filtros de regras efetuando a comparação com as ARNs construídas pelo método de Pandey et al. (2009).

O objetivo deste Relatório Técnico é apresentar a influência das medidas objetivas assimétricas na seleção das regras de associação que são utilizadas na construção das redes de regras de associação, bem como fazer o estudo do impacto da selação de regras na etapa de extração do conhecimento.

Visando apresentar a pesquisa realizada na elaboração deste relatório técnico, além desse capítulo introdutório, no qual foi apresentado uma visão geral, este documento foi estruturado da seguinte forma:

Capítulo 2: Uso de Medidas Objetivas na construção das ARNs: Avaliação Experimental - São apresentadas as medidas utilizadas para a construção das novas ARNs, bem como os experimentos relacionados a cada uma dela. Os resultados obtidos foram comparados a ARNs sem o filtro das medidas objetivas a fim de que a hipótese motivadora deste trabalho fosse comprovada

Capítulo 3: Considerações Finais - Neste capítulo são elaboradas as principais conclusões a respeito do trabalho elucidado neste relatório técnico. Além de descrever a importância do mesmo em sua área do conhecimento e trabalhos futuros.

CAPÍTULO

2

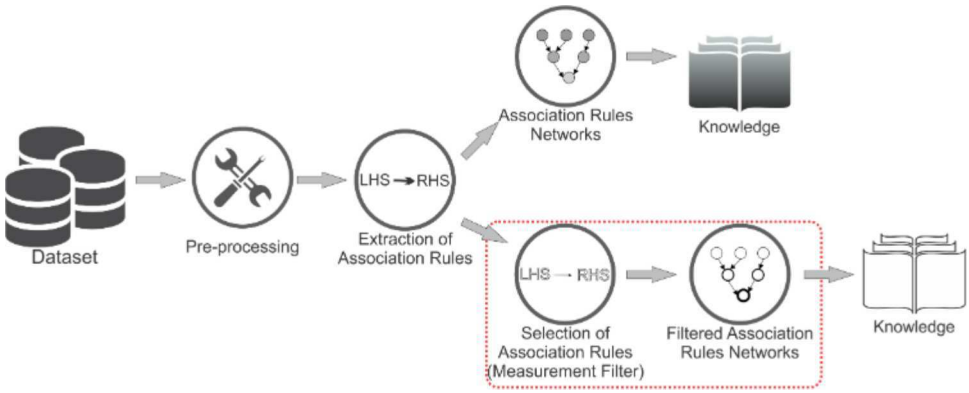
USO DE MEDIDAS OBJETIVAS NA CONSTRUÇÃO DAS ARNS: AVALIAÇÃO EXPERIMENTAL

O uso de medidas objetivas assimétricas de interesse pode auxiliar na descoberta de conhecimentos em abordagens que utilizam apenas medidas suporte e confiança, principalmente com o uso de redes como ferramenta para este processo.

2.1 Materiais e Métodos

Na Figura 1 são descritas as etapas executadas para estudo e exploração do uso de medidas objetivas assimétricas na construção das ARNs deste relatório técnico, dando ênfase à fase de seleção das regras, na qual foram estabelecidos os principais experimentos. Todas as fases da pesquisa foram formuladas com o intuito de comprovar a extração do conhecimento com o uso de ARNs com regras filtradas por medidas objetivas de interesse.

Figura 1 – Metodologia aplicada nos experimentos deste relatório técnico



Fonte: Elaborada pelo autor.

Os estudos abordados neste relatório foram feitos para analisar a capacidade de influência das medidas assimétricas de regras de associação na construção das ARNs, bem como a validação da hipótese de que cada medida pode ser utilizada para extrair conhecimentos diferentes dos dados ao se comparar com a ARN proposta por (PANDEY et al., 2009).

2.1.1 Dataset

Para analisar regras de associação por meio de ARNs com os filtros da medidas objetivas, um dataset foi selecionados da UCI [1]. O dataset selecionado foi o lenses (CENDROWSKA, 1987) por se tratar de uma base de dados de baixa complexidade.

O dataset lenses contém informações relativas a certos tipos de lente de contato e algumas características dos pacientes e foi usado para demonstrar os resultados neste relatório técnico. O dataset tem cinco atributos: idade (com 3 valores possíveis), prescrição oftalmológica (com 2 valores possíveis), astigmático (com 2 valores possíveis), taxa de produção de lágrimas (com 2 valores possíveis) e a classe da lente (com 3 valores possíveis).

2.1.2 Pré-processamento

Uma etapa de pré-processamento foi realizada com os dados oriundos do dataset Lenses. As transações estão disponíveis em uma matriz, na qual, cada linha representava um paciente e cada coluna um atributo. O dataset foi processado de forma que cada transação tenha os conteúdos no formato "atributo = valor". Usando o atributo "astigmatic", por exemplo, seus valores foram alterados para "astigmatic = yes", nos casos em que o valor encontrado no dataset é "1" e "astigmatic = no", nos casos em que o valor encontrado é "0". Essa alteração foi feita com o intuito de melhorar a leitura das regras após construção das ARNs

2.1.3 Extração de regras de associação

O algoritmo Apriori-TID implementado em Java® foi usado para gerar as regras de associação, conforme descrito anteriormente. Uma vez que o dataset tem um pequeno número de atributos, o suporte mínimo foi definido como 0.0, portanto, todos os valores poderiam ser considerados. A confiança mínima foi definida para 0.25 para evitar considerar todas as combinações possíveis. Além disso, o tamanho da regra foi especificado em dois, considerando conjuntos unitários para LHS e RHS. Usando essa configuração, foram obtidas 60 regras de associação candidatas para o dataset "lenses".

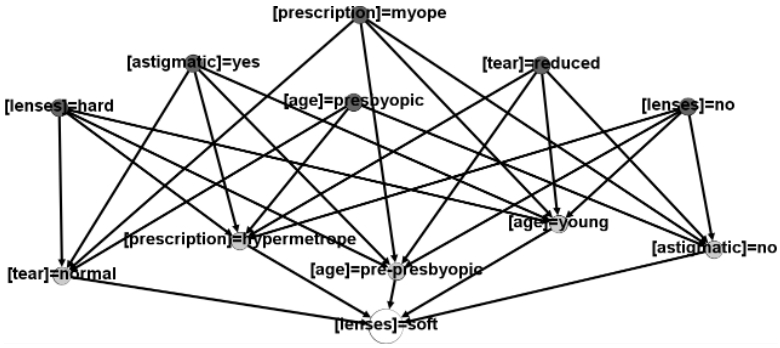
2.1.4 Construção da ARN

Foram construídas 3 (três) ARNs, considerando como itens objetivos, os três valores do atributo "classe de lente": "lenses = soft" (Figura 2), "lenses = hard" (Figura 3) e "lenses = no"

[1] http://archive.ics.uci.edu/ml/

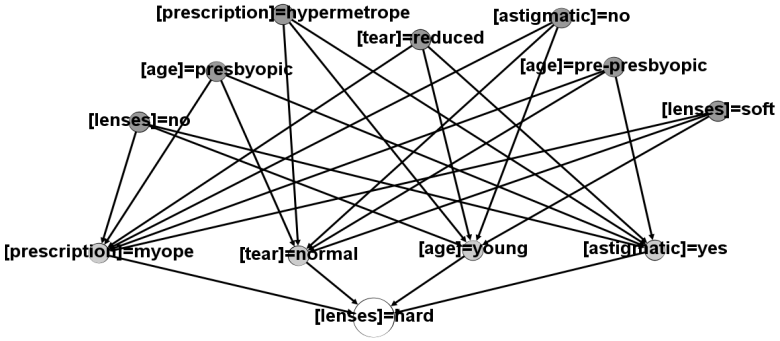
(Figura 4). Esses itens foram selecionados uma vez que o atributo classe é o principal objetivo de estudo em tarefas de classificação com este dataset. Nestes exemplos, não é necessário podar os links devido aos valores dos atributos no dataset. As redes foram representadas graficamente com o uso do software Gephi (BASTIAN; HEYMANN; JACOMY, 2009).

Figura 2 – ARN com "lenses=soft" como item objetivo



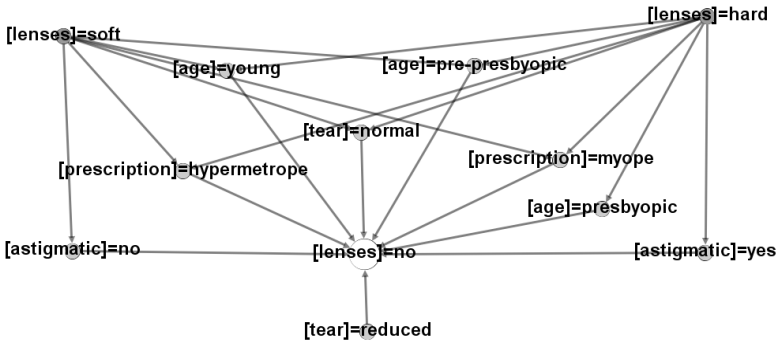
Fonte: Elaborada pelo autor.

Figura 3 – ARN com "lenses=hard" como item objetivo



Fonte: Elaborada pelo autor.

Figura 4 – ARN com "lenses=no" como item objetivo



Fonte: Elaborada pelo autor.

Para a avaliação das medidas objetivas de interesse das regras de associação, optou-se por verificar apenas medidas assimétricas, devido a estrutura direcionada da rede.

2.1.5 Seleção das Regras de Associação (Filtros das Medidas)

Após a extração das regras, foi implementado um script para cálculo das medidas selecionadas, bem como a filtragem das regras de acordo com a finalidade e o padrão estatístico conceitual de cada medida, conforme apresentados na Tabela 1. Para cada medida foram excluídas as regras que não apresentavam influência estatística entre os elementos, ou que não possuíam um valor numérico válido

Tabela 1 – Filtros das Medidas Objetivas Assimétricas utilizadas nos experimentos

Medida	Intervalo	Filtro
Added Value (SAHAR, 2003)	[-1;1]	AV = 0
Certainty Factor (SHORTLIFFE; BUCHANAN, 1975)	[-1;1]	CF = 0
Convicção (BRIN et al., 1997)	[0;∞[	Conv = 1
Gini Index (FISHER, 1996)	[0;0,5]	GI = 0
J-Measure (GOODMAN; SMYTH, 1991)	[0;1]	J = 0 e J = NaN
Laplace (SMALDON; FREITAS, 2006)	[0;1]	Lap = 0
Gain (FUKUDA et al., 1996)	[0;1]	Gain ≥ 0

Fonte: Dados da pesquisa.

Utilizando-se os valores descritos na Tabela 1, as regras foram filtradas. Por exemplo, o filtro da medida Added Value (AV) excluiu todas as regras que possuíam valor nulo para esta medida, selecionando assim as regras com relevante influência entre os itemsets conforme o conceito adotado pela métrica AV.

2.1.6 Construção das ARNs com regras filtradas

Foram selecionados os mesmos itens objetivo das ARNs construídas inicialmente, a fim de que a comparação possa ser estabelecida e a avaliação do impacto de cada medida possa ser evidenciado. A construção das ARNs com regras filtradas seguiram o processo convencional descrito por (PANDEY et al., 2009), porém com o uso das regras de associação selecionadas pelos filtros de suas respectivas medidas.

2.1.7 Experimento de Validação

Após o estudo da influência de cada medida objetiva em redes construídas com regras selecionadas por meio de minsup e minconf, com o intuito de validar o uso de cada métrica, foi realizado mais um experimento no qual a seleção das regras seria feita diretamente pelo script filtro implementado, i.e. atribuindo-se um valor nulo para minconf as regras são extraídas e selecionadas pelo filtro da respectiva medida a ser estudada. Foi-se construída uma nova ARN com regras filtradas e foi realizada a comparação com as ARNs sem regras filtradas.

2.2 Resultados e Discussão

Após a construção das ARNs sem filtragem, foram elaboradas as ARNs filtradas com cada medida de acordo com os valores apresentados na Tabela 1 e efetuada a comparação com o uso da construção gráfica de cada rede. Verificou-se o impacto que a filtragem das regras fez na estrutura de rede, bem como na seleção das regras.

2.2.1 ARN com filtro Added Value (AV-ARN)

A Medida Added Value (AV) indica o grau de influência do antecedente (LHS) em relação ao consequente (RHS). 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.

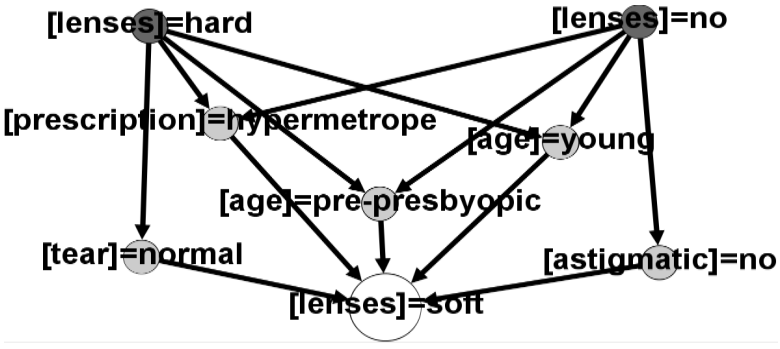
Conforme descrito na Tabela 1, o script do filtro AV = 0 realizou a exclusão de todas as regras de associação que possuem valor zero para a Medida Added Value.

As Redes de Regra de Associação filtradas pela medida Added Value foram construídas (AV-ARN) e a extração do conhecimento foi avaliada.

Analisando-se a Figura 5 e comparando com a Figura 2, percebe-se uma grande mudança na estrutura da rede, principalmente no nível L = 2, no qual o número de nós diminuiu de 6 (seis) para 2 (dois), ou seja, na construção da ARN, são inseridas arestas que aumenta a granularidade da rede, mas que não obrigatoriamente indicam algum tipo de influência verdadeira.

Com o filtro Added Value, forma-se a rede apenas com arestas formadas de regras de associação que possuem um real índice de influência.

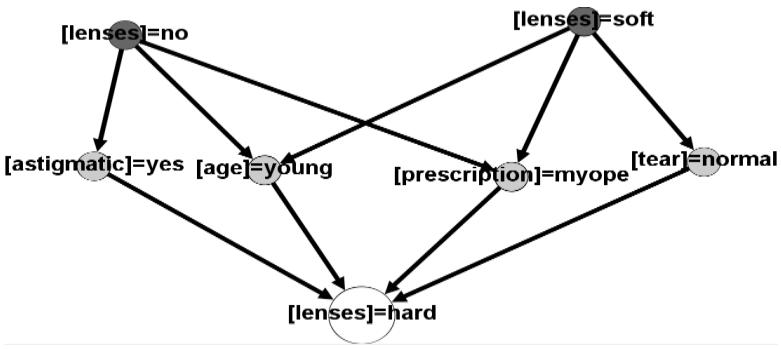
Figura 5 – ARN filtrada pela medida de Added Value com "lenses=soft" como item objetivo



Fonte: Elaborada pelo autor.

O mesmo ocorre quando analisa-se a AV-ARN com "lenses=hard" como item objetivo (Figura 6). O número de nós diminuiu, pois foram retiradas as regras que não possuem dependência estatística, provocando assim, uma melhor visualização da informação com o aumento da possibilidade de leitura do hipergrafo gerado,e por consequência, facilitando a extração do conhecimento pela formulação de hipóteses com maiores probabilidades de serem verdadeiras.

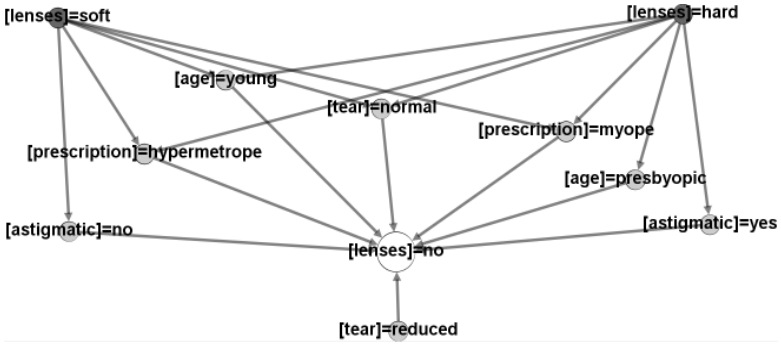
Figura 6 – ARN filtrada pela medida de Added Value com "lenses=hard" como item objetivo



Fonte: Elaborada pelo autor.

Nas AV-ARNs anteriores percebeu-se alteração no mais alto nível da rede, o que poderia ser um empecilho para a comprovação da eficiência da técnica proposta, porém, na AV-ARN com "lenses=no" como item objetivo (Figura 7) houve a eliminação de um nó com nível L = 1 ([age]=pre-presbyopic), mesmo existindo nós de nível maior. A seleção das regras ocorre em nós de todos os níveis, sendo o fator mais importante a influência estatística gerada pela métrica Added Value. Sendo assim, extração do conhecimento ocorre de modo mais seguro quanto a confiabilidade da influência dos elementos envolvidos em cada regra.

Figura 7 – ARN filtrada pela medida de Added Value com "lenses=no" como item objetivo

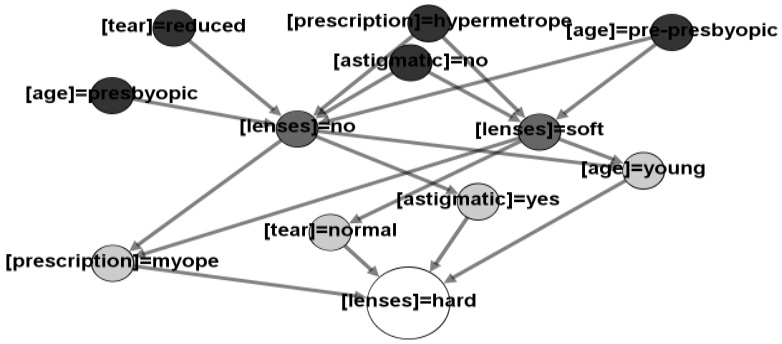


Fonte: Elaborada pelo autor.

Após a análise das AV-ARNs formuladas, foi estruturado mais um experimento, no qual a rede seria construída apenas com as regras selecionadas pelo filtro Added Value sem a utilização de um parâmetro de minconf, ou seja, considerando-se minconf = 0 e selecionando o item-objetivo "lenses=hard" (Figura 8).

Analisando a Figura 8, infere-se que a medida Added Value aumentou a confiabilidade das informações, já que surgiu um nível a mais, ao se comparar com a Figura 6. Os nós deste nível estão conectados àqueles formados pelos itens que formam as classes derivadas do dataset, ou seja, pode-se, mesmo com um único alvo, perceber também as influências nas outras possibilidades de resultado para aquela variável objetivo, embora isto não seja uma garantia para outros datasets. No exemplo, "[lenses]=hard" é o item objetivo, mas os nós "[age]=presbyopic"

Figura 8 – ARN com minconf = 0 e filtrada pela medida de Added Value com "lenses=hard" como item objetivo



Fonte: Elaborada pelo autor.

e "[tear]=reduced" se conectam diretamente a "[lenses]=no", indicando que estes parâmetros influenciam muito mais para que uma pessoa não use lente, do que o uso de uma lente rígida, como é observado na ARN (Figura 3).

2.2.2 ARN com filtro Certainty Factor (CF-ARN)

A medida objetiva Certainty Factor (CF) foi elaborada a fim de se mensurar o impacto de LHS em RHS, ou seja, o quanto o antecessor (LHS) influencia na presença do respectivo RHS. As regras foram eliminadas conforme o critério apresentado na Tabela 1.

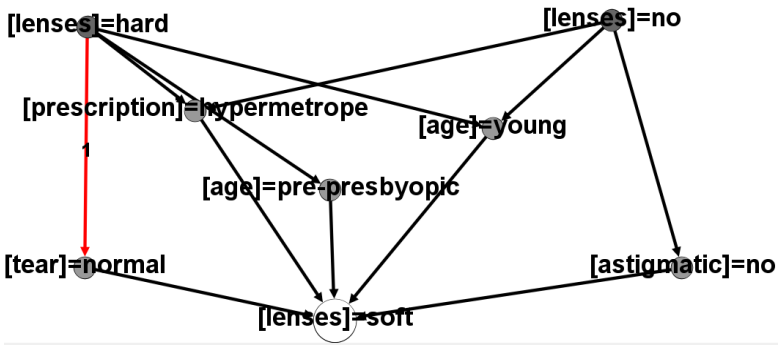
Na construção das Redes foram destacadas as arestas formadas pelas regras que possuíam CF = 1 e CF = -1, por se caracterizarem uma dependência estatística positiva e negativa, respectivamente, favorecendo assim a formulação de hipóteses com maiores probabilidades de serem verdadeiras a respeito da influência de cada característica dentro da base de dados estudada, pois consegue-se detectar o impacto real entre os itemsets das regras.

Observando a Figura 9 percebe-se uma conexão de dependência direta entre "[lenses = hard]" e "[tear] = normal", e ao comparar a CF-ARN, ARN com filtro CF, com a ARN (Figura 2), percebe-se que o número de regras diminuiu, decrementando a granularidade da rede e favorecendo a visualização da informação. Restaram na rede apenas as regras (arestas) com verdadeira relevância estatística de influência avaliada pela medida CF.

Quando analisa-se a rede com "lenses=hard" como item objetivo (Figura 10), a aresta que representa um impacto de LHS em RHS maior é formada pela regra "[lenses]=soft" ⇒ "[tear] = normal". Quando comparamos com a ARN (Figura 3), percebe-se uma queda considerável do número de nós de segundo nível da rede (L=2), o que demonstra a capacidade de filtragem da medida CF no que diz respeito a seleção de regras.

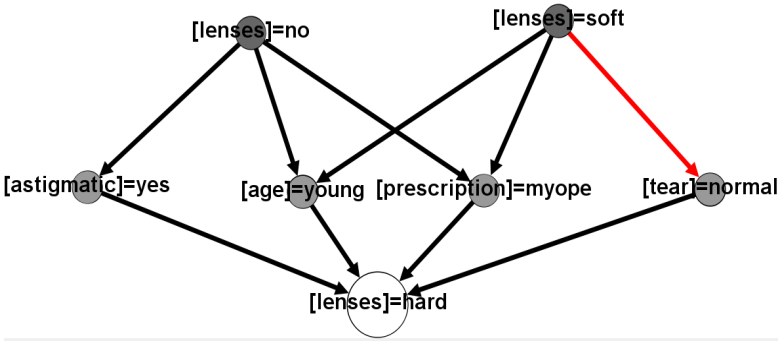
Na Figura 11 observa-se uma rede com 1 (um) nó a menos no nível um (L=1) que a ARN (Figura 4), demonstrando a potencialidade da medida CF. Outro fator importante é a presença de arestas com CF = 1 e CF = -1, o que gera uma hipótese de dependência total entre os

Figura 9 – ARN filtrada pela medida de CF com "lenses=soft" como item objetivo



Fonte: Elaborada pelo autor.

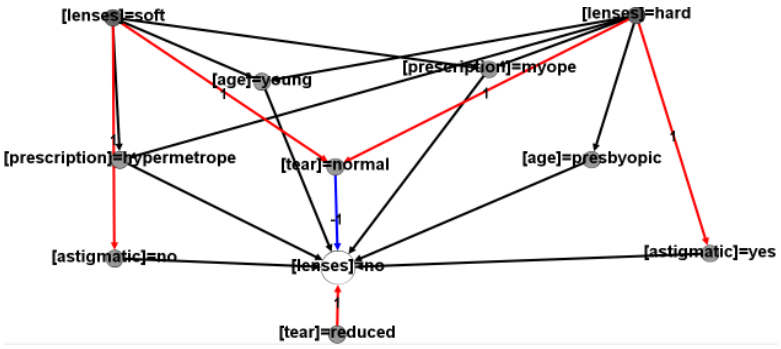
Figura 10 – ARN filtrada pela medida de CF com "lenses=hard" como item objetivo



Fonte: Elaborada pelo autor.

elementos da regra. O nó "[tear] = normal" se conecta ao nó objetivo por meio de uma influência inversa de dependência, i.e. inversamente proporcional ao nó objetivo. Se considerarmos que o nó "[tear] = reduced" conecta-se com dependência direta, pode-se concluir que a característica de produção de lagrimas (tear) é um fator extremamente relevante para a seleção do tipo de lente de contato. Elabora-se então a hipótese: "quanto mais normal for a produção de lágrimas, menor a necessidade do uso de lentes, bem como, se a produção de lágrimas for reduzida, o uso de lentes é totalmente recomendado".

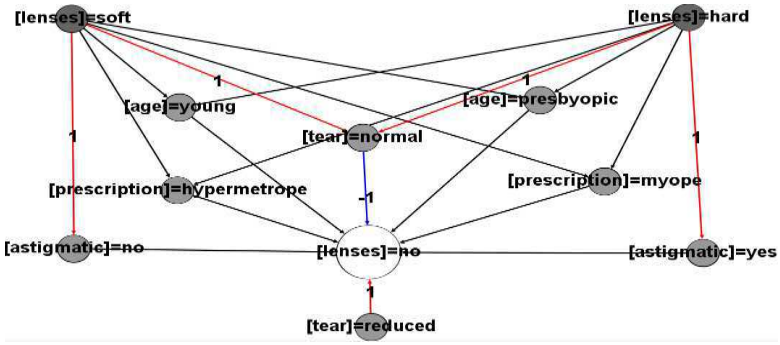
Figura 11 – ARN filtrada pela medida de CF com "lenses=no" como item objetivo



Fonte: Elaborada pelo autor.

Foi elaborado mais um experimento com o uso apenas da medida Certainty Factor como parâmetro de seleção das regras e o resultado obtido encontra-se na Figura 12. Nesta rede percebe-se os mesmos padrões já encontrados em redes anteriores, porém com o impacto do conhecimento de fatores de dependências positivos e negativos similares à rede da Figura 11. Pode-se então depreender que a medida de confiança perde totalmente a influência quando se utiliza a medida CF para a seleção das regras e construção das redes.

Figura 12 – ARN com minconf = 0 e filtrada pela medida de Certainty Factor com "lenses=no" como item objetivo



Fonte: Elaborada pelo autor.

2.2.3 ARN com Filtro Convicção (Conv-ARN)

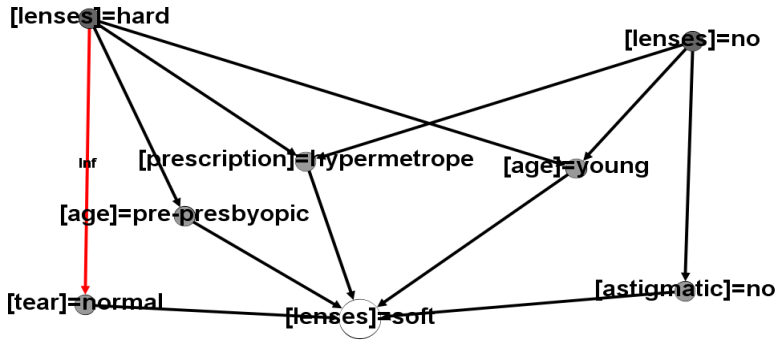
Convicção é uma medida intuitivamente derivada da Lift, porém com uma característica de assimetria. Utiliza-se conceitos de lógica proposicional para gerar uma medida capaz de substituir a confiança, já que esta não leva em consideração o sup(LHS). o Filtro Convicção fez a exclusão das regras de associação com valor Conv = 1 (Tabela 1).

A implicação lógica entre os elementos da regra são identificados quando o valor de Conv = ∞, portanto foi feito um destaque nas arestas que representam esta situação, tornando a observação mais objetiva e eficiente. Não foi possível nenhum outro tipo de comparação entre as regras pois a medida de Convicção possui uma grande desvantagem de estar em um intervalo ilimitado de valores (BERZAL et al., 2002).

Na Figura 13 é mostrada a Conv-ARN, ARN com filtro de Convicção, construída com "lenses=soft" como item objetivo. Observa-se uma diminuição da granularidade da rede ao se comparar com a rede da Figura 2, gerando uma melhor observação das informações, bem como, uma melhor geração de hipóteses. A aresta entre "[lenses = hard]" e "[tear] = normal" foi destacada indicando uma implicação lógica entre estes elementos.

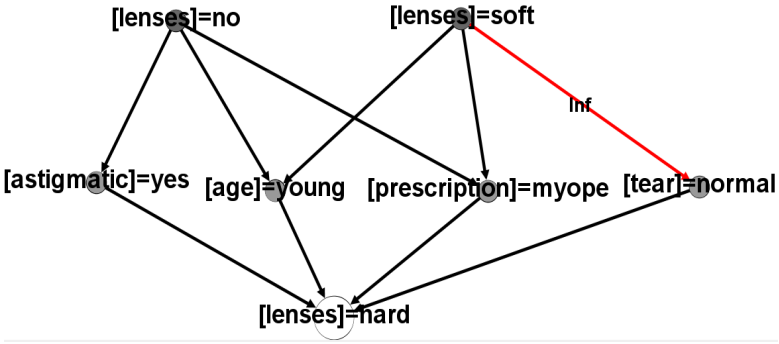
Também houve diminuição da densidade da rede quando utilizamos "lenses=hard" como item objetivo (Figura 14), tornando a mesma de mais fácil entendimento ao eliminar 5 (cinco) nós de nível dois (L = 2) por fazerem parte de regras eliminadas pelo filtro Convicção. Uma regra foi destacada indicando implicação lógica por se tratar de um valor Conv = ∞.

Figura 13 – ARN filtrada pela medida de Convicção com "lenses=soft" como item objetivo



Fonte: Elaborada pelo autor.

Figura 14 – ARN filtrada pela medida de Convicção com "lenses=hard" como item objetivo



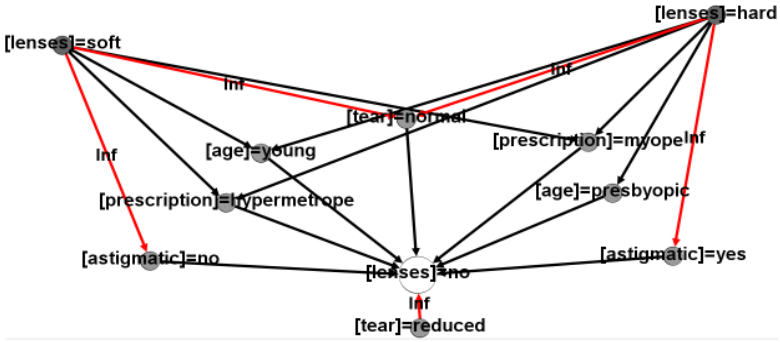
Fonte: Elaborada pelo autor.

Quando compara-se a ARN com "lenses=no" como item objetivo (Figura 4) com a que recebeu as regras filtradas pela medida Convicção (Figura 15) observa-se que houve a retirada de 1 (um) nó de nível um (1), ou seja, um elemento que estava diretamente ligado ao item objetivo, mas que não produzia nenhum tipo de influência estatística, podendo-se elaborar uma hipótese com grande possibilidade de ser falsa, o que acarretaria em um conhecimento equivocado.

Foram destacadas 5 (cinco) arestas indicativas de implicação lógica. Quando comparada a ARN formuladas com as regras filtradas pela medida CF (Figura 11), percebe-se a incapacidade da medida de convicção indicar implicações lógicas de dependência inversa.

Efetuando-se mais um experimento selecionando regras de associação apenas pelo uso da medida de Convicção com o filtro estabelecido anteriormente, elaborou-se a Conv-ARN da Figura 16 com "lenses=soft" como item objetivo. Comparando esta rede com a ARN (Figura 2) percebe-se o surgimento de mais um nível na rede, e com ele a informação de uma implicação lógica (Conv = ∞) entre "[tear]=reduced" e "[lenses]=soft" o que trás um ganho de conhecimento para o comportamento de uma regra para outro elemento que não seja o objetivo, porém, um fator negativo nesta rede é o aparecimento de outras arestas com valores pequenos da medida de confiança, o que torna a rede não muito confiável em níveis mais baixos. Assim, infere-se a possibilidade de que a aliança do uso de mais uma medida provocaria um resultado melhor para

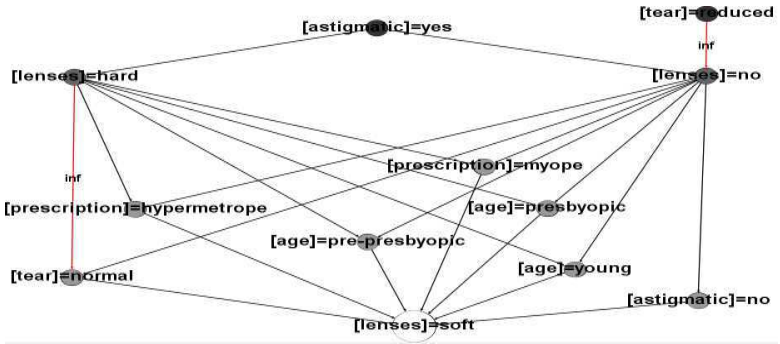
Figura 15 – ARN filtrada pela medida de Convicção com "lenses=no" como item objetivo



Fonte: Elaborada pelo autor.

a geração automática de hipóteses.

Figura 16 – ARN com minconf = 0 e filtrada pela medida de Convicção com "lenses=soft" como item objetivo



Fonte: Elaborada pelo autor.

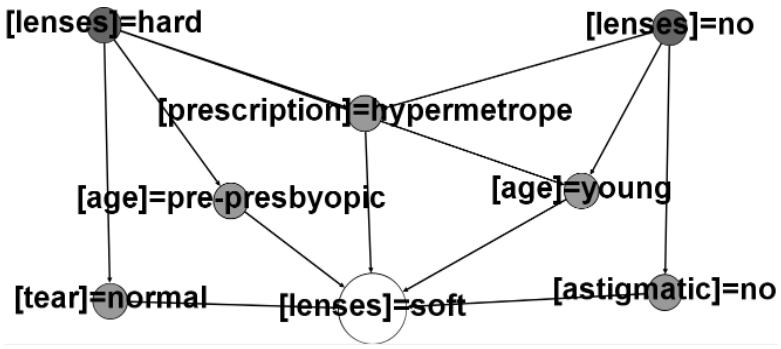
2.2.4 ARN com outro Gini Index (GI-ARN)

Gini Index (GI) é uma medida que relaciona uma variável meta com uma preditora, portanto, uma grande candidata ao uso de redes direcionadas como as ARNs. Fürnkranz e Flach (2005) provaram que as medidas Gini são equivalentes à precisão, no sentido de que elas dão rankings idênticos ou inversos para qualquer conjunto de regras. Tan e Kumar (2000) afirma que os valores variam de GI = 0, quando não há correlação estatística nenhuma, ou seja, as variáveis são independentes, até GI = 0,5, quando há uma correlação perfeita. Sendo assim, o filtro GI realizou a exclusão de todas as regras de associação que possuíam Gi nulo (Tabela 1).

Na GI-ARN, ARN com filtro GI, da Figura 17, pode-se observar uma diminuição da densidade da rede ao compará-la com a ARN (Figura 2) o que proporciona uma melhor visualização das informações e, por conseguinte, uma melhor formulação de hipóteses. Não foi encontrada nenhuma regra com GI = 0,5 (valor máximo).

Ao compararmos a rede que recebeu o filtro GI (Figura 18) e a ARN (Figura 3) que

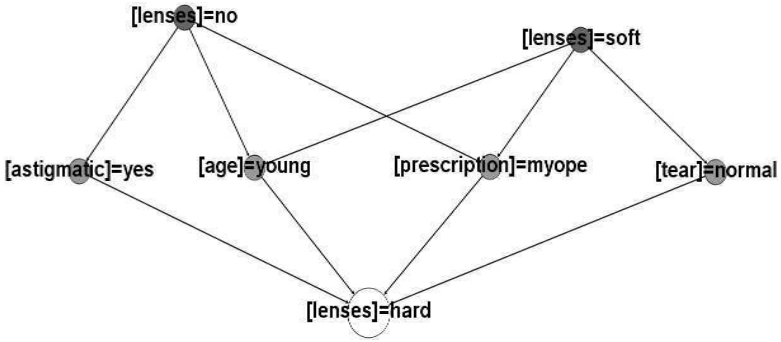
Figura 17 – ARN filtrada pela medida de Gini Index com "lenses=soft" como item objetivo



Fonte: Elaborada pelo autor.

possuem "lenses=hard" como item objetivo, percebe-se o mesmo ganho no decréscimo da granularidade da rede, mesmo que não tenha ocorrido a presença de regras com correlação perfeita (GI = 0,5).

Figura 18 – ARN filtrada pela medida de Gini Index com "lenses=hard" como item objetivo

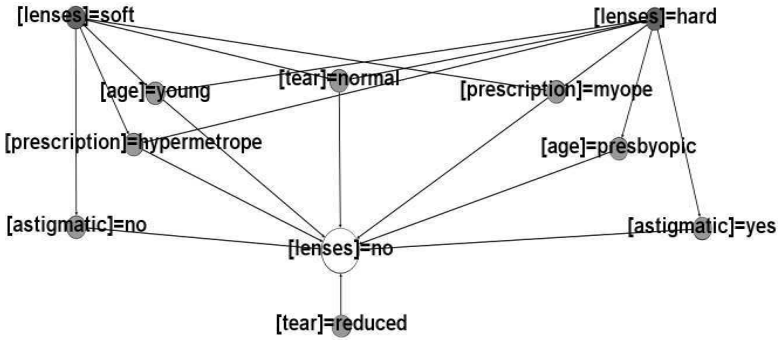


Fonte: Elaborada pelo autor.

Quando analisa-se as regras que formam a rede filtrada pela medida GI com "lenses=no" como item objetivo (Figura 19), é encontrado o mesmo valor para as regras "[tear]=normal" ⇒ "[lenses]=no" e "[tear]=reduced" ⇒ "[lenses]=no". Desta forma, pode-se perceber que ambas possuem o mesmo grau de importância quando se utiliza essa medida de interesse, porém sem indicação do sentido da influência (direta ou inversa). Neste caso, o uso de outra métrica (CF, por exemplo), poderia indicar estas correlações com mais facilidade. A medida GI realizou a filtragem em 1 (um) nó do nível um (L = 1) da rede, o que indica o potencial desta medida para seleção de regras de interesse.

Após a análise das ARNs filtradas com a medida Gini Index, foi executado mais um experimento, no qual a rede é construída apenas com as regras selecionadas pelo filtro Gini Index sem a utilização de um parâmetro de minconf, ou seja, atribuindo-se minconf = 0. Optou-se para utilizar como item-objetivo "lenses=hard" (Figura 20). Comparando esta GI-ARN com a ARN (Figura 3) percebe-se uma perda no detalhamento devido ao aumento na densidade da rede, e por consequência, piorando a leitura do grafo e gerando mais dificuldades para a formulação de

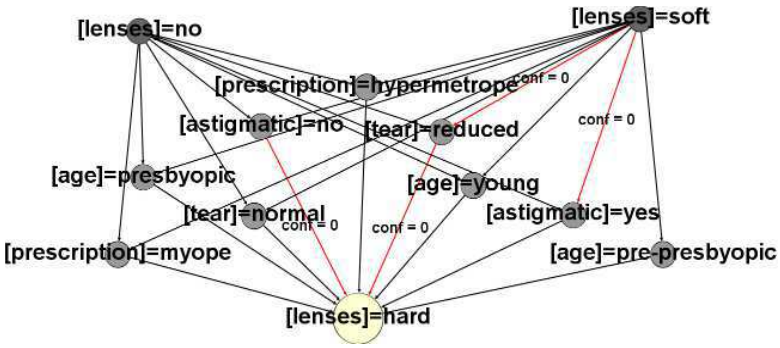
Figura 19 – ARN filtrada pela medida de Gini Index com "lenses=no" como item objetivo



Fonte: Elaborada pelo autor.

hipóteses. Além disso, foram geradas arestas com confiança = 0, que foram destacadas na rede, indicando que o uso apenas da Medida GI não deve ser implementado, pois ocasionaria em erros no estudo dos dados. Desta forma, a medida de Gini Index deve ser utilizada como complemento ao estudo das regras e não apenas como única medida de seleção das mesmas.

Figura 20 – ARN com minconf = 0 e filtrada pela medida de Gini Index com "lenses=hard" como item objetivo



Fonte: Elaborada pelo autor.

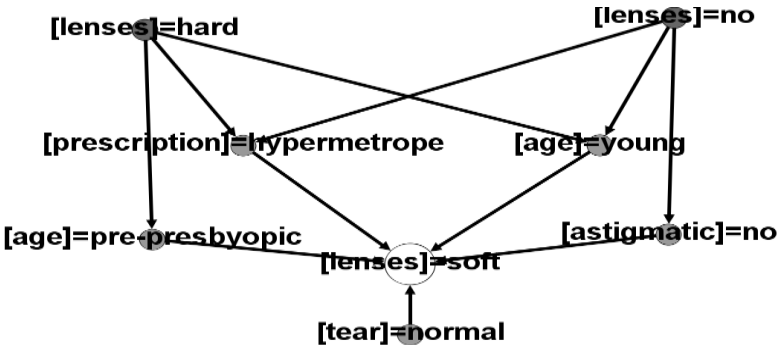
2.2.5 ARN com Filtro J-Measure (J-ARN)

A medida J-Measure é baseada em conceitos da teoria da informação (GENG; HAMILTON, 2006) indicando um estudo mais detalhado das influências dos conteúdos de cada variável envolvida. A Equação J-Measure é formulada por dois termos, sendo que o primeiro indica a generalidade da regra, e o segundo indica a discriminação, vinculando o valor agregado à presença de RHS e LHS em conjunto. Nesse caso, mensura-se o impacto de LHS em RHS indicando que a soma dos conteúdos da informação (de um conjunto de regras com LHS exclusivos e excludentes) deve ser igual à conhecida informação mútua comum entre as duas variáveis (GOODMAN; SMYTH, 1991). O Filtro J (Tabela 1) efetuou a eliminação das regras de associação com J = 0 e quando não foi encontrado um valor numérico para a mesma (J = NaN ou Not a Number).

As regras com J = 1, i.e. quando RHS é encontrado apenas na presença de LHS, receberam destaque para otimização da extração do conhecimento.

Após a seleção das regras pela confiança e pelo filtro J-Measure, foi selecionado como item objetivo "lenses=soft" (Figura 21). Analisando-se o grafo, pode-se inferir que a granularidade da rede teve um alto decréscimo quando comparada com a ARN (Figura 2) permanecendo apenas arestas com valores da Medida J no intervalo]0;1]. Uma das principais mudanças ocorreu com o nó "[tear] = normal" que, embora tenha permanecido no nível um (L=1), ficou sem nenhum nó antecessor, o que indica que a informação por ele gerada é apenas influenciadora, e não recebe influência de nenhum outro elemento.

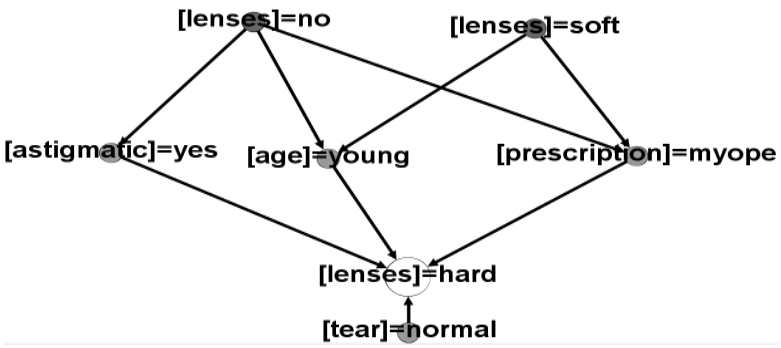
Figura 21 – ARN filtrada pela medida de J-Measure com "lenses=soft" como item objetivo



Fonte: Elaborada pelo autor.

Foi elaborada uma J-ARN, ARN com filtro J-Measure, com "lenses=hard" como item objetivo (Figura 22). Ao compará-la com a ARN de mesmo item objetivo (Figura 3) mensura-se uma granularidade menor da rede o que possibilita uma melhor formulação de hipóteses, já que a informação está sendo previamente avaliada pelo cálculo de J-Measure. O número de nós no nível 1 (L = 1) permaneceu inalterado, mas o nó "[tear] = normal" surgiu sem antecessor, indicando que a produção normal de lágrimas não recebe influência de nenhuma outra informação.

Figura 22 – ARN filtrada pela medida de J-Measure com "lenses=hard" como item objetivo

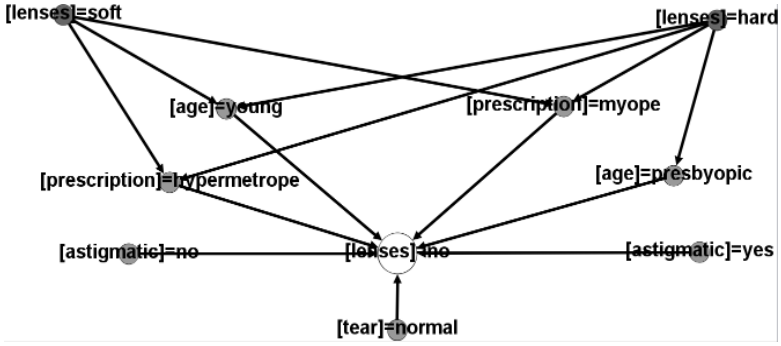


Fonte: Elaborada pelo autor.

Observando a J-ARN da Figura 23 com "lenses=no" como item objetivo e comparando com a ARN (Figura 4), infere-se que a densidade da rede ficou menor, pois algumas arestas

foram excluídas pelo filtro e nós como "[tear] = normal", "[astigmatic] = no" e "[astigmatic] = yes" ficaram sem antecessor. Além disso, um dos nós de maior destaque "[tear] = reduced" em ARNs filtradas anteriormente foi excluído, gerando uma análise negativa a respeito do uso desta métrica em uma metodologia de construção de ARNs

Figura 23 – ARN filtrada pela medida de J-Measure com "lenses=no" como item objetivo



Fonte: Elaborada pelo autor.

Foi executado mais um experimento, no qual foi utilizada apenas a seleção das regras pelo uso do filtro J-Measure, ou seja, não foi utilizada a seleção pela medida de confiança (minconf = 0). Como resultado, foi gerado o grafo da Figura 24 que, ao compará-lo com a ARN da Figura 4 pode-se perceber diferenças consideráveis entre as mesmas, como por exemplo, a menor densidade da rede pelo decréscimo do número de arestas, mas com o aparecimento de mais um nível na rede, pois o nó "[age] = pre-presbyopic" subiu do nível um (L = 1) para o nível três (L = 3) da rede.

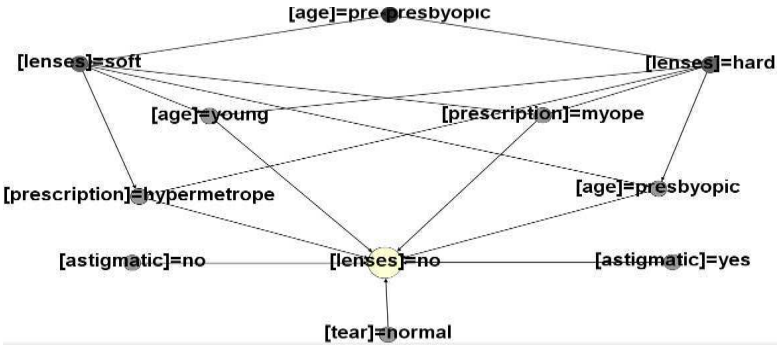
Outra alteração foi o desaparecimento do nó "[tear] = reduced", bem como das ligações dos antecessores de "[tear] = normal", "[astigmatic] = no" e "[astigmatic] = yes", essas informações eram sempre presentes, mesmo com o uso de filtros de outras medidas, levando-se a necessidade de um estudo mais aprofundado para comparação entre as mesmas. O não aparecimento destas arestas demonstra que a medida J não é interessante para ser utilizada na construção de ARNS, pois não consegue descrever relações importantes entre os itemsets analisados.

2.2.6 ARN com filtro Laplace (Laplace-ARN)

A medida de Laplace (Lap) é uma das mais utilizadas para selecionar pares atributo-valor (GENG; HAMILTON, 2006). Conceitualmente, a métrica de Laplace é tendenciosa para regras mais gerais com maior precisão preditiva do que a confiança. Uma das propriedades da medida de Laplace é a capacidade de seleção de regras de maior grau de interesse, uma vez que só precisa-se verificar o conjunto de regras com valores maiores que minsup e minconf. O filtro Lap foi executado com a exclusão das regras com valor nulo conforme a Tabela 1.

Selecionando o filtro Lap e "lenses=soft" como item objetivo, foi construída a Laplace-ARN, ARN com filtro Laplace, da Figura 25, que quando comparada com a ARN (Figura 2) não

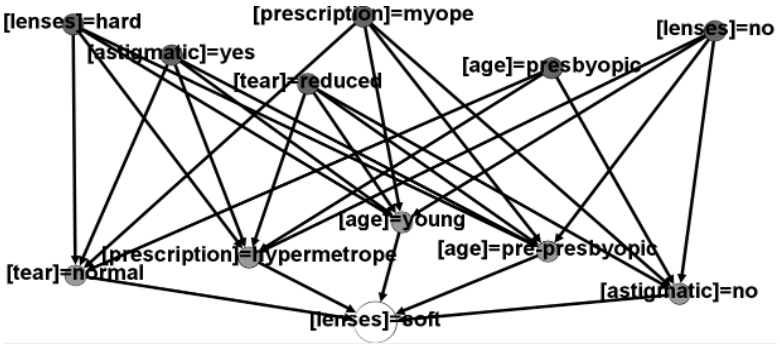
Figura 24 – ARN com minconf = 0 e filtrada pela medida de J-Measure com "lenses=no" como item objetivo



Fonte: Elaborada pelo autor.

demonstra nenhum tipo de alteração, pois o filtro dessa medida objetiva não excluiu nenhuma regra para o dataset estudado, i.e. não gerou valores de Lap = 0, o que ocasionou em uma manutenção de todos os nós e arestas da rede, não trazendo nenhum ganho para a extração do conhecimento por este tipo de filtragem.

Figura 25 – ARN filtrada pela medida de Laplace com "lenses=soft" como item objetivo

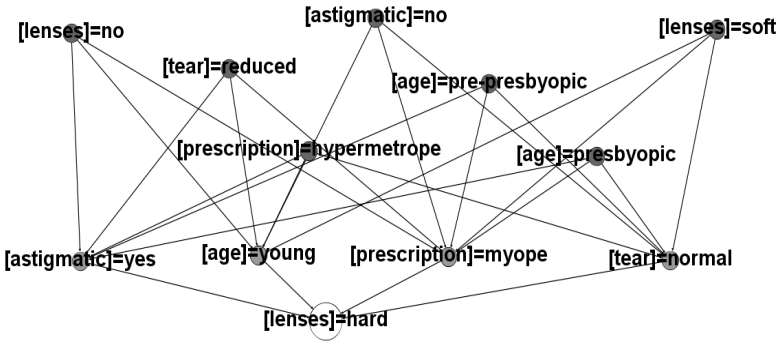


Fonte: Elaborada pelo autor.

O mesmo ocorre quando seleciona-se o filtro Lap e "lenses=hard" como item objetivo (Figura 26) e com "lenses=no" como item objetivo (Figura 27) quando comparadas com as respectivas ARNs (Figura 3 e Figura 4), ou seja, nenhuma alteração ocorreu na visualização da rede, pois não houveram regras com Lap = 0 e consequentemente, nenhuma regra foi excluída.

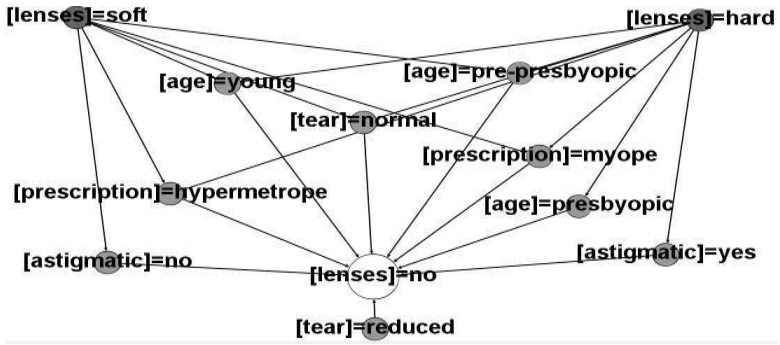
Para uma avaliação mais ampla da medida de Laplace, foi executado mais um experimento, no qual foi utilizado apenas o filtro Lap para seleção das regras, ou seja, não houve limitação da medida de confiança (minconf = 0). O resultado pode ser observado na Figura 28. Ao fazer a relação entre o grafo obtido nesse experimento com a ARN (Figura 3), percebe-se um aumento considerável na densidade da rede, pois foram acrescentadas 9 (nove) novas arestas, proporcionando uma maior dificuldade na elaboração de hipóteses, bem como, levando à conclusões equivocadas, pois o filtro da medida Laplace não conseguiu realizar a seleção das regras, sendo inseridas algumas que possuíam confiança = 0 (em destaque na Figura 3).

Figura 26 – ARN filtrada pela medida de Laplace com "lenses=hard" como item objetivo



Fonte: Elaborada pelo autor.

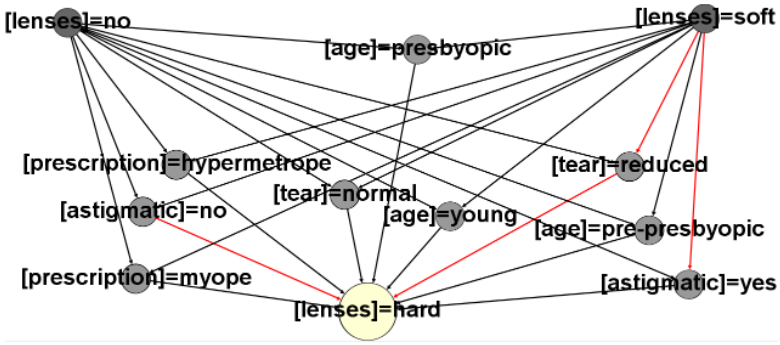
Figura 27 – ARN filtrada pela medida de Laplace com "lenses=no" como item objetivo



Fonte: Elaborada pelo autor.

Houve um aumento considerável no número de nós de nível um (L = 1), i.e. nós que estão diretamente conectados ao item objetivo, o que teoricamente indicariam uma maior relevância de informação, mas que não é traduzido na visualização da rede, pois trás equívocos de relacionamentos. Pode-se então perceber que a Medida Laplace poderá possuir importância se for utilizada como complemento e não como uso exclusivo na seleção das regras de associação para construção de novas ARNs.

Figura 28 – ARN com minconf = 0 e filtrada pela medida de Laplace com "lenses=hard" como item objetivo



Fonte: Elaborada pelo autor.

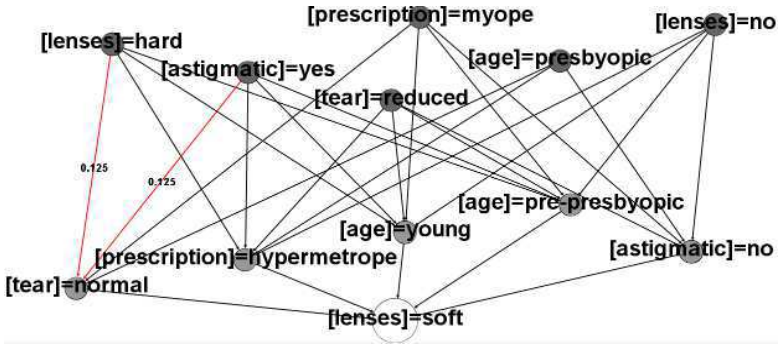
2.2.7 ARN com Filtro Gain (Gain-ARN)

A medida de ganho (Gain) determina o grau de interesse das regras de associação, realizando uma interpretação similar à medida de confiança, porém dando uma importância maior ao grau de influência de LHS sobre RHS, e ao mesmo tempo fazendo a seleção das regras de modo normalizado, na qual, valores de Gain = 0, indicam exatamente as regras que possuem a confiança igual ao valor de minconf atribuído na fase de extração (Tabela 1).

Estudando a medida Gain pode-se inferir que o seu uso realiza a seleção automática pela confiança, ao mesmo tempo que proporciona uma métrica do grau de relevância da influência de LHS sobre RHS, com isso, pode-se obter valores de Gain iguais, mesmo com confianças diferentes e vice-versa. Para cálculo da medida Gain, é necessária a definição de minconf.

Utilizando a medida Gain e atribuindo ao item objetivo "lenses=soft", foi construída a rede da Figura 29 e, comparando-a com a ARN (Figura 2) percebe-se que a rede manteve sua mesma estrutura de conexões, o que corrobora com a ideia que a medida Gain funciona de modo similar a confiança. Destaca-se apenas duas arestas para exemplificação de que regras com confianças diferentes podem possuir o mesmo valor de Gain. Nesse caso, as regras "[lenses]=hard" ⇒ "[tear]=normal" e "[astigmatic]=yes" ⇒ "[tear]=normal" possuem confianças 1 e 0,5, respectivamente, porém com o mesmo valor de Gain = 0,125.

Figura 29 – ARN filtrada pela medida de Gain com "lenses=soft" como item objetivo

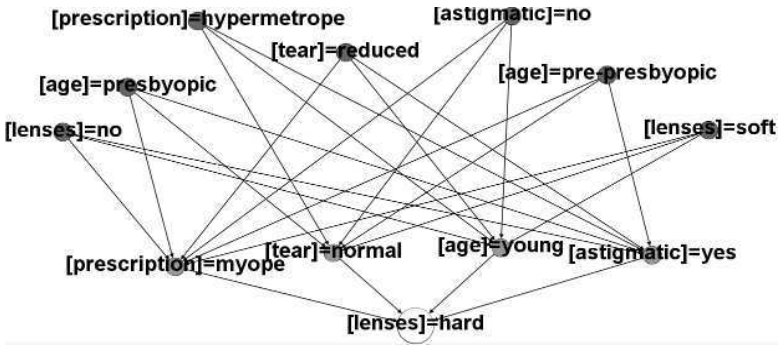


Fonte: Elaborada pelo autor.

Uma Gain-ARN, ARN com filtro Gain, com "lenses=hard" como item objetivo foi construída (Figura 30). Fazendo um paralelo com a ARN de mesmo item objetivo (Figura 3), observa-se uma rede equivalente em todos os aspectos, provando a semelhança entre as medidas Gain e confiança, bem como reforçando a possibilidade de substituição entre as mesmas.

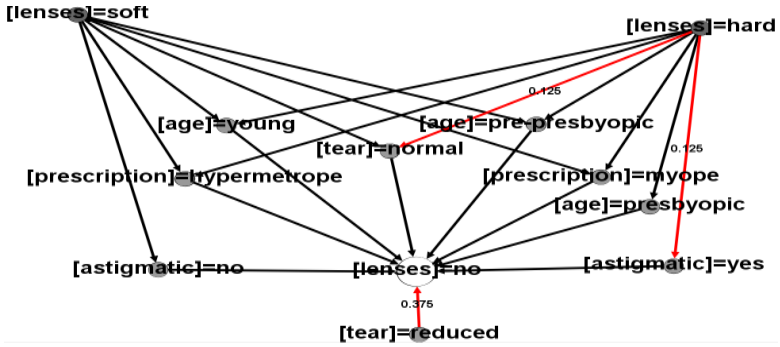
Ao analisar a rede com "lenses=no" como item objetivo que foi filtrada pela medida Gain (Figura 31) comparando-a com a ARN (Figura 4), percebe-se a mesma equivalência das anteriores. Destaca-se as arestas das regras "[lenses]=hard" ⇒ "[astigmatic]=yes", "[lenses]=hard" ⇒ "[tear]=normal" e "[tear]=reduced ⇒ [lenses]=no" como exemplos de regras que possuem o mesmo valor de confiança = 1 com valores de Gain diferentes, sendo 0,125 para as duas primeiras e 0,375 para a última, sendo portanto regras com graus de influência estatística diferentes.

Figura 30 – ARN filtrada pela medida de Gain com "lenses=hard" como item objetivo



Fonte: Elaborada pelo autor.

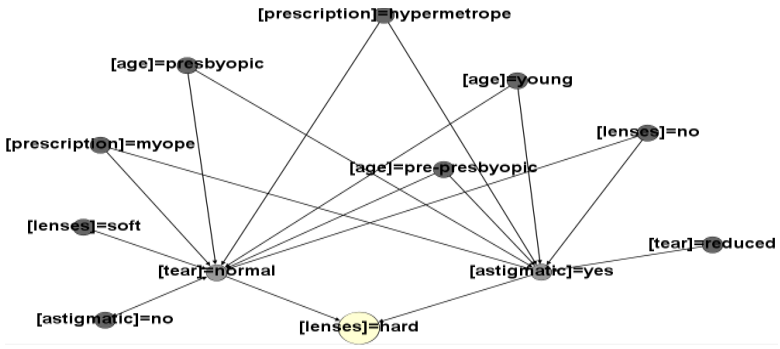
Figura 31 – ARN filtrada pela medida de Gain com "lenses=no" como item objetivo



Fonte: Elaborada pelo autor.

Para avaliação mais objetiva da medida Gain, efetuou-se um experimento, no qual a seleção das regras foi feita apenas por esta medida (minconf = 0), realizando um filtro com uma variável denominada de mingain (ganho mínimo) com o valor de 0,1, ou seja, este valor indica como 10% o grau mínimo de influência desejado entre os termos de uma regra. Foi estabelecido como item objetivo "lenses=hard" e elaborada a construção gráfica da rede (Figura 32).

Figura 32 – ARN com minconf = 0 e filtrada pela medida de Gain com "lenses=hard" como item objetivo



Fonte: Elaborada pelo autor.

Comparando a Gain-ARN com o uso exclusivo da medida Gain com a ARN (Figura

3), percebe-se uma estrutura totalmente diferente, com variações na quantidade de nós em cada camada da rede, bem como na distribuição das arestas, demonstrando a capacidade de prever conexões com maior influência que aquelas percebidas pelo uso exclusivo da medida de confiança. Houve uma melhoria na seleção das regras, bem como uma queda da granularidade da rede, possibilitando uma elaboração de hipóteses mais eficaz, já que foram selecionadas apenas as regras com maior índice de interesse.

2.2.8 Síntese

Após a análise do impacto das medidas objetivas assimétricas na construção das Redes de Regras de Associação, infere-se sobre a possibilidade do uso das mesmas para otimização da extração do conhecimento no processo de mineração de regras de associação. Em sua maioria, os filtros das medidas possibilitaram a formatação de redes com menor densidade e granularidade. Redes com menor número de arestas são mais fáceis de serem observadas otimizando a formação de hipóteses. As hipóteses são elaboradas sobre possíveis relações entre os itens de um dataset e, nesse estudo, direcionados a um item objetivo.

CAPÍTULO

3

CONSIDERAÇÕES FINAIS

Com o intuito de facilitar a extração do conhecimento com a elaboração de hipóteses com maior probabilidade de serem verdadeiras por meio do uso de grafos (redes), faz-se necessário reduzir o número de resultados (regras) extraídos,e para isso, muitas medidas de interesse podem ser utilizadas. Neste relatório técnico, foram avaliadas algumas medidas objetivas usadas na mineração de dados, dando enfoque especial às assimétricas devido à natureza direcionada das Redes de Regras de Associação (ARNs). Foram realizadas as descrições de cada medida, bem como a análise do impacto das mesmas na seleção das regras de associação que serão utilizadas para a construção de ARNs.

Com base na forma dos padrões gerados pelo método de mineração de dados, foram construídas ARNs pelo método de Pandey et al. (2009), e ARNs filtradas com uso das medidas selecionadas. As medidas objetivas são baseadas na teoria da probabilidade, nas estatísticas e na teoria da informação. Assim como, possuem princípios e fundamentos rígidos e suas propriedades podem ser analisadas e comparadas.

Foram examinadas as propriedades de medidas objetivas, o impacto na construção das ARNs filtradas e a influência na elaboração de hipóteses, por meio das variações da densidade e granularidade da rede. Como resultado, foi verificada uma melhoria da visualização das informações, bem como a consequente extração do conhecimento, pois as redes foram construídas apenas com regras que apresentavam dependências estatísticas entre os itens das regras(LHS e RHS).

Medidas objetivas possuem propriedades específicas que proporcionam análises diferenciadas das regras. Sendo assim, a escolha do filtro da medida é importante para obtenção de melhores resultados.

Medidas subjetivas e semânticas podem ser estudadas em trabalhos futuros, bem como o uso de métricas relacionadas diretamente às redes.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining Association Rules between Sets of Items in Large Databases. Special Interest Group on Management of Data, 22(2), v. 22, n. 2, p. 207–216, 1994. ISSN 01635808. Citado nas páginas 15 e 16.

AGRAWAL, R.; SRIKANT, R. 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, p. 487–499, 1994. Citado na página 16.

ARIDHI, S.; NGUIFO, E. M. Big Graph Mining: Frameworks and Techniques. Big Data Research, Elsevier Inc., v. 1, p. 1–10, 2016. ISSN 22145796. Disponível em: <http://dx.doi.org/10.1016/j.bdr.2016.07.002>. Citado na página 16.

BARALIS, E.; CAGLIERO, L.; MAHOTO, N.; FIORI, A. GraphSum: Discovering correlations among multiple terms for graph-based summarization. Information Sciences, Elsevier Inc., v. 249, p. 96–109, 2013. ISSN 00200255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2013.06.046>. Citado na página 16.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. Third International AAAI Conference on Weblogs and Social Media, p. 361–362, 2009. ISSN 14753898. Citado na página 21.

BERZAL, F.; BLANCO, I.; SÁNCHEZ, D.; VILA, M.-A. Measuring the accuracy and interest of association rules: A new framework. Intelligent Data Analysis, v. 6, p. 221–235, 2002. ISSN 1088467X. Citado na página 27.

BRIN, S.; MOTWANI, R.; ULMAN, J. D.; TSUR, S. Dynamic itemset counting and implication rules for market basket data. Proc. of the ACM SIGMOD Intl. Conf. on Management of Data, p. 255–264, 1997. Citado na página 22.

CENDROWSKA, J. PRISM: An algorithm for inducing modular rules. International Journal of Man-Machine Studies, v. 27, n. 4, p. 349–370, 1987. ISSN 00207373. Citado na página 20.

FISHER, D. Iterative Optimization and Simpli cation of Hierarchical Clusterings. Journal of Arti cial Intelligence Research, v. 4, p. 147–179, 1996. Citado na página 22.

FUKUDA, T.; MORIMOTO, Y.; MORISHITA, S.; TOKUYAMA, T. Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization. In: SIGMOD '96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data. [S.l.: s.n.], 1996. p. 13–23. Citado na página 22.

FÜRNKRANZ, J.; FLACH, P. A. ROC n' rule learning - Towards a better understanding of covering algorithms. Machine Learning, v. 58, n. 1, p. 39–77, 2005. ISSN 08856125. Citado na página 29.

GENG, L.; HAMILTON, H. J. Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys, v. 38, n. 3, p. 1–32, 2006. ISSN 03600300. Citado nas páginas 31 e 33.

GOODMAN, R. M.; SMYTH, P. Rule Induction Using Information Theory. Knowledge Discovery in Databases, n. 9, 1991. Citado nas páginas 22 e 31.

KAKKAD, U. K.; ALUVALU, R. Association rule mining using Apriori algorithm: a survey. IOSR Journal of Computer Engineering (IOSR-JCE), v. 16, n. 2, p. 112–118, 2013. Citado na página 16.

LIU, X.; ZHAI, K.; PEDRYCZ, W. An improved association rules mining method. Expert Systems with Applications, Elsevier Ltd, v. 39, n. 1, p. 1362–1374, 2012. ISSN 09574174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2011.08.018>. Citado na página 16.

NEWMAN, M. Networks: An introduction. The Journal of analytical psychology, v. 55, p. 617–635, 2010. ISSN 00218774. Citado na página 16.

NGUYEN, K. N. T.; CERF, L.; PLANTEVIT, M.; BOULICAUT, J. F. Discovering descriptive rules in relational dynamic graphs. Intelligent Data Analysis, v. 17, n. 1, p. 49–69, 2013. ISSN 1088467X. Citado na página 16.

PANDEY, G.; CHAWLA, S.; POON, S.; ARUNASALAM, B.; DAVIS, J. G. Association Rules Network: Definition and Applications Gaurav. Statistical Analysis and Data Mining, v. 1, n. 4, p. 260–179, 2009. ISSN 19321864. Citado nas páginas 17, 20, 22 e 39.

SAHAR, S. What Is Interesting: Studies on Interestingness in Knowledge Discovery. 200 p. Tese (Phd Thes) — Tel-Aviv University The, 2003. Citado na página 22.

SHORTLIFFE, E. H.; BUCHANAN, B. G. A Model of Inexact Reasoning in Medicine. Mathematical Biosciences, v. 23, p. 351–379, 1975. Citado na página 22.

SMALDON, J.; FREITAS, A. A. A New Version of the Ant-Miner Algorithm Discovering Unordered Rule Sets. In: Proceedings of the 8th annual conference on Genetic and evolutionary computation GECCO 06. [s.n.], 2006. p. 43. ISBN 1595931864. Disponível em: <http://kar.kent.ac.uk/14462/>. Citado na página 22.

TAN, P.-n.; KUMAR, V. Interestingness Measures for Association Patters: A Perspective. In: KDD Workshop on Postprocessing in Machine Learning and Data Mining. [S.l.: s.n.], 2000. v. 6, n. 0, p. 1–9. Citado na página 29.

WENG, C. H. Identifying association rules of specific later-marketed products. Applied Soft Computing Journal, Elsevier B.V., v. 38, p. 518–529, 2016. ISSN 15684946. Citado nas páginas 15 e 16.

ZANIN, M.; PAPO, D.; SOUSA, P. A.; MENASALVAS, E.; NICCHI, A.; KUBIK, E.; BOCCALETTI, S. Combining complex networks and data mining: Why and how. Physics Reports, Elsevier B.V., v. 635, p. 1–44, 2016. ISSN 03701573. Disponível em: <http://dx.doi.org/10.1016/j.physrep.2016.04.005>. Citado na página 15.