Asymmetric objective measures applied to filter Association Rules Networks

Dario Brito Calçada, Renan de Padua, Solange Oliveira Rezende
Institute of Mathematics and Computer Science
Universidade de São Paulo
São Carlos, Brasil
Email: dariobcalcada@usp.br, solange@icmc.usp.br

***Resumo*—In this paper, the Filtered-Association Rules Network (*Filtered*-ARN) is presented to structure, prune, and analyze a set of association rules to construct candidate hypotheses. The *Filtered*-ARN algorithm selects association rules with the use of asymmetric objective measures, *Added Value* and *Gain*, then builds a network allowing more exploration information. The *Filtered*-ARN was validated using three datasets: Lenses and Soybean Large, both available online for a text and a real dataset with data on organic fertilization (*Green Manure*). The results were validated by comparing the *Filtered*-ARN with the conventional ARN and also comparing the results with the decision tree. The approach presented promising results, showing its ability to explain a set of objective items and the aid to build more consolidated hypotheses by guaranteeing statistical dependence with the use of objective measures.**

***Index Terms*—Association rules, Networks, Association rules networks, Data mining, Graphs, Objective measures**

I. Introdução

A mineração de dados é frequentemente descrita como o processo de descobrir padrões "interessantes" em grandes bancos de dados [1]. Dada a grande quantidade de dados digitais que estão sendo constantemente gerados e armazenados, a mineração de dados oferece uma solução para o problema de rapidamente resumir e buscar relacionamentos não óbvios dados.

Mineração de dados pode ser usada como uma metodologia para descobrir hipóteses ou teorias candidatas. Essa abordagem permite explorar avanços recentes em técnicas de pesquisa computacionalmente eficientes em conjunto com métodos estatísticos tradicionais, que continuam sendo a base da verificação e validação de teorias [2].

O ponto de partida do processo de mineração vem das observações (eventos) que acionam o pesquisador para acelerar os estudos conceituais e chegar a uma estrutura na qual o processo subjacente (que está gerando os eventos) pode ser elucidado. A mineração de regras de associação [3]–[5] é usada a fim de encontrar padrões interessantes na forma de regras A $\Rightarrow$ B, na qual A e B podem ser atributos, itens ou mais geralmente "objetos de dados". Considerando-se que fosse conhecido de antemão que A e B estão correlacionados, em um sentido estatístico, então a descoberta da regra A $\Rightarrow$ B apenas confirma o conhecimento prévio e não apresenta informação nova. Por outro lado, caso nunca foi identificada a correlação entre A e B, a descoberta da regra A $\Rightarrow$ B sugere que A e B sejam pares candidatos a serem validados estatisticamente (para correlação).

Como os conjuntos de dados geralmente estão aumentando, tanto em quantidade quanto em dimensionalidade, enumerar todas as combinações possíveis de A e B e depois verificar sua correlação não é computacionalmente viável. Assim, a mineração de regras de associação pode ser vista como um mecanismo para oferecer teorias candidatas para serem validadas. Neste artigo, é apresentado um método de mineração de regras de associação que utiliza medidas objetivas aliadas a uma estrutura de rede para otimização da formação de hipóteses.

Algoritmos que descobrem regras de associação usam medidas de interesse que são capazes de avaliar a qualidade de uma regra. Entre essas medidas, o suporte e a confiança se destacam, embora *lift*, *gain*, *certainty factor*, *added value* ou *leverage* também sejam indicadores que fornecem informações úteis sobre as regras extraídas [6].

Como os algoritmos de regra de associação são capazes de extrair todas as regras de associação de acordo com um suporte mínimo e valor mínimo de confiança, o número de regras extraídas geralmente supera a capacidade de exploração do usuário. Várias abordagens foram propostas para guiar o usuário na exploração de regras. No entanto, a grande maioria dessas abordagens se concentra na exploração de acordo com a regra inteira e não considerando explorações que possam se concentrar em um pequeno conjunto de itens ou em um item alvo.

Por outro lado, para facilitar a extração do conhecimento, muitos processos de mineração utilizam técnicas de redes para visualização dos dados [7]. Aliando medidas objetivas com estrutura de rede para visualização, é proposta neste artigo a *Filtered*-Association Rules Network (*Filtered*-ARN). A *Filtered*-ARN usa medidas objetivas assimétricas com a Association Rules Network (ARN) proposta por Pandey [8] para estruturar e auxiliar na análise das regras extraídas em um *dataset*.

Ao utilizar medidas objetivas assimétricas, a *Filtered*-ARN permite a exploração mais concisa das regras por selecionar apenas aquelas em que o antecessor da regra influenciam estatisticamente o sucessor. A *Filtered*-ARN expande os recursos presentes na ARN, adicionando algumas propriedades como,

cálculo de influência estatística e uso de uma medida de ganho entre os elementos das regras.

A estrutura da rede é a mesma que a ARN, porém a seleção de regras com influência estatística comprovada promove a identificação de hipóteses com maior probabilidade de serem verdadeiras.

O objetivo central da *Filtered*-ARN é apresentar um grafo com regras de associação que possuem uma maior chance de serem interessantes para a análise realizada pelo especialista do domínio. A principal diferença entre a *Filtered*-ARN e a ARN convencional é que, na *Filtered*-ARN, o usuário pode visualizar um conjunto de itens que promovem uma influência estatística em vez de elementos que apenas se relacionam com o item objetivo. Ao permitir isso, a *Filtered*-ARN apresenta regras que indicam hipóteses com comprovação de dependência entre antecedente e consequente.

A fim de validar as *Filtered*-ARN, foram realizados 3 estudos de caso e os resultados foram comparados com a ARN convencional e com um algoritmo de árvore de decisão, pois os mesmos podem ser utilizados para visualização de graus de dependência entre elementos de um *dataset*. Os resultados demonstraram que a *Filtered*-ARN é capaz de descrever os elementos que influenciam o item objetivo de modo mais conciso se comparado a ARN, permitindo que o usuário observe os casos em que um item interfere estatisticamente em um item objetivo. Além disso, a *Filtered*-ARN foi comparada a um algoritmo de árvore de decisão, com o objetivo de analisar a explicação dos dados. Tal comparação demonstra que a *Filtered*-ARN apresenta uma estrutura mais eficiente que a árvore de decisão, tornando mais fácil para o usuário entender o conhecimento extraído.

O artigo está organizado da seguinte forma. Na Seção II é fornecida uma visão geral da mineração de regras de associação, apresentando a definição e algumas medidas objetivas conhecidas na literatura, além da definição de ARN. Na Seção III são apresentadas algumas pesquisas que inspiraram a proposta deste artigo. A *Filtered*-ARN é apresentada com todas as suas definições e princípios na Seção IV. Na Seção V os estudos de caso são apresentados, com o objetivo de validar a *Filtered*-ARN e compará-la à ARN bem como à árvore de decisão. Por fim, as conclusões e alguns trabalhos futuros são apresentados na Seção VI.

II. Mineração de Regras de Associação

O objetivo da mineração de dados é encontrar modelos para prever o futuro ou entender o passado [1]. A descoberta de regras de associação é uma técnica de mineração de dados, que procura identificar determinados padrões de dados em *datasets*, permitindo, após a sua interpretação, adquirir conhecimento específico acerca do problema em análise [9].

Definição 1 Seja $I = \{i_1, i_2, ..., i_n\}$ um conjunto de objetos denominados itens que podem assumir valores binários 0 ou 1 (falso ou verdadeiro), 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 que está contido numa transação D, se todos os itens do conjunto têm valor "verdadeiro" na transação, ou seja, fazem parte dessa mesma transação. Uma regra de associação R pode ser representada por uma expressão da forma: $A \Rightarrow B$, 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, utilizando-as, posteriormente, como binárias. A é denominado de antecedente (LHS - *Left Hand Side*) da regra e B o consequente (RHS - *Right Hand Side*).

Ao se tratar de mineração de regras de associação, pode-se dividir este processo em três etapas [3], [10], [11]:

- **Pré-processamento:** preparação da base para a etapa de extração, podendo ocorrer a remoção de itens não interessantes.
- **Extração de padrões:** são efetuados os cálculos de medidas, construção dos *itemsets* frequentes e construção das regras de associação propriamente ditas.
- **Pós-processamento:** efetua-se a remoção de regras não interessantes e redução do número de regras a serem exploradas pelo usuário.

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) [3].

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

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

O suporte pode ser descrito como a probabilidade de que uma transação qualquer satisfaça tanto LHS quanto RHS, ao passo que a confiança é a probabilidade de que uma transação satisfaça RHS, dado que ela satisfaz LHS.

De acordo com Agrawal [3], o problema de se extrair todas as regras de associação pode ser decomposto em duas partes:

- Encontrar todos os conjuntos de itens que possuam um suporte de transações acima de um limite mínimo informado, denominados de *itemsets* frequentes.
- Gerar as regras de associação a partir dos conjuntos de itens frequentes. Deve-se selecionar apenas as regras que possuam o grau de suporte e confiança mínimos.

Assim, dado um conjunto de transações, o problema de mineração por 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 como suporte_mínimo (*minsup*) e confiança_mínima (*minconf*), respectivamente.

O algoritmo mais conhecido para obtenção de regras de associação é o *Apriori* [3]. O algoritmo emprega busca em profundidade e gera conjuntos de itens candidatos (padrões) de k elementos a partir de conjuntos de itens de $k-1$ elementos. Os padrões não frequentes são eliminados. Toda a base de

dados é percorrida e os conjuntos de itens frequentes são obtidos a partir dos conjuntos de itens candidatos.

Dada a dimensão das bases de dados atuais, o número de regras descobertas pode ser tão elevado, que quase transforma a sua interpretação em um novo problema de mineração. Deste modo, é importante o entendimento das regras de associação e a busca de melhores maneiras para interpretá-las [12].

A seleção dos *itemsets* de interesse torna o processo de mineração de regras de associação complexo, pois a definição de interesse é muito subjetiva, além de estar diretamente conectada ao objetivo do processo de mineração e ao seu respectivo *dataset* [4].

As medidas de interesse desempenham um papel importante na extração e/ou seleção de regras interessantes de associação. Essas medidas são usadas para encontrar padrões com base na necessidade do usuário, uma vez que o grande número de regras de associação geradas pelo algoritmo de mineração de padrões pode não ser útil como um todo. Portanto, há uma necessidade de filtrar as regras [13].

Além das medidas usuais de suporte (*sup*) e confiança (*conf*), outras medidas para a regra (LHS $\Rightarrow$ RHS) podem ser calculadas. Algumas medidas simétricas são muito utilizadas como *Lift* [14], *Rule Interest* [15] e Teste do χ^2 (qui-quadrado) [16].

Definição 3 Medidas de interesse objetivas podem ser classificadas como simétricas ou assimétricas, isto é, uma medida M é simétrica se M($A \Rightarrow B$) = M($B \Rightarrow A$) [17].

Uma das funções das medidas objetivas de interesse é demonstrar como os itens influenciam uns aos outros. Esta influência pode ser de modo direto, quando os itens variam de modo diretamente proporcional, ou inverso, quando a elevação da incidência de um item leva à diminuição de outro.

Para este trabalho, foram selecionadas as medidas *Added Value* (AV) e *Gain* para construção da abordagem proposta por se tratarem de medidas assimétricas, sendo relacionadas diretamente com a natureza direcionada das ARNs.

Definição 4 ***Added Value*** **[-1;1]:** a medida *Added Value* (AV), descrita na 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 [18]. Se $AV > 0$, a frequência de RHS aumenta na presença de LHS. Sendo $AV < 0$, a frequência de RHS diminui na presença de LHS. Se $AV = 0$, tem-se uma coincidência aleatória, ou seja, a frequência de LHS não altera a frequência de RHS.

$$AV = P(RHS|LHS) - P(RHS) = conf(LHS \Rightarrow RHS) - P(RHS) \quad (3)$$

Definição 5 ***Gain*** **[0;1]:** é uma medida proposta por Fukuda [19] (Equação 4) que forma um *trade-off* entre suporte e confiança, auxiliando na seleção das regras de acordo com a frequência da mesma e o valor da confiança mínima.

$$Gain = [conf(LHS \Rightarrow RHS) - minconf].P(LHS) \quad (4)$$

Por meio da Equação 4, percebe-se que a medida objetiva *Gain* funciona como uma normalização da medida confiança. Quando o valor de $Gain = 0$ a confiança da regra é igual a confiança mínima ($conf(LHS \Rightarrow RHS) = minconf$).

A vantagem desta medida sobre a confiança é que pode-se calcular com mais exatidão a influência do elemento antecessor sobre o sucessor, podendo também a medida *gain*, ser utilizada para selecionar regras.

III. Trabalhos Relacionados

Conforme apresentado, as regras de associação são geralmente extraídas usando um limite mínimo de suporte e confiança.

Hahsler e Karpienko [20] apresentam um método de visualização interativa por meio de uma representação de matriz agrupada, que permite explorar e interpretar intuitivamente cenários altamente complexos. Os grupos de regras geradas são selecionadas com o uso da medida *Lift* e aninhados formando uma hierarquia que pode ser explorada interativamente até a regra individual.

No trabalho de Deng [21] é demonstrado o uso de classificadores associativos que consistem em um conjunto de regras ordenado e representados como um modelo de árvore. Além disso, ele também propôs um algoritmo para transformar uma árvore em um conjunto de regras ordenadas com condições de regras concisas.

Ao incorporar operadores de negação e disjunção em antecedentes de regras, Kim [22] oferece um poder expressivo ao descrever os interesses dos usuários como antecedentes. Nesse trabalho é demonstrado o uso de três elementos: (1) um modelo conceitual; (2) três algoritmos, chamados de família de algoritmos CULTIVATION; (3) um sistema baseado, que é uma implementação em Java da abordagem.

Os trabalhos apresentados funcionam processando as regras e modelando-as de maneira que facilita o entendimento do usuário. No entanto, às vezes, o usuário quer analisar o comportamento de um item específico. Esses trabalhos podem reduzir o número de regras a serem analisadas pelo usuário, mas não explicam como um item específico interage com o *dataset* inteiro. Este item de exploração pode ser extremamente útil na construção de hipóteses sobre os dados.

Com o objetivo de facilitar a obtenção do conhecimento por meio das regras de associação, foram propostas algumas abordagens que aliam o uso de redes à mineração de regras de associação nas três principais fases de mineração (Pré-processamento, Extração e Pós-processamento) [23].

Pode-se citar, a abordagem de pós-processamento com o uso de redes e aprendizado transdutivo, no qual são selecionadas algumas regras a serem classificadas pelo usuário, direcionando-se esforços do mesmo com as regras consideradas de maior impacto na rede, de acordo com alguma medida de rede [24]. Uma operação importante no pós-processamento das regras é a poda, que consiste na eliminação de regras de não-interesse, com isso o uso de hipergrafos direcionados é uma abordagem eficiente no auxílio deste processo [25].

Unindo a utilização de redes como auxílio da mineração de regras de associação, principalmente em sua etapa de pós-processamento, bem como a poda das regras para otimização

da extração do conhecimento, Pandey [8] apresentou as Redes de Regras de Associação.

Redes de Regras de Associação possuem uma estrutura que permite sintetizar, podar, e analisar um conjunto de regras de associação para a construção de hipóteses candidatas.

A ideia central da ARN é que as regras de associação descobertas pelo algoritmo de mineração podem ser sintetizada, podadas, e integradas no contexto de objetivos específicos da pesquisa. Em particular, se houver uma variável de interesse ("objetivo"), pode-se formar uma rede com as variáveis mais relevantes e relacionadas ao objetivo, e, logo após, elaborar uma estrutura que pode ser testada usando métodos estatísticos.

Conforme Chawla [25] descreve, ARNs utilizam como representação um hipergrafo inversamente-direcionado (B-graph), que após os processos de poda, podem ir transformando a ARN conforme o objetivo. Para a criação da ARN, são realizadas quatro etapas:

(1) Passo A: dado um *dataset* D, o suporte mínimo e a confiança mínima, deve-se primeiro extrair todas as regras de associação utilizando um algoritmo padrão como *Apriori* [4], *Apriori-Tid* [5] ou *FP-Growth* [26].
(2) Passo B: escolher um item frequente Z, que será representado no conjunto de regras como o nó objetivo, e construir um *B-graph* que flui de forma recursiva para Z.
(3) Passo C: realizar a poda do *B-graph* gerado no Passo B retirando hiperciclos e hiper-arestas reversas. O *B-graph* resultante é chamado de ARN.
(4) Passo D: encontrar caminhos mais curtos entre o nó objetivo e os demais nós no nível mais superior (uma variante da distância entre extremidades) da ARN. O conjunto destes caminhos representa a rede exploratória para o nó objetivo.

IV. *Filtered*-Association Rules Networks

Em geral, uma rede R é representada como R = (V, E), na qual V é um conjunto de vértices (ou nós) e E é um conjunto de arestas (ou links), que ligam alguns pares de vértices em V. Estatisticamente, um grafo pode ser caracterizado por valores derivados,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 [27].

Dada uma rede R = (V, E), vários links e auto-conexões não são permitidas dependendo do tipo de rede que está sendo implementada. Se R é uma rede dirigida (RD), considere o conjunto universal, denotado por U, contendo todas as $|V| * (|V| - 1)$ potenciais ligações dirigidas entre par de nós em V, no qual |V| denota o número de elementos em V. Se R é uma rede não dirigida, 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 [28].

Newman [29] afirma que uma variedade de sistemas pode ser representada como redes em que os dados podem ser reunidos seguindo algum critério. A função do sistema que a rede representa pode indicar qual a forma ideal da rede. Algumas abordagens promovem a análise de uma rede de acordo com um item (nó) objetivo.

Existem casos em que, a mineração de regras de associação é feita com o objetivo de explicar itens predeterminados. A ARN apresenta uma exploração guiada por um único item objetivo. Esta exploração remove todas as regras que não são interessantes no contexto do item objetivo, de acordo com as métricas mínimas de suporte e confiança, mostrando ao usuário apenas as regras relevantes, porém sem a certeza da dependência estatística entre os elementos das regras.

Para exemplificar, considere o *dataset Lenses*[1]. Se o usuário construir uma ARN com o atributo "[lenses]=hard", com o objetivo de descobrir quais sintomas levam ao paciente que utiliza uma lente rígida, o atributo "[prescription]=myope" aparece conectado diretamente ao nó objetivo (Nível = 1). Esse conhecimento pode direcionar o usuário a pensar que um paciente com miopia tem maior probabilidade de usar uma lente rígida.

No entanto, a ARN pode, às vezes, apresentar relações que não possuem influência entre os elementos da regra. Ao calcularmos o valor *Added Value* da regra "[prescription]=myope $\Rightarrow$ [lenses]=hard" encontramos $AV = 0$, o que afere uma total independência entre os elementos constituintes dessa regra, portanto sendo uma hipótese equivocada quanto ao comportamento de pacientes que necessitam de lentes rígidas.

Com o objetivo de permitir uma exploração completa e levando em conta a relação entre um conjunto de itens objetivos, neste artigo é proposta a *Filtered*-Association Rules Network (*Filtered*-ARN), que permite a exploração de um item objetivo com análise de dependência entre os elementos das regras.

Definição 6 Dado um conjunto de regras de associação R, contendo regras de *itemsets* unitários, e um item objetivo Z, a *Filtered*-ARN é uma RD que modela todas as regras relacionadas ao item em Z, como:

1. Cada aresta modela uma regra r $\subset$ R.
2. A partir de qualquer ponto da rede, é sempre possível alcançar pelo menos 1 vértice representando um item Z.
3. Dado um vértice v $\subset$ *Filtered*-ARN, como v $\not\subset$ Z. Não há caminho de qualquer item Z para v.
4. Se existe uma regra r como RHS(r)$\subset$ Z, então a regra r $\subset$ *Filtered*-ARN.

No algoritmo *Filtered*-ARN faz-se uso de filtros com medidas objetivas assimétricas (*Added Value* e *Gain*), construindo o gráfico de acordo com as regras selecionadas.

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

(1) Passo A: similar ao primeiro passo de todos os processos de mineração de regras de associação. Extração das regras com corte por suporte e confiança mínimos.
(2) Passo B: Efetuar o cálculo das medidas objetivas assimétricas *Added Value* e *Gain*, realizando a exclusão de todas as regras com $AV = 0$ e por um ganho mínimo (*mingain*).

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

(3) Passo C: escolher um item frequente Z, que será representado no conjunto de regras como o nó objetivo, e construir um *B-graph* que flui de forma recursiva para Z conforme a metodologia de Pandey [8].

A primeira etapa consiste na fase de mineração de regras de associação. A única restrição adicionada a essa etapa, se comparada a uma mineração de regra de associação convencional, é que as regras devem possuir conjuntos unitários no antecedente e consequente ($|LHS| = 1$ e $|RHS| = 1$). Essa restrição foi adicionada para facilitar a modelagem da *Filtered*-ARN.

O segundo passo é a filtragem das regras, i.e. a seleção das regras que possuem elementos com dependência estatística e definição do ganho mínimo de influência (*mingain*). Esta etapa guiará toda a exploração, pois definirá as regras de interesse que serão utilizadas com o item objetivo do qual a rede será construída.

No último passo, o usuário deve selecionar o item que deseja entender no *dataset*. Logo após, é efetuada a construção da *Filtered*-ARN. Esta etapa é responsável por obter todas as regras que estão direta ou indiretamente relacionadas ao item objetivo e modelá-las. A construção da *Filtered*-ARN é feita recursivamente. Primeiro, o item selecionado como item objetivo é modelado no gráfico (Nível = 0). Então, todas as regras que o item LHS não estão no gráfico e possuem o item do RHS no nível 0 são modelados na rede. O mesmo processo é feito para todos os itens no nível 1, nível 2 e assim por diante. Até que não haja mais regras a serem modeladas.

Além disso, a *Filtered*-ARN é construída de acordo com os níveis de seus vértices.

Definição 7 O nível de um determinado vértice v $\subset$ *Filtered*-ARN é o número de arestas necessárias para acessar o item Z.

Por exemplo, o item Z têm nível zero (Nível = 0), pois ele não precisa passar por nenhuma aresta para alcançar o próprio item Z. Itens na parte LHS de regras que possuem RHS $\subset$ Z terá o nível um (Nível = 1). Eles estão uma aresta longe dos itens em Z.

Conforme descrito, a *Filtered*-ARN é construída conectando vértices em um gráfico de acordo com a regra de associação que está representando. A *Filtered*-ARN é estabelecida como sendo uma RD, a prova é a seguinte:

Prova: a *Filtered*-ARN é construída conectando vértices do nível X aos vértices no nível $X-1$. Assim, todas as conexões na *Filtered*-ARN são direcionadas. Suponha que haja um ciclo A $\rightarrow$ B $\rightarrow$ C $\rightarrow$ A, de acordo com as regras de construção da *Filtered*-ARN, Nível (A) = Nível (B) + 1, Nível (B) = Nível (C) + 1 e Nível (C) = Nível (A) + 1. Nível (C) = Nível (A) - 2, portanto, não é possível criar um ciclo.

Outra propriedade importante é que, a partir de qualquer vértice da *Filtered*-ARN, é possível atingir um vértice no nível 0 (item objetivo).

Prova: suponha que exista um vértice v_x, no nível N > 0, que não esteja conectado a nenhum vértice nível zero (Nível = 0). A *Filtered*-ARN é construída do nível zero aos níveis mais altos, sempre modelando as regras que têm o RHS em um nível inferior.

Assim, o vértice v_x só será modelado se um vértice em um nível inferior àquele que v_x está conectado, também estiver conectado a um vértice no nível 0. Então, todos os nós na *Filtered*-ARN estão conectados a, pelo menos, um vértice em nível 0.

A exclusão das regras de dependência estatística nula (AV = 0) é uma propriedade fundamental das *Filtered*-ARNs, pois garantem que todos os itens que participam da rede geram algum tipo de influência no item objetivo.

Prova: a medida objetiva *Added Value* (Equação 3) é o resultado da diferença entre a confiança $conf(LHS \Rightarrow RHS)$ e o $sup(RHS)$. Essa medida torna-se zero (AV = 0) quando a confiança da regra e o suporte do consequente são iguais. Como a confiança da regra é uma probabilidade condicionada à presença de RHS e o suporte é a probabilidade de RHS, significa que o item antecedente não influencia em nenhum momento na presença do consequente, provando a total independência estatística dos mesmos.

Todas essas propriedades são importantes para a análise da *Filtered*-ARN, pois garantem que todos os itens modelados sempre apontem para o item objetivo selecionado com algum tipo de relação de influência. As propriedades garantem que toda a *Filtered*-ARN estará explorando o item definido em Z, permitindo que o usuário entenda sua ocorrência e construção de hipóteses a partir dos dados.

Para validação da *Filtered*-ARN, foram utilizados dois *datasets* da UCI [2] (*Lenses* e *SoyBean Large*), além de um *dataset* real (Green Manure) coletado na Embrapa Meio Norte, na cidade de Parnaíba, estado do Piauí. Para a seleção das regras foi estabelecido um valor de ganho mínimo (*mingain*) igual a 0,1, efetuando-se a filtragem das regras que possuem AV = 0. Para a construção da rede foi usado o processo descrito por Pandey [8].

Como o objetivo da *Filtered*-ARN é analisar as correlações no *dataset* de modo ao usuário obter informações fidedignas, e construir hipóteses com grande probabilidade de serem verdadeiras, a *Filtered*-ARN é avaliada com dois outros métodos. Primeiro, faz-se uma comparação entre a *Filtered*-ARN e a ARN convencional, analisando as diferenças entre elas e os prós e contras de usar cada uma dessas abordagens. Depois disso, a *Filtered*-ARN é comparada a um algoritmo de árvore de decisão. O objetivo dessa comparação é confrontar os resultados e discutir qual deles produz uma estrutura melhor para análise do *dataset* de acordo com um conjunto de itens.

V. Estudos de Caso

Com o objetivo de validar a abordagem *Filtered*-ARN e apresentar sua capacidade de exploração de *datasets*, foram realizados alguns experimentos em *datasets* conhecidos, bem como em um *dataset* real. Os experimentos foram focados em apresentar as diferenças dos métodos, mostrando como a *Filtered*-ARN é capaz de permitir uma exploração mais ampla

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

dos dados, dando ao usuário uma compreensão completa dos mesmos. Além disso, um algoritmo de árvore de decisão foi utilizado nos *datasets* a fim de comparar seu resultado com a saída da *Filtered*-ARN.

O primeiro estudo foi utilizando o *dataset Lenses*, disponível na UCI[3]. Nesse *dataset*, cada linha representa os atributos de um paciente e a lente de contato que foi prescrita para ele. O objetivo principal do *dataset* é descrever quais características implicam na prescrição de cada tipo de lente de contato. É importante lembrar que este *dataset* é uma versão simplificada do problema, os *insights* obtidos a partir dele podem não representar as correlações reais do cenário de prescrição de lentes de contato.

O segundo estudo foi feito no *dataset Soybean (Large)*, também disponível na UCI[4]. O *dataset* relata os resultados de uma pesquisa de características de doenças em plantações de soja. Existem 19 classes, apenas as 15 primeiras foram usadas em trabalhos anteriores. As últimas quatro classes são pouco exploradas, pois têm poucos exemplos. Existem 35 atributos categóricos, alguns nominais e alguns ordenados. Os valores dos atributos são codificados numericamente, com o primeiro valor codificado como " 0 ", o segundo como " 1, " e assim por diante. O *dataset* possui campos "vazios".

Para a escolha dos *datasets*, foi elaborada, neste trabalho, uma proposta para o cálculo da taxa de complexidade utilizando os dados encontrados em Gupta [30]. Na presente pesquisa o autor realizou a comparação de 9 algoritmos de aprendizado de máquina com 11 *datasets* oriundos da UCI. Com os resultados gerados neste trabalho, foi proposto um cálculo que permite comparar o grau de complexidade de cada *dataset* de acordo com a definição.

Definição 8 Proposta de Taxa de Complexidade A taxa de complexidade de um *dataset* é calculado pela razão entre a média aritmética do tempo de complexidade (Tc) gerado por algoritmos de aprendizado de máquina e a média aritmética da acurácia (Acc) destes mesmos algoritmos multiplicado pelo fator de correção 1000 (Equação 5).

$$TxComplexidade = \frac{Tc}{Acc} * 1000 \tag{5}$$

Assim, os 11 *datasets* da UCI foram classificadas de acordo com a Tabela I.

Para a realização dos experimentos apresentados neste trabalho, foram selecionados dois *datasets*, sendo um de baixa complexidade (LENSES) e outro de alta complexidade (SOYBEAN LARGE).

Além dos 2 *datasets* da UCI, foi utilizado para validação das *Filtered*-ARNs uma base de dados coletada na EMBRAPA Meio Norte, na cidade de Parnaíba, estado do Piauí [31]. O terceiro estudo foi elaborado com o *dataset* de Decomposição de Adubação Orgânica (Green Manure). Este *dataset* relaciona parâmetros de produção de leguminosas para uso como adubos orgânicos com o tempo de meia vida de cada planta utilizada nos experimentos. Com os 3 *datasets* foram realizadas as mesmas etapas para construção das *Filtered*-ARNs.

Tabela I
TAXA DE COMPLEXIDADE

Dataset	**Taxa de Complexidade**
LENSES	1,804
LABOR	2,305
IRIS	3,065
LUNG CANCER	4,765
VOTE	5,355
HAYES-ROTH	5,686
TEACHING ASSISTANT	6,030
STATLOG	8,256
GLASS IDENTIFICATION	8,897
SOYBEAN LARGE	123,469
ABALONE	1192,960

Nesses experimentos, as regras foram extraídas com o uso do algoritmo *Apriori-TID* implementado em Java.

As regras extraídas foram filtradas pela medida *Added Value*, sendo excluídas todas aquelas que possuem valor nulo (AV = 0), que indicam ausência de influência estatística, e também foi selecionado o ganho mínimo de 0,01, para que o maior número de regras possível pudesse ser analisado

As redes de regras foram construídas graficamente com o uso do software *Gephi* [32]. O *Gephi* é uma aplicação *open source* específica para construção de redes e está disponível no link https://gephi.org/users/download/.

As árvores de decisão foram geradas pelo algoritmo J48, disponível no *software* Weka[5], usando a configuração padrão. Nas seções seguintes são explanados os resultados com cada *dataset*.

A. Lenses dadaset

Como o *dataset* tem um pequeno número de atributos, o suporte mínimo foi definido como 0, portanto, todos os valores foram considerados. A confiança mínima foi definida para 0,25, possibilitando o estudo de classes que ocorreram pelo menos 1 em cada 4 vezes. Além disso, o tamanho da regra foi especificado em dois, considerando um item no LHS e um item no RHS. Usando essa configuração, foram obtidas 99 regras de associação candidatas para o *dataset* "*lenses*". Após a etapa de filtragem restaram 60 regras.

A *Filtered*-ARN foi gerada considerando como item objetivos: "[lenses]=hard". O resultado pode ser visto na Figura 1. A rede gerada possui 4 níveis: nível 0, nível 1, nível 2 e nível 3. A *Filtered*-ARN tem apenas 2 itens relacionados ao item objetivo, que são: "[tear]=normal" e "[astigmatic]=yes" Estas regras podem ser capazes de influenciar no item "[lenses]=hard" e é interessante investigar, pois são os únicos parâmetros que geram uma influência no item objetivo, portanto com um alto grau de probabilidade de geração de hipóteses verdadeiras.

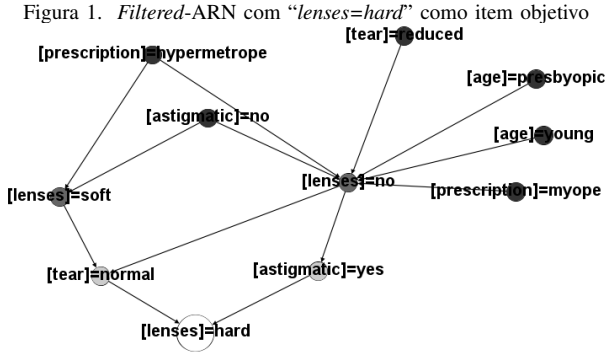
Pode-se observar na *Filtered*-ARN gerada que os nós de nível 2 são as outras classes que indicam o tipo de

[3]https://archive.ics.uci.edu/ml/datasets/lenses

[4]http://archive.ics.uci.edu/ml/datasets/Soybean+%28Large%29

[5]http://www.cs.waikato.ac.nz/ml/weka/

Figura 1. *Filtered*-ARN com "*lenses=hard*" como item objetivo

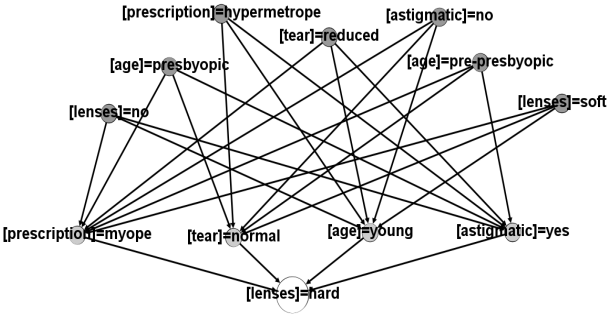


lente "[lenses]=soft" e "[lenses]=no", indicando que os elementos do nível 1 sofrem interferência desses itens. Quando analisa-se os nós de nível 3 três, perecebe-se que algumas regras se destacam como: "[tear]=reduced"⇒ "[lenses]=no", "[age]=presbyopic"⇒ "[lenses]=no", "[age]=young"⇒"[lenses]=no" e "[prescription]=myope"⇒ "[lenses]=no". Estas regras possuem conexão apenas com a classe "[lenses]=no", o que gera hipóteses para a construção de uma nova *Filtered*-ARN, provocando um novo direcionamento na exploração do conhecimento.

Na Figura 2 é apresentada a ARN com "[lenses]=hard" como item objetivo. Pode-se observar que a rede possui apenas 3 níveis, porém com uma estrutura totalmente diferente da *Filtered*-ARN. Como a construção das redes é diretamente induzida pelo item RHS das regras, percebe-se que regras do nível 1 sem influência comprovada ($AV = 0$) aparecem, como "[prescription]=myope"⇒ "[lenses]=hard" e "[age]=young"⇒"[lenses]=hard", o que leva a um geração de hipóteses equivocadas a respeito do uso de lentes rígidas.

Outra diferença notória é o aumento do número de nós de nível 3, sem a distinção de dependências entre os mesmos, o que impossibilita o direcionamento dos estudos, pois a rede demonstra o mesmo grau de importância para todos os itens do mesmo nível.

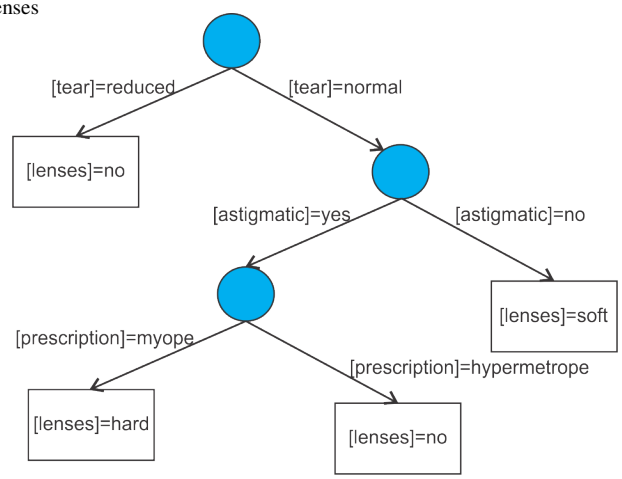
Figura 2. ARN com "*lenses=hard*" como item objetivo



A árvore de decisão é apresentada na Figura 3. O algoritmo J48 obteve 20/24 classificações corretas (83,33%) e perdeu 4/24 exemplos (16,67%). A saída completa pode ser vista em *https://goo.gl/V882D2*.

Comparando a saída da *Filtered*-ARN com a árvore de

Figura 3. Árvore de decisão construída com o algoritmo J48 com o *dataset* Lenses



decisão (Figura 3), é possível perceber diferenças na explicação dos itens objetivos. Ambos têm o "[tear]=reduced" e "[prescription]=hypermetrope" conectados diretamente ao "[lenses]=no", porém na *Filtered*-ARN outras regras surgem como a codição de idade ([age]).

Na árvore, é explicado que "[prescription]=myope" se liga diretamente à "[lenses]=hard", o que é um equívoco por se tratar de uma condição sem influência (AV = 0), o que pode ser inferido na *Filtered*-ARN.

B. Soybean Large dataset

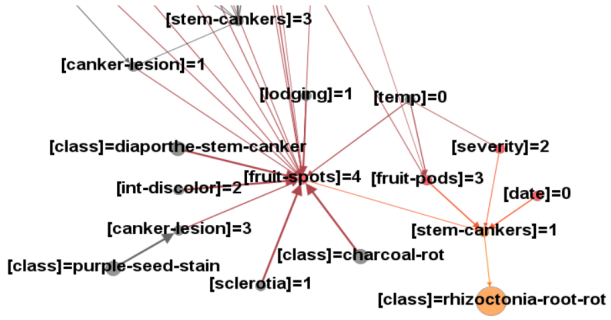
Para continuar a validação dos benefícios do uso da *Filtered*-ARN, algumas experiências foram realizadas no *dataset* de plantações de soja (Soybean Large), por se tratar de um *dataset* com alto grau de complexidade (Tabela I).

Uma vez que o *dataset* possui 19 classes, 36 atributos e 307 instâncias, sendo algumas com campos vazios, o suporte mínimo foi definido como 0,03 para evitar o aparecimento de classes com informações incompletas. A confiança mínima foi definida para 0,45, para a geração de uma proporção mais coerente com o item objetivo "[classe] = rhizoctonia", o qual foi selecionado, em relação aos itens das demais classes. Além disso, o tamanho da regra foi especificado em dois, considerando um item no LHS e um item no RHS. Usando essa configuração, foram obtidas 4223 regras de associação candidatas para o *dataset* "*Soybean*". Após a etapa de filtragem restaram 4019 regras.

A *Filtered*-ARN foi gerada considerando como item objetivos: "[classe] = rhizoctonia" o que significa plantas com um tipo específico de fungo. O resultado da rede, com os nós de nível 1 e nível 2, pode ser visto na Figura 4. O arquivo contendo a saída completa pode ser vista em *https://goo.gl/V882D2*.

Analisando-se os itens de nível 1 da *Filtered*-ARN percebe-se que apenas 1 elemento provoca influência direta no item objetivo, "[stem-cankers] = 1".Esta relação pode ser capaz de descrever o item objetivo com um alto grau de dependência, ocasionando a formação de uma hipóteses com grande probabilidade de ser verdadeira.

Figura 4. *Filtered*-ARN com "[class]=rhizoctonia" como item objetivo

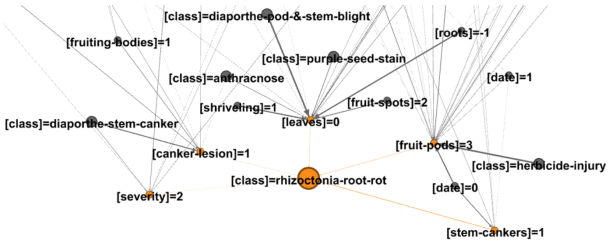


Na Figura 5 é apresentado um corte da ARN com "[class]=rhizoctonia" como item objetivo e nós de nível 1, nível 2 e nível 3 sem antecessores. O arquivo contendo a saída completa pode ser encontrado em *https://goo.gl/V882D2*. Pode-se observar que a rede possui um número bem maior de itens conectados diretamente ao objetivo. Além do item "[stem-cankers] = 1", outros 4 elementos formam o nível 1 da rede, porém sem nenhuma garantia de dependência, o que pode ocasionar em formação de hipóteses falsas.

Os elementos "[canker-lesion] = 1" e "[severity] = 2", que fazem parte do nível 1 da ARN, são observados no nível 2 na *Filtered*-ARN, deduzindo-se com isso que eles podem afetar o item objetivo, mas de um modo mais indireto que o indicado pela ARN tradicional. Os outros itens observados no nível 1 da ARN, "[leaves] = 0" e "[fruit-pods] = 3", fazem parte do nível 3 da *Filtered*-ARN, o que diminui drasticamente a probabilidade de uma hipótese ser gerada diretamente com o item objetivo.

Na *Filtered*-ARN pode-se observar outros itens de nível 2, "[fruit-spots] = 4" e "[date] = 0", que influen diretamente na única condição ligada diretamente ao item objetivo.

Figura 5. ARN com "[class]=rhizoctonia" como item objetivo



A árvore de decisão gerada para validação possui 69 folhas e obteve 87,58% de precisão. Como a árvore é muito longa, não foi gerada imagem da mesma. O arquivo contendo a saída completa pode ser visto em *https://goo.gl/V882D2*. Ao analisar a árvore, é difícil entender até mesmo o comportamento do item "[class]=rhizoctonia". A árvore que explica as classes ficou mais complexa de entender do que a *Filtered*-ARN.

C. Green Manure dataset

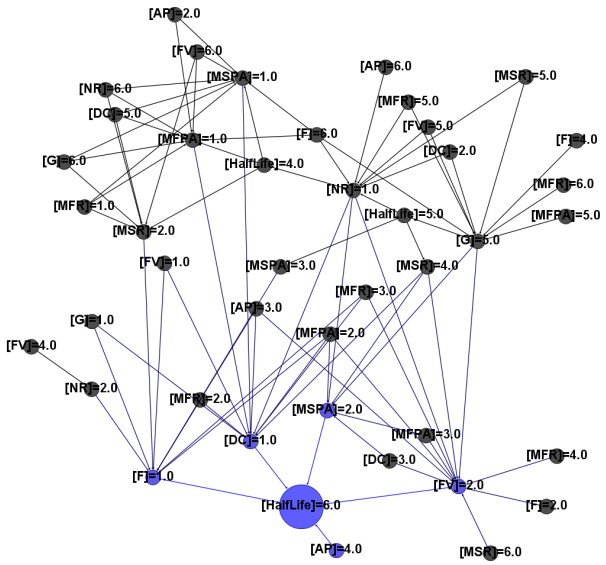
Com o intuito de validar os benefícios do uso da *Filtered*-ARN em dados reais, algumas experiências foram realizadas no *dataset* de Adubo Orgânico (*Green Manure*) coletado junto a EMBRAPA Meio Norte. Este *dataset* já foi utilizado por [31] na construção de ARNs para exploração das informações.

Uma vez que o *dataset* possui 6 classes relacionados à meia-vida de decomposição de leguminosas, 11 atributos e 28 instâncias, sendo os atributos formados por categorias de valores referente à pesquisa de campo executada pela EMBRAPA Meio Norte, o suporte mínimo foi definido como 0,3 e a confiança mínima foi definida para 0,5 para aferição comparativa com os estudos de [31]. Além disso, o tamanho da regra foi especificado em dois, considerando um item no LHS e um item no RHS. Usando essa configuração, foram obtidas 64 regras de associação candidatas para o *dataset* "*Green Manure*". Após a etapa de filtragem restaram 50 regras.

A *Filtered*-ARN foi gerada considerando como item objetivos: "[HalfLife]=6" o que significa leguminosas com maiores taxa de meia-vida por possuírem uma taxa de decomposição elevada. O resultado da rede pode ser visto na Figura 6. O arquivo contendo a saída completa pode ser visto em *https://goo.gl/V882D2*.

Analisando-se os itens de nível 1 da *Filtered*-ARN percebe-se que 5 itens estão relacionados ao item objetivo, destacando-se "[AP] = 4", parâmetro relacionado a altura da planta no florescimento, que não possui nenhum antecessor. Estas regras podem ser capazes de descrever quais as prováveis características das plantas com maior tempo de meia-vida, sendo os únicos parâmetros que geram uma influência no item objetivo, portanto gerando hipóteses com um alto grau de probabilidade de serem verdadeiras.

Figura 6. *Filtered*-ARN com "*HalfLife*=6" como item objetivo.

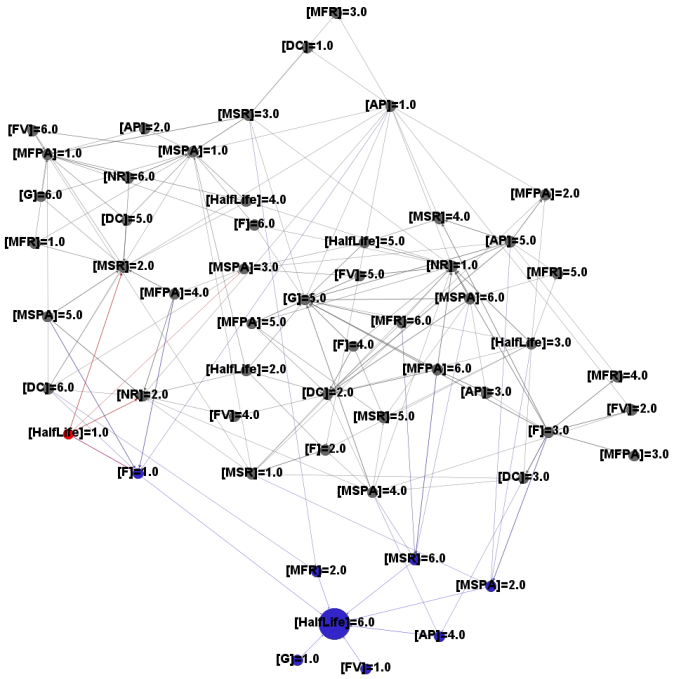


Na Figura 7 é apresentada a ARN com "[HalfLife]=6" como item objetivo. Pode-se observar que a rede possui uma estrutura diferente da *Filtered*-ARN. O número de itens conectados diretamente ao objetivo aumentou para 7. Percebe-se que alguns elementos são mantidos em ambas a redes conectados

ao item objetivo: “[MSPA] = 2.0”,“[F] = 1.0” e “[AP] = 4.0”, sendo que este último deixou de ter antecessores na *Filtered*-ARN, aumentando o grau de importância do mesmo para o estudo. O parâmetro “[FV]” sofreu uma mudança de categoria, na ARN percebia uma categoria “1.0” e na *Filtered*-ARN esse valor foi alterado para a categoria “2.0”, indicando que, embora exista uma relação entre estes elementos, a influência inicia na segunda faixa de valores. As demais regras com elementos de nível 1 da ARN foram eliminados na *Filtered*-ARN.

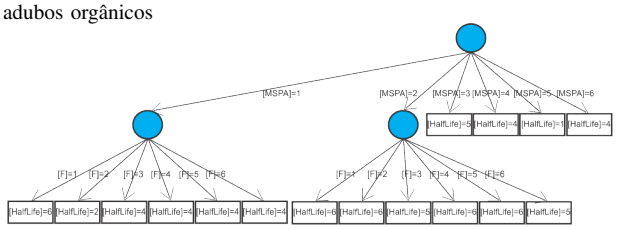
Outra diferença notória é o aparecimento de um novo parâmetro de influência relacionado ao diâmetro do coletor ([DC]). Na *Filtered*-ARN este parâmetro na categoria “1.0” aparece com influência direta ao item objetivo, o que não poderia ser percebido pela ARN convencional.

Figura 7. ARN com “*HalfLife*=6” como item objetivo. Adaptado de [31].



Além de comparar com a ARN, a *Filtered*-ARN também foi comparada a uma árvore de decisão. A árvore de decisão gerada pelo algoritmo J48 é apresentada na Figura 8. O algoritmo J48 obteve apenas 39,28% de classificações corretas formando uma árvore com 16 folhas.

Figura 8. Árvore de decisão construída com o algoritmo J48 com o *dataset* de adubos orgânicos



Comparando a saída da *Filtered*-ARN com a árvore de decisão (Figura 8), é possível perceber diferenças na explicação dos itens objetivos. Para o objetivo “[HalfLife] = 6.0” é observado na árvore apenas a possibilidade direta do parâmetro [F] referente à floração da planta, sendo obtido os valores “[F] = 1.0” quando “[MSPA] = 1.0” e “[F] = 1.0 ou 2.0 ou 4.0 ou 5.0” quando “[MSPA] = 2.0”. As outras condições paramétricas para obtenção de meias-vidas mais elevadas são totalmente ignoradas pelo algoritmo J48. Desta forma, a aplicação do algoritmo *Filtered*-ARN possibilita um estudo mais amplo de parâmetros, otimizando a extração de conhecimento por meio do estudo de hipóteses mais relevantes e com maior probabilidade de serem verdadeiras.

VI. Conclusão e Trabalhos Futuros

Neste artigo foi apresentada a proposta de *Filtered*-ARN, um método capaz de modelar as regras de associação, previamente selecionadas por meio de medidas objetivas assimétricas, de acordo com um item objetivo previamente definido. As regras selecionadas para a construção da rede são aquelas que possuem uma dependência estatística comprovada pela medida AV. O item objetivo é utilizado como norteador da exploração e é escolhido de acordo com a problemática que se deseja formular hipóteses. A *Filtered*-ARN cria um hipergrafo direcionado, modelando as regras de associação que tem o item objetivo no RHS recursivamente. Ele visa explicar a correlação entre os itens no *dataset* com o item objetivo.

Três estudos de caso foram desenvolvidos para validar a capacidade da *Filtered*-ARN. Dois *datasets* artificiais foram explorados: *Lenses* e *Soybean Large*, além de um *dataset* real: *Green Manure*. O objetivo foi descrever a ocorrência de regras que influenciam estatisticamente a classe desejada, visando encontrar os itens que melhor explicam a ocorrência do item da classe. Além da *Filtered*-ARN, foram aplicados a ARN convencional e o algoritmo J48 para comparação.

Os resultados demonstraram que a ARN é boa para descrever relações com um item objetivo, no entanto, ele não mostra ao usuário quais os casos em que os itens verdadeiramente influenciam estatisticamente o item objetivo. É importante destacar esses casos porque algumas explorações são feitas com item objetivo e hipóteses equivocadas são elaboradas ocasionando uma demora na extração do conhecimento, e o mesmo ainda pode ser falso. A *Filtered*-ARN é muito bem sucedida em apresentar esses casos, permitindo que o usuário veja os itens que produzem variações no item objetivo.

Também foi comparado o algoritmo *Filtered*-ARN ao J48, que é um algoritmo de árvore de decisão disponível no Weka. Como a árvore de decisão realiza apenas a classificação dos elementos, não é muito boa em explicar as relações. A árvore de decisão gerada no *dataset Soybean* é um ótimo exemplo disso, como a árvore conseguiu 87,58% de precisão, mas a árvore ficou muito grande (69 folhas), fica muito difícil para um usuário entender as correlações. A *Filtered*-ARN obteve bons resultados, descrevendo os dados usando as regras de associação extraídas e selecionadas, mostrando ao usuário os

itens que são inteiramente responsáveis pela ocorrência do item objetivo e que geram influência no mesmo.

Além de apresentar conhecimentos interessantes, existem várias melhorias que podem ser feitas na *Filtered*-ARN para ajudar o usuário a identificar os itens que não são interessantes para sua exploração. O desenvolvimento de uma abordagem com mais de um item objetivo com o intuito de relacionar aqueles que verdadeiramente não concorrem entre si pode auxiliar na identificação de possíveis itens interessantes.

É importante destacar a variabilidade dos valores do ganho mínimo (*mingain*), esse tipo de medida interfere diretamente na estrutura da rede elaborada, portanto pode-se fazer uma correlação das medidas objetivas das regras de associação com medidas de centralidade da rede gerada, a fim de auxiliar na seleção dos melhores parâmetros de extração.

Em relação à estrutura *Filtered*-ARN, é interessante analisar o efeito de alguns algoritmos de construção de redes no resultado final, visando otimizar certas características e permitindo ao usuário manipular a construção da *Filtered*-ARN.

ACKNOWLEDGMENT

Agradecimento a CAPES pelo financiamento e a EMBRAPA Meio Norte pelos dados relacionados à adubação orgânica.

REFERÊNCIAS

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

[2] M. Vinaya and K. Shah, "Performance Evaluation of Distributed Association Rule Mining Algorithms," *Procedia - Procedia Computer Science*, vol. 79, pp. 127–134, 2016.

[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Special Interest Group on Management of Data, 22(2)*, vol. 22, no. 2, pp. 207–216, 1994.

[4] R. Agrawal and R. Srikant., "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)*, pp. 487–499, 1994.

[5] R. Agrawal and J. C. Shafer, "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 962–969, 1996.

[6] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. C. Riquelme, "Selecting the best measures to discover quantitative association rules," *Neurocomputing*, vol. 126, pp. 3–14, 2014.

[7] C. Kim, H. Lee, H. Seol, and C. Lee, "Identifying core technologies based on technological cross-impacts: An association rule mining (ARM) and analytic network process (ANP) approach," *Expert Systems with Applications*, vol. 38, no. 10, pp. 12 559–12 564, 2011.

[8] G. Pandey, S. Chawla, S. Poon, B. Arunasalam, and J. G. Davis, "Association Rules Network: Definition and Applications," *Statistical Analysis and Data Mining*, vol. 1, no. 4, pp. 260–179, 2009.

[9] T. Le and B. Vo, "The lattice-based approaches for mining association rules: a review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 6, no. 4, pp. 140–151, 2016.

[10] C. C. Aggarwal, C. Procopiuc, and P. S. Yu, "Finding localized associations in market basket data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 51–62, 2002.

[11] L. T. T. Nguyen and N. T. Nguyen, "Updating mined class association rules for record insertion," *Applied Intelligence*, vol. 42, no. 4, pp. 707–721, 2015.

[12] M. A. Domingues and S. O. Rezende, "Post-processing of Association Rules using Taxonomies," *Proceedings of the 12th Portuguese Conference on Artificial Intelligence (EPIA 2005)*, pp. 192–197, 2005.

[13] D. J. Prajapati, S. Garg, and N. C. Chauhan, "MapReduce Based Multilevel Consistent and Inconsistent Association Rule Detection from Big Data Using Interestingness Measures," *Big Data Research*, vol. 9, pp. 18–27, 2017.

[14] S. Brin, R. Motwani, J. D. Ulman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, pp. 255–264, 1997.

[15] G. Piatetsky-Shapiro, "Discovery, Analysis and Presentation of Strong Rules," *Knowledge Discovery in Databases, AAAI/MIT Press*, pp. 229–248, 1991.

[16] B. Liu, W. Hsu, and Y. Ma, "Pruning and summarizing the discovered associations," *Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data mining*, pp. 125–134, 1999.

[17] P.-N. Tan, M. Steinbach, and V. Kumar, "Association Analysis: Basic Concepts and Algorithms," in *Introduction to Data mining*, 2005, pp. 327–414.

[18] S. Sahar, "What Is Interesting: Studies on Interestingness in Knowledge Discovery," Phd Thes, Tel-Aviv University The, 2003.

[19] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, "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*, 1996, pp. 13–23.

[20] M. Hahsler and R. Karpienko, "Visualizing association rules in hierarchical groups," *Journal of Business Economics*, vol. 87, no. 3, pp. 317–335, 2017.

[21] H. Deng, G. Runger, E. Tuv, and W. Bannister, "CBC: An associative classifier with a small number of rules," *Decision Support Systems*, vol. 59, no. 1, pp. 163–170, 2014.

[22] E. H. Kim, H. G. Kim, S. H. Hwang, and S. I. Lee, "FARM: An FCA-based Association Rule Miner," *Knowledge-Based Systems*, vol. 85, pp. 277–297, 2015.

[23] M. Rashid, I. Gondal, and J. Kamruzzaman, "Mining Associated Patterns from Wireless Sensor Networks," *IEEE TRANSACTIONS ON COMPUTERS*, vol. 64, no. 7, pp. 1998–2011, 2015.

[24] R. d. Padua, S. O. Rezende, and V. O. D. Carvalho, "Post-processing association rules using networks and transductive learning," *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014*, pp. 318–323, 2014.

[25] S. Chawla, "Feature Selection, Association Rules Network and Theory Building." *JMLR: Workshop and Conference Proceedings - The Fourth Workshop on Feature Selection in Data Mining*, vol. 10, pp. 14–21, 2010.

[26] G. Grahne and J. Zhu, "Fast algorithms for frequent itemset mining using FP-trees," *IEEE Trans Knowl Data Eng*, vol. 17, pp. 1347–1362, 2005.

[27] D. F. Nettleton, "Data mining of social networks represented as graphs," *Computer Science Review*, vol. 7, no. 1, pp. 1–34, 2013.

[28] J. C. Valverde-Rebaza and A. De Andrade Lopes, "Link prediction in online social networks using group information," Ph.D. dissertation, Universidade de São Paulo, 2014.

[29] M. Newman, *Networks: An introduction*, 1st ed. New York, USA: Oxford University Press, 2010, vol. 55.

[30] A. Gupta, "Classification of Complex UCI Datasets Using Machine Learning Algorithms Using Hadoop," *International Journal of Scetific & Techology Research*, vol. 4, no. 05, pp. 85–94, 2015.

[31] D. B. Calçada, S. O. Rezende, and M. S. Teodoro, "Analysis of decomposition parameters of green manure in the Brazilian Northeast with Association Rules Networks," in *I International Conference on Agro BigData and Decision Support Systems in Agriculture*, Montevideo, Uruguay, 2017, pp. 63–65.

[32] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," *Third International AAAI Conference on Weblogs and Social Media*, pp. 361–362, 2009.