PÓS-PROCESSAMENTO EM KDD

Cristian Simioni Milani, PUCPR, E-mail: cristiansimionimilani@gmail.com
Deborah Ribeiro Carvalho, PUCPR, E-mail: ribeiro.carvalho@pucpr.br

Resumo. Apesar dos esforços em utilizar o processo KDD – *Knowledge Discovery in Databases* com o objetivo de potencializar o uso das bases de dados disponíveis para apoiar o processo de apoio à decisão ainda são poucos que se valem desta tecnologia no dia a dia. O processo KDD prevê três etapas: Pré-processamento, Mineração de Dados e Pós-processamento. São diversas as dificuldades apontadas quando da utilização do KDD, mas a grande maioria reside na etapa de Pós-processamento. Este artigo se propõe a apresentar algumas estratégias para a etapa de Pós-processamento, que facilitem a avaliação dos padrões não apenas em nível conceitual, mas também facilitando a compreensão a partir de exemplos de fácil compreensão.

Palavras-chave: Mineração de Dados, Pós-processamento, Apoio à Decisão.

POST-PROCESSING IN KDD

Abstract. Despite efforts to use the KDD process (Knowledge Discovery in Database) in order to maximize the use of databases to support the decision process there are few people who adopt this technology in daily life. The KDD process has three phases: Pre-processing, Data Mining and Post-processing. There are several difficulties when users choose KDD and most of them occur during Post-processing. This article aims to present Post-processing strategies to facilitate the evaluation of patterns, not only at the conceptual level, but also using examples.

Keywords: Data Mining, Postprocessing, Decision Support.

1. INTRODUÇÃO

Diversos fatores relacionados à Tecnologia da Informação favoreceram a ampliação do volume armazenado em bases de dados, demandando assim por formas mais eficientes para um melhor aproveitamento do potencial de informações que podem ser extraídas.

Quando se dispõe de grandes volumes é frequente que as técnicas tradicionais de extração de informações sejam insuficientes para orientar o processo decisório, demandando assim pela busca por formas alternativas que permitam uma melhor otimização do uso destas bases.

Uma das alternativas para otimizar o uso de bases de dados é a partir do processo *Knowledge Discovery in Databases* – KDD, o qual compreende em uma de suas fases a Mineração de Dados, na qual ocorre a aplicação de algoritmos com a finalidade específica de identificar padrões válidos, novos, potencialmente úteis e compreensíveis (Fayyad et al., 1996).

Um padrão é definido como um tipo de modelo de uma declaração. Uma instância de um padrão é uma declaração em uma linguagem de alto nível que descreve uma informação preferencialmente interessante, descoberta nos dados de acordo com algum critério estabelecido (Klosgen, 1992).

Além da Mineração de Dados o processo KDD compreende as etapas de Pré-processamento e Pós-processamento. O Pré-processamento é uma etapa em geral trabalhosa em função dos dados disponíveis não estarem de tal forma organizados a permitir a aplicação direta aos algoritmos de mineração.

A etapa de extração de padrões, Mineração de Dados, é a mais direcionada ao cumprimento dos objetivos, ou seja, busca por padrões que apoiem o processo decisório, ou seja, problema de gestão, que motivou o processo KDD (Fayyad et al. 1996). Na etapa de Mineração de Dados podem ser identificadas três tarefas principais: classificação, descoberta de regras de associação e agrupamento.

O Pós-processamento tem como principal objetivo apoiar na verificação de até que ponto estes padrões contribuem na solução do problema inicialmente identificado. Por exemplo, no experimento conduzido por Kobus (2006), ao oferecer para análise do médico especialista as regras de associação descobertas, recebeu como resposta "qual é a sequência de eventos?". Isso porque, por princípio, o algoritmo que descobre regras de associação do tipo <se> A <então> B, não se preocupa se o evento A cronologicamente antecedeu ou não a ocorrência do evento B.

Diferentemente do contexto no qual os algoritmos que descobrem regras de associação foram originalmente propostos, na área da Saúde fica evidenciado que ou o algoritmo que descobre as regras de associação considera a sequenciação de eventos durante a etapa de descoberta, ou os padrões após descobertos são pós-processados recebendo esta indicação.

Para operacionalizar o Pós-processamento existem várias estratégias propostas, entre elas eliminar a redundância, generalizar, identificar no conjunto aqueles com maior potencial de serem interessantes etc. Novamente, no projeto desenvolvido por Kobus (2006) foi necessário que a autora generalizasse os padrões descobertos manualmente, dado o fato de não dispor de uma ferramenta que o fizesse de forma automática.

Nas experimentações realizadas por Dallagassa (2009) foi perceptível a necessidade adicional que os padrões descobertos considerassem (ou demonstrassem) a janela de tempo decorrida entre dois eventos associados. Novamente, o especialista ao avaliar os padrões descobertos, necessitava desta informação temporal para identificar uma possível situação de causa e efeito. Gomes e Hauth (2010) propuseram e implementaram um algoritmo que durante a etapa de Mineração de Dados, a partir de janelas de tempo pré-definidas pelo especialista, descobre regras de associação considerando as possibilidades de intervalos temporais, denominados restrições de tempo.

Porém a despeito da existência de várias propostas para pós-processar os padrões descobertos poucos são os relatos do quanto estas estratégias de fato contribuem para agregar valor ao que o especialista já conhecia anteriormente sobre o problema de gestão. Uma exceção é encontrada no trabalho relatado por Zhang e seus colegas (2009) que acompanharam uma das principais aplicações para a Motorola. O objetivo era encontrar as causas de falhas de chamadas do telefone celular a partir dos dados de log de uso. Uma das constatações foi que apesar da ampla utilização das regras de associação, os usuários raramente consideram uma única regra como sendo interessante por si só. Uma regra só é interessante no contexto de outras regras. Além disso, em muitos casos, cada regra individual pode não ser interessante, mas um grupo delas pode representar uma parte importante do conhecimento. Sendo assim para tratar esta questão foi proposta para análise das regras como operações OLAP e de mineração de impressões gerais. Esta abordagem permite ao usuário explorar o espaço de conhecimento para encontrar facilmente o conhecimento útil e de forma sistemática, bem como fornecer uma estrutura para visualização, denominado Mapa de Oportunidades.

Em geral os esforços relatados na literatura são para pós-processar os padrões descobertos representados na forma de regras de associação, porém existem outros

formatos de apresentação, como por exemplo, árvores de decisão, que tem grande potencial para apoiar profissionais de diversas áreas. Por exemplo, Vianna et al (2009) e Von-Stein (2010) realizaram experimentos no contexto da saúde, a partir de padrões expressos na forma de árvores de decisão, que poderiam ter contribuições mais relevantes se naquele momento estivessem disponíveis estratégias para pós-processar os padrões descobertos naquele formato.

Neste artigo são apresentadas situações de exemplo de duas etapas do processo KDD: Mineração de Dados e o Pós-processamento dos padrões (conhecimento) descobertos voltados para a área da Saúde. A contribuição decorre do fato do texto buscar não apenas apresentar as estratégias, mas também demonstrar com exemplos de fácil compreensão, permitindo assim que o leitor ao replicar o comportamento dos algoritmos entenda melhor o processo.

As várias tarefas desenvolvidas em Mineração de Dados têm como objetivo primário a predição e / ou a descrição. A predição usa atributos para predizer os valores futuros de uma ou mais variáveis (atributos) de interesse, em geral descobre padrões no formato de classificadores. A descrição contempla o que foi descoberto nos dados sob o ponto de vista da interpretação humana (Fayyad et al.1996), tendo os padrões descobertos representados por regras de associação ou agrupamentos.

'Para a predição existe a tarefa de classificação que oportuniza encontrar um modelo que descreva as diversas classes envolvidas no contexto, com o objetivo de classificar (predizer) uma classe às instâncias ainda não classificadas. Por exemplo, na tarefa de classificação, pode-se ter uma aplicação financeira na qual um banco poderia classificar seus clientes em duas classes: "crédito ruim" ou "crédito bom". Em uma aplicação de medicina, um médico poderia classificar alguns de seus pacientes em duas classes: "tem" ou "não tem" uma determinada doença.

A fim de contribuir para a compreensibilidade do conhecimento descoberto (relação entre os atributos e as classes), esse conhecimento é geralmente representado na forma de regras "se"... (condições) ... "então"... (classe) ..., cuja interpretação é: "se" os valores dos atributos satisfazem as condições da regra "então" o exemplo pertence à classe prevista pela regra.

Para a descrição, conta-se com a tarefa de descoberta de regras de associação e agrupamento. As regras de associação são expressões $X \rightarrow Y$ (lidas como: SE (X) ENTÃO (Y)). O significado de cada regra desta natureza é de que os conjuntos de itens X e Y frequentemente ocorrem juntos em uma mesma transação (registro). (Agrawal et al, 1993).

A tarefa de agrupamento consiste na identificação de um conjunto finito de grupos, classes ou clusters, baseados nos atributos de objetos não previamente classificados. Por exemplo, um conjunto de pacientes pode ser agrupado em várias classes (grupos) baseadas nas similaridades dos seus sintomas, e os sintomas comuns aos pacientes de cada grupo podem ser usados para descrever à qual classe um novo paciente pertencerá. Assim, um dado paciente seria atribuído ao cluster cujos pacientes têm sintomas o mais parecido possível com os sintomas daquele dado paciente. Dessa forma, a tarefa de agrupamento, cujo resultado é a identificação de novas classes, pode ser realizada como pré-processamento para realização da tarefa de classificação (Kubat et al., 1998).

2. PÓS-PROCESSAMENTO

Existem várias estratégias propostas na literatura para pós-processar o conhecimento descoberto, entre elas a atribuição de medidas de potencial grau de interesse as quais

são organizadas em dois grupos, ditas *user-driven* e *data-driven* (Silberschatz & Tuzhilin, 1996), (Freitas, 1998). Outra estratégia é proposta por Hussain et al. (2000) que constitui um método que identifica, a partir de um conjunto de padrões descobertos, um subconjunto de regras que representam regras de exceção e, além disso, atribui uma medida de interesse para cada regra. A tabela 1 mostra a estrutura geral das regras de exceção. Nesta tabela A, B e C são conjuntos não-vazios de itens de dados associados e o símbolo "¬" denota a negação lógica. É importante observar que uma regra de exceção é uma especialização de uma regra geral e uma regra de exceção associa a um item de dados que nega aquele identificado pela regra geral. Este método assume que regras de senso comum representam padrões conhecidos pelo usuário, tendo em vista que aquelas regras têm uma grande cobertura, ao contrário das regras de exceção, que em geral são desconhecidas, uma vez que elas têm baixa cobertura. Sendo assim, as regras de exceção tendem a ser surpreendentes, dado o fato de representarem uma contradição em relação à regra de senso comum. É importante observar que a regra de referência auxilia na explicação da causa da regra de exceção.

$A \rightarrow C$ regra geral (alta cobertura e alta confiança)
$A, B \rightarrow \neg C$ regra de exceção (baixa cobertura, alta confiança)
$B \rightarrow \neg C$ regra de referência (baixa cobertura e/ou baixa confiança)

Tabela 1. Estrutura das Regras de Exceção.

Formalmente, a medida proposta por Hussain et al. (2000) é definida da seguinte forma:

$$\Pr(AC) \cdot \log \frac{\Pr(AC)}{\Pr(AC)} + \Pr(AB\neg C) \cdot \log \frac{\Pr(AB\neg C)}{\Pr(A\neg C) \cdot \Pr(B\neg C)}$$

Quanto maior o valor da medida de interesse, maior é a chance de a regra ser surpreendente.

Outra estratégia para o Pós-processamento é o filtro de regras de associação, que objetiva selecionar aquelas que associem alguns elementos previamente selecionados (ou descartados) pelo especialista. Esta estratégia, a partir da redução do conjunto de regras, além de facilitar a análise elo especialista também melhora significativamente o desempenho de algoritmos que a partir deste conjunto reduzido venha executar outras estratégias de Pós-processamento. A parametrização do processo de filtro é a partir dos identificadores dos itens de dados que compõem as regras, conforme exemplo:

Regra 1: $A \rightarrow C$
Regra 2: $A, B \rightarrow D$
Regra 3: $C \rightarrow E$
Regra 4: $A, D \rightarrow C$

Supondo que o especialista não esteja interessando em regras em que apresentem o item "D", independentemente se este consta do antecedente ou do consequente, apenas as Regras 1 e 3 serão selecionadas. Vale destacar novamente que essa estratégia somente é interessante quando o especialista tem conhecimento prévio do que deseja que seja contemplado ou eliminado, caso contrário, padrões interessantes podem ser eliminados. Uma alternativa a esta funcionalidade seria eliminar os itens de dados da

base, porém esta alternativa onera computacionalmente a etapa de Pré-processamento, exigindo que um novo conjunto de dados seja construído a cada experimento.

Existem ainda outras estratégias para a eliminação de regras extraídas que não agregam novos conhecimentos ao especialista, como por exemplo, a eliminação de redundância. Por exemplo, a partir do conjunto das Regras 1, 2, 3 e 4 percebe-se uma redundância da Regra 4 em relação a Regra 1, ou seja, a Regra 4 já está contemplada pela Regra 1, desta forma a Regra 4 pode ser eliminada do conjunto.

No que se refere aos padrões descobertos pela tarefa de classificação também existem diversas formas para pós-processar os padrões extraídos, entre elas: transcrição da árvore de decisão em regras, eliminação de redundância e atribuição de medidas de interesse, por exemplo, a partir de generalizações sucessivas.

A transcrição da árvore de decisão visa facilitar a compreensibilidade dos padrões extraídos e é realizada da seguinte forma: a quantidade de nós-folha caracteriza o tamanho do conjunto de regra de associação; o caminho até o nó-folha é o antecedente da regra; o nó-folha (ou classe) é o consequente da regra. A figura 1 apresenta um exemplo de árvore de decisão e na tabela 2 às respectivas regras transcritas.

Em muitos casos, ao transcrever a árvore de decisão em regras de associação ocorre a situação de redundância. É importante destacar que a redundância gerada por árvore de decisão é diferente da gerada pelas regras de associação. Nesta tarefa, as redundâncias ocorrem entre itens do antecedente da regra e não entre as regras, como na tarefa de regras de associação.

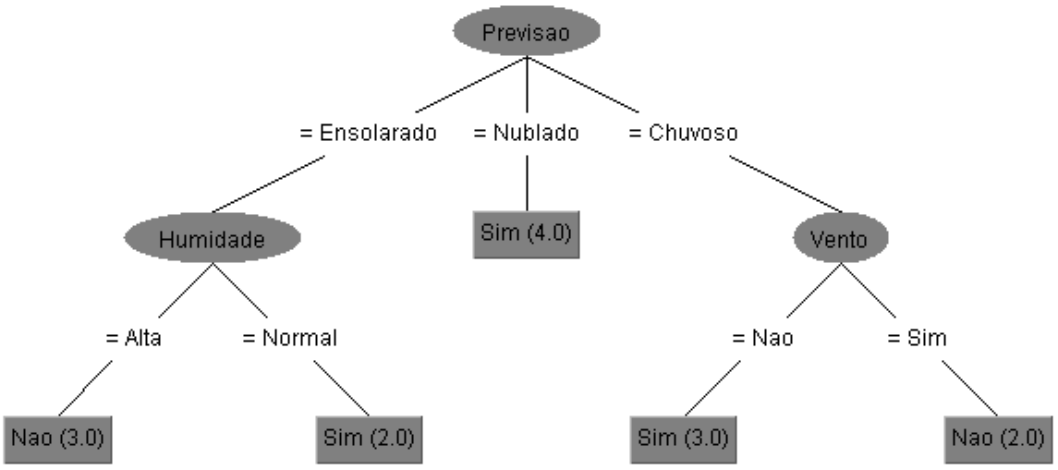

Figura 1. Exemplo de Árvore de Decisão

Num	Regra Transcrita
1	SE previsão=ensolarado E humidade=alta ENTÃO classe=não
2	SE previsão=ensolarado E humidade=normal ENTÃO classe=sim
3	SE previsão=nublado ENTÃO classe=sim
4	SE previsão=chuvoso E vento=sim ENTÃO classe=não
5	SE previsão=chuvoso E vento=não ENTÃO classe=sim

Tabela 2. Regras Resultantes de Transcrição da Árvore de Decisão.

Além destas técnicas pra melhorar a compreensibilidade das regras extraídas da árvore decisão, podem ser atribuídos graus de interesse para cada regra. Existem diversas formas de atribuir essa medida, dentre elas, uma medida baseada em generalizações sucessivas. Essa medida esta baseada na seguinte fórmula (Carvalho, 2005):

$$\frac{numero\ de\ generalizações\ que\ alteram\ o\ consequente}{total\ de\ condições\ da\ regra}$$

Entende-se por "generalização que altera o consequente" o ato de suprimir gradativamente as condições (atributo-condição-valor) do antecedente da regra e a respectiva classe predita ser alterada em relação à regra original. A partir do desmembramento do antecedente da Regra 2 (Tabela 2) tem-se as seguintes condições C1: previsao=ensolarado e C2: humidade=normal.

Para calcular a respectiva medida para a Regra 2 foram suprimidas, uma por vez, cada uma das duas condições, ou seja, foram realizadas generalizações sucessivas sobre o antecedente da regra. Quando da supressão da C1 a classe predita foi alterada de "sim" para "não" e quando da supressão de C2 a classe permaneceu a mesma. Logo, o total de generalizações que alteraram a classe predita é igual 1 e o total de condições da regra é igual a 2, portando, o grau de interesse desta regra é 0.5.

3. MATERIAIS E MÉTODOS

As tarefas de Mineração de Dados, descoberta de regras de associação e classificação foram aplicadas a partir dos algoritmos APRIORI (BORGELT, 2004) e J48 (HALL et al, 2009), respectivamente. Para realizar os experimentos, foram utilizadas duas bases de dados, ambas disponibilizadas no ambiente WEKA. Para a tarefa de descoberta de regras de associação foi utilizada a base "*weather*" (Tabela 3), já para a tarefa de classificação foi utilizada a base "*iris*". A escolha se deve ao fato abordarem um domínio de fácil compreensão, possibilitando ao leitor a replicação dos resultados.

Previsão	Temperatura	Humidade	Vento	Jogar
Ensolarado	Quente	Alta	Não	Não
Ensolarado	Quente	Alta	Sim	Não
Nublado	Quente	Alta	Não	Sim
Chuvoso	Moderada	Alta	Não	Sim
Chuvoso	Frio	Normal	Não	Sim
Chuvoso	Frio	Normal	Sim	Não
Nublado	Frio	Normal	Sim	Sim
Ensolarado	Moderada	Alta	Não	Não
Ensolarado	Frio	Normal	Não	Sim
Chuvoso	Moderada	Normal	Não	Sim
Ensolarado	Moderada	Normal	Sim	Sim
Nublado	Moderada	Alta	Sim	Sim
Nublado	Quente	Normal	Não	Sim
Chuvoso	Moderada	Alta	Sim	Não

Tabela 3. Base de Dados "*weather*".

O conjunto de dados *"íris"* contendo 150 instâncias e 5 atributos, sendo um deles a classe a ser prevista é apresentado de forma sucinta na tabela 4.

comprimentoSepala	larguraSepala	comprimentoPetala	larguraPetala	Classe
5.1	3.5	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica

Tabela 4. Instâncias da Base "*íris*".

Considerando a tarefa de descoberta de regras de associação, foram utilizadas quatro técnicas de Pós-processamento, a saber: filtro de regras de associação, eliminação de redundâncias, um "descobridor de regras de exceção" (DRE) que dependendo da opção do usuário atribui ou não o grau de interesse proposto por (SUZUKI, 2000) para cada par de regras exceção encontrado.

Essas técnicas foram escolhidas por terem sido destacadas como importantes em experimentos anteriormente realizados (DALAGASSA, 2009), (KOBUS, 2006).

Para a tarefa de classificação, foram adotadas três técnicas: transcrição da árvore de decisão nas respectivas regras, eliminação de redundância e atribuição de grau de interesse por generalizações sucessivas. A transcrição da árvore de decisão constitui requisito fundamental para facilitar o Pós-processamento que passa a trabalhar sobre estruturas de baixa complexidade de representação.

4. RESULTADOS

Sobre a base de dados demonstrada na Tabela 3, composta por cinco atributos e quatorze instâncias, foi aplicado o algoritmo APRIORI (BORGELT, 2004), com suporte mínimo de 1% e confiança mínima de 60%, sendo extraídas 313 regras de associação (CONJ1). Apesar do CONJ1 não ser considerado muito grande, analisar 313 regras não é uma tarefa simples.

O CONJ1 foi submetido ao processo e eliminação de regras redundantes, no qual foram reduzidas 192 regras, restando assim, apenas 121 regras do conjunto inicial (CONJ2). Um exemplo de redundância nas regras extraídas desta base é:

Regra A: Se temperatura=quente então humidade=alta
Regra B: Se temperatura=quente e previsão=ensolarado então humidade=alta
Onde a Regra B é eliminada por ser redundante em relação à Regra A.

Sobre o CONJ2 foi submetido ao Filtro de Regras de Associação selecionando apenas as regras que apresentassem no consequente o item de dado "jogar", independentemente do seu valor. Desta forma o CONJ2 foi reduzido para 80 regras (CONJ3).

Por fim, o CONJ3 foi submetido ao DRE com atribuição do grau de interesse. Vale destacar que o CONJ1 (conjunto inicial) poderia ter sido submetido ao DRE, entretanto padrões não relacionados ao foco comporiam o conjunto final de padrões a ser oferecido ao especialista, o que não seria desejável dado que exigiria um esforço adicional para a análise e interpretação dos resultados. Entre as 80 regras (CONJ3) foi possível identificar 10 pares de regras gerais e suas respectivas regras de exceção (CONJ4).

A seguir é apresentado um destes pares (CONJ4):

Regra Geral:
Regra G: Se jogar=sim então vento=não
Regras de Exceção:
Regra E1: Se jogar=sim e temperatura=frio e previsão=nublado então vento=sim / Grau de Interesse: 0.086
Regra E2: Se jogar=sim e humidade=alta e temperatura=moderada e previsão=nublada então vento=sim / Grau de Interesse: 0.121

Analisando a Regra G e suas regras de exceção (E1 e E2), percebe-se que a Regra E2 possui um Grau de Interesse maior, portanto, tem mais chance de ser mais interessante para o especialista.

O gráfico 1 apresenta a dinâmica da cardinalidade dos subconjuntos criados a partir das sucessivas aplicações das estratégias de Pós-processamento sobre as regras de associação descobertas (CONJ1).

Esta base foi submetida ao algoritmo J48 (HALL et al, 2009) o qual descobriu a árvore de decisão (Figura 2), contendo 11 nós folhas, que representam 11 regras, pois o percurso entre o nó raiz e o nó folha caracteriza uma regra.

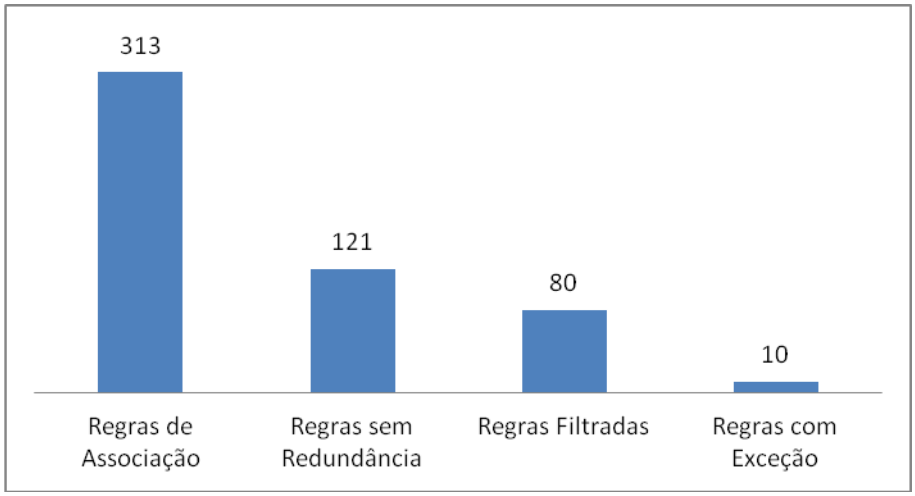

Gráfico 1. Número de regras resultantes após aplicação sucessiva das estratégias de Pós-processamento.

```
larguraPetala <= 0.6: Iris-setosa (50.0)
larguraPetala > 0.6
|   larguraPetala <= 1.7
|   |   comprimentoPetala <= 4.9: Iris-versicolor (48.0/1.0)
|   |   comprimentoPetala > 4.9
|   |   |   larguraPetala <= 1.5
|   |   |   |   larguraPetala <= 1.35: Iris-versicolor (0.0|28.0)
|   |   |   |   larguraPetala > 1.35
|   |   |   |   |   comprimentoSepala <= 5.95: Iris-versicolor (0.0|25.0)
|   |   |   |   |   comprimentoSepala > 5.95: Iris-virginica (3.0)
|   |   |   larguraPetala > 1.5
|   |   |   |   comprimentoPetala <= 5.85
|   |   |   |   |   comprimentoSepala <= 7.25
|   |   |   |   |   |   larguraSepada <= 2.65: Iris-virginica (0.0|7.0)
|   |   |   |   |   |   larguraSepada > 2.65
|   |   |   |   |   |   |   larguraSepada <= 3.05: Iris-versicolor (3.0/1.0)
|   |   |   |   |   |   |   larguraSepada > 3.05: Iris-virginica (0.0|19.0/2.0)
|   |   |   |   |   comprimentoSepala > 7.25: Iris-virginica (0.0|8.0)
|   |   |   |   comprimentoPetala > 5.85: Iris-virginica (0.0|13.0)
|   larguraPetala > 1.7: Iris-virginica (46.0/1.0)
```

Figura 2. Árvore Gerada pelo Algoritmo J48 para a Base "*iris*".

Esta árvore foi submetida ao programa PAD – Pós-processamento de Árvores de Decisão resultando em 11 regras transformadas, entre elas:

Regra T1: SE larguraPetala>0.6 E larguraPetala>1.7 ENTÃO Iris-virginica
Regra T2: SE larguraPetala>0.6 E larguraPetala<=1.7 E comprimentoPetala <=4.9 ENTÃO Iris-versicolor
Regra T3: SE larguraPetala<=0.6 ENTÃO Iris-setosa

Ao analisar a Regra T1 é possível perceber a redundância em função das duas condições construídas sobre o mesmo atributo "largura Petala", ou seja, após a eliminação de redundâncias a Regra T1 passa a ser:

Regra T1: SE larguraPetala>1.7 ENTÃO Iris-virginica

A partir da aplicação desta estratégia de Pós-processamento o número médio de condições por regra passou de 5.2, para 3.90. Ou seja, houve uma redução da ordem de 25% no número médio de condições a ser analisado pelo usuário, o que facilita a descoberta de conhecimento, principalmente em relação ao tempo de análise. Na tabela 5 são listados os atributos e suas respectivas condições descobertas que mais apresentaram redundâncias.

Atributo-valor	Ocorrências	Eliminações
larguraPetala>0.6	10	8
larguraPetala<=1.7	9	3
comprimentoPetala>4.9	8	1
larguraPetala>1.5	5	0
comprimentoPetala<=5.85	4	0
larguraPetala<=1.5	3	1
comprimentoSepala<=7.25	3	0
larguraSepada>2.65	2	1
larguraPetala>1.35	2	0
larguraSepada<=3.05	1	0
larguraSepada>3.05	1	0
larguraSepada<=2.65	1	0
comprimentoSepala<=5.95	1	0
comprimentoSepala>5.95	1	0
comprimentoSepala>7.25	1	0
larguraPetala<=1.35	1	0
comprimentoPetala>5.85	1	0
comprimentoPetala<=4.9	1	0
larguraPetala>1.7	1	0

Tabela 5. Condições com maior frequência de redundância.

A partir da análise da tabela 5 frente à árvore descoberta (Figura 2) é possível perceber que as condições dispostas nos níveis mais próximos ao nó-raiz, tendem a compor redundâncias, dado o fato de constarem em um número maior de ramificações (regras).

O conjunto de 11 regras após a eliminação das redundâncias foi submetido à estratégia de atribuição de grau de interesse a partir de generalizações sucessivas. Das 11 regras, a que apresentou maior grau de interesse foi a regra T4:

Regra T4: Se larguraPetala <= 1.7 e comprimentoPetala > 4.9 e larguraPetala>1.5 e comprimentoPetala <= 5.85 e comprimentoSepala > 7.25 então Iris-virginica, com um Grau de Interesse de 0.4.

As Regras T1, T2 e T3, apresentaram grau de interesse 0 (zero). Estas tenderiam a ser regras com baixo potencial de interesse ao usuário, entretanto, vale destacar que essas medidas são *data-driven*, levando em consideração apenas a estrutura do conhecimento extraído e o conjunto de dados.

5. CONCLUSÃO E TRABALHOS FUTUROS

Este artigo descreveu e relatou resultados de experimentações de estratégias de Pós-processamento em KDD, sobre conjuntos de dados de baixa complexidade, disponíveis em repositório públicos, permitindo assim que interessados em aprofundar o conhecimento repliquem tais experimentações, não apenas nas bases propostas, mas também em bases disponíveis em seus espaços de trabalho e/ou pesquisa. Os algoritmos de Mineração de Dados são públicos e podem ser obtidos a partir das referências indicadas. Os algoritmos de Pós-processamento também podem ser obtidos a partir de contato via endereço eletrônico com qualquer um dos autores deste artigo.

Foram demonstradas experimentações sobre padrões descobertos a partir de duas das três tarefas mais usuais: descoberta de regras de associação e classificação. Com relação à descoberta de regras de associação, mesmo considerando um conjunto original

contendo 14 instâncias, foram descobertas 313 regras. Apenas a partir da eliminação de regras redundantes, ou seja, sem potencial para agregar conhecimento do usuário este conjunto teve uma redução para 121 regras, ou seja, apenas 38% do conjunto original de regras descobertas. Fica fácil perceber que o usuário seria "poupado" de avaliação de 192 regras. Também foi demonstrado que o usuário ao sinalizar o item de dado de maior interesse também permite uma redução ainda maior, chegando ao limite de ter apenas um conjunto de 10 pares de regras gerais e suas respectivas exceções. Ou seja, sai de 313 regras para apenas 10 pares.

Quanto ao Pós-processamento de padrões descobertos e posteriormente representados na forma de árvore de decisão, também foi possível perceber que a simples transformação da árvore em regras e posterior eliminação de condições redundantes reduziu o número médio de condições por regra da ordem de 25%. Vale destacar que em Dalagassa (2009) esta atividade foi desenvolvida de forma manual sobre uma árvore de decisão contendo mais de 1000 ramificações que além da transformação em regras, também tiveram eliminadas as condições redundantes.

O presente trabalho também apresenta, testa e disponibiliza uma ferramenta que além de transformar, eliminar condições redundantes, também descobre regras de exceção e atribui seus respectivos graus de interesse.

Desta forma, este artigo permite novas iniciativas sejam oportunizadas com o objetivo de popularizar ainda mais a utilização do processo KDD no dia a dia das instituições que dispõem de conjuntos de dados e desejam melhor utilizar o seu potencial para apoiar o processo decisório.

Como trabalhos futuros fica a sugestão que novas ferramentas experimentações sejam realizadas envolvendo estratégias mais fortemente relacionadas a critérios *user-driven*, bem como sejam realizadas experimentações sobre bases de dados reais com posterior avaliação por parte de especialistas da área.

REFERÊNCIAS

AGRAWAL R.; IMIELINSKI T.; SWAMI A. Mining Associations between Sets of Items in Massive Databases. Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data, Washington D.C., May 1993, p.207-216.

BORGELT C. APRIORI – ASSOCIATION RULE INDUCTION. 2004. Disponível em: http://www.borgelt.net/apriori.html.

CARVALHO D.R. Algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados [tese]. Rio de Janeiro: COPPE - UFRJ; 2005.

CARVALHO D. R.; MOSER, A. D.; DA SILVA, V. A; DALLAGASSA, M. R. Mineração de Dados Aplicada à Fisioterapia. Fisioter Mov. 2012, jul/set; 25(3):595-605.

DALLAGASSA M. R. Concepção de uma metodologia para identificação de beneficiários com indicativos de diabetes mellitus tipo 2 [dissertação] Curitiba: Pontifícia Universidade Católica do Paraná; 2009.

FAYYAD U.; PIATETSKY-SHAPIRO G.; SMYTH P.; UTHURUSAMY R. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence. Menlo Park, CA: MIT Press. 1996.

FREITAS A. A. On objective measures of rule surprisingness. Principles of Data Mining & Knowledge Discovery (Proc. 2nd European Symp., PKDD'98. Nantes, France, Sep. 1998). LNAI 1510, 1998. 1-9. Springer-Verlag.

GOMES H., HAUGT L. G. Mineração de Dados Temporal: Descobertas de Regras De Causa e Efeito. [trabalho de conclusão de curso]. Curitiba: Universidade Tuiuti do; 2010.

HALL M.; FRANK E.; HOLMES G.; PFHRINGER B.; REUTEMANN P.; WITTEN I. an. WEKA - THE WEKA DATA MINING SOFTWARE. 2009. Disponível em: http://www.cs.waikato.ac.nz/ml/weka

HUSSAIN F.; LIU H.; SUZUKI E.; LU H. EXCEPTION RULE MINING WITH RELATIVE INTERESTINGNESS MEASURE. PAKDD. 2000; 1805(1): 86-97.

KLOSGEN, W. Patterns for Knowledge Discovery in Databases. Proc. Of Machine Learning. UK. 1992, p. 1-9.

KOBUS L. C. G. Aplicação da descoberta de conhecimento em base de dados para identificação de usuários com doenças cardiovasculares elegíveis para programas de gerenciamento de caso [dissertação de mestrado]. Curitiba: Pontifícia Universidade Católica do Paraná; 2006.

KUBAT M.; BRATKO I.; MICHALSKI R. S. A Review of Machine Learning Methods, in Michalski, R.S., Bratko, I. and Kubat, M. (Eds.), Machine Learning and Data Mining: Methods and Applications, London: John, 1998.

SILBERSCHATZ A.; TUZHILIN A. What makes patterns interesting in knowledge discovery systems. IEEE Trans. Knowledge & Data Eng. 8(6). 1996.

VIANNA R.C.X.F.; MORO C.M.C.B.; MOISES S.J; CARVALHO D.R.; NIEVOLA J.C. Mineração de dados e características da mortalidade infantil. Cadernos de Saúde Pública. 2010;26(3):535-42.

VON STEIN Jr. A.; MALUCELLI A.; BASTOS L.C.; CARVALHO D.R.; CUBAS M.R.; PARAISO E.C. Classificação de microareas homogêneas de risco com uso de mineração de dados. Revista de Saúde Pública, 2010;44(2):292-300.

ZHANG L.; LIU B.; BENKLER J.; ZHOU C. Finding Actionable Knowledge via Automated Comparison . International Conference on Data Engineering - ICDE , 2009. p. 1419-1430.