Espacios. Vol. 37 (Nº 06) Año 2016. Pág. 24

Criação de Conhecimento através de Data Mining: Um estudo sobre regras de associação em uma base de dados do Weka

Knowledge creation through Data Mining: A study of association rules in a Weka database

Jovani Taveira de SOUZA 1; Antônio Carlos de FRANCISCO 2; João Luiz KOVALESKI 3; Bruno Aparecido OLIVEIRA 4; Álamo Alexandre da Silva BATISTA 5; Maria Helene Giovanetti CANTERI

Recibido: 20/10/15 • Aprobado: 25/11/2015

Contenido

RESUMO:
Objetivou-se com este estudo apresentar uma das tarefas de aplicação utilizada pelo Data Mining. Para isso, foi preciso utilizar uma base de dados fornecida pela biblioteca do Weka. A regra utilizada foi a regra de associação. A metodologia é baseada no processo KDD (Knowledge Discovery in Database), que visa a preparação dos dados, procura por padrões, avaliação e refinamento. Como resultado evidenciamos a íntima relação entre os sintomas prévios e os diagnósticos realizados. A adoção das técnicas de KDD apresentou-se como uma estratégia eficaz para descoberta de conhecimento em base de dados, extraindo informações ocultas e gerando dados privilegiados.
Palavras chave: Data Mining, Regra de Associação, KDD.

ABSTRACT:
The objective of this study present an application of the tasks used by Data Mining. For this, we need to use a database provided by the Weka library. The rule used was the association rule. The methodology is based on the KDD process, aimed at preparing the data, looking for patterns, evaluation and refinement. As a result we noted the close relationship between the previous symptoms and those difficulties. The adoption of KDD techniques presented itself as an effective strategy for knowledge discovery in databases, extracting hidden information and generating privileged data.
Key-words: Data Mining, association rules, KDD.

1. Introdução

Com o aumento gradativo de informações em banco de dados surgiu a necessidade de criar técnicas e ferramentas capazes de transformar esses dados brutos em conhecimento útil e aplicado. Visando esse estudo, a área denominada de Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases* –KDD) estudou essas técnicas, juntamente com a técnica capaz de transformar todos os dados gerados em conhecimento útil, intitulado como *Data Mining* ou Mineração de Dados (FAYYAD et al., 1996). O processo de KDD está concentrado no mapeamento de dados brutos em modelos mais compactos, genéricos ou úteis que os dados originais (MELO, 2010).

Nesse aspecto, a Mineração de Dados contribui com essa descoberta de conhecimento, pois,

através de técnicas e algoritmos, ajuda a buscar correlações importantes entre os dados (FAYYAD et al., 1996).

Com as técnicas de *Data Mining,* grandes empresas têm conquistado de forma inteligente um diferencial frente aos concorrentes, tais como: prevenção de tendências, aprimoramento de produtos ou serviços, dependente do perfil de consumo dos mesmos e, além de tudo, maximando seus lucros em face da grande competividade entre eles. Grandes empresas nacionais e internacionais, como Itaú, Telefônica, Sprint, Golden Cross, entre outras, aderem a essa técnica nos dias atuais (VIANA, 2013).

Devido à grande quantidade de informações em base de dados, ferramentas computacionais, especialmente os modelos quantitativos de análise de dados, foram requeridas para identificar elementos importantes e necessários para tomada de decisão (PRAHALAD; KRISHNAN, 2008).

Suporte computacional em atendimento médico é a maneira mais viável para se trabalhar com grandes quantidades de dados, visto que os atendimentos geram grandes massas de dados com informações digitalizadas dos pacientes. Essa massas de dados auxiliam o apoio ao diagnóstico médico (COSTA; TRAINA, 2012). Os dados, se utilizados adequadamente, influenciam diretamente a maneira de como o paciente será tratado, isto é, a forma de cuidado da saúde do paciente. Os dados usados adequadamente podem conceber informações que auxiliem na prevenção e no combate às doenças (LAVRAC et al., 2000).

Sendo assim, este artigo estuda a aplicação da tarefa regras de associação a partir de *Data Mining* em uma base de dados escolhida do software Weka (*Waikato Enviroment for Knowledge Analysis*), software utilizado para realização deste estudo e que será melhor detalhado na Seção 3. A base escolhida será a base *Breast Cancer*, na qual objetiva-se mostrar pacientes com câncer de mama que possuem ou não recorrência de sintoma após o tratamento. No entanto, o artigo em questão apresenta apenas a aplicação da regra e os resultados a partir dessa aplicação, sem se aprofundar em questões específicas da base de dados.

Tendo em vista a utilização das tarefas de aplicação de *Data Mining*, é possível extrair informações ocultas e gerar dados que favorecem a descoberta de conhecimento.

O artigo está organizado da seguinte maneira: na primeira seção, é apresentada a introdução, na qual se exibe um breve resumo sobre o assunto, seguida do referencial teórico, que relata o processo de descoberta de conhecimento em banco de dados, fases do processo de KDD, algoritmos utilizados e regras empregadas. A seção 3 apresenta os procedimentos metodológicos utilizados no trabalho. Na seção 4 são evidenciados os resultados obtidos e, em seguida, são relatadas as conclusões do trabalho.

Pretende-se com esse estudo articular não só os fundamentos teóricos referentes à descoberta de conhecimento, mas também ressaltar a importância e os benefícios gerados a partir dessas técnicas, especificamente a regra de associação por *Data Mining*.

2. Referencial Teórico

2.1 KDD- *Knowledge Discovery in Databases*

O processo de KDD (*Knowledge Discovery in Databases*) surgiu em 1989 como um novo ramo da computação que visava à extração de conhecimento, de maneira automatizada, e através do mesmo, explorar as crescentes bases de dados, criando, assim, relações de interesse e reconhecimento de padrões existentes, através da modelagem de fenômenos do mundo real (GOLDSCHIMIDT; PASSOS, 2005).

Junto às bases de dados, existem informações relevantes contidas em nível gerencial e estratégico, que em hipótese nenhuma podem ser descobertas por sistemas de gerenciamento de base de dados tradicionais (DIAS et al., 2015).

Concomitantemente ao processo de KDD, encontra-se a técnica de *Data Mining* (Mineração de

Dados), que, segundo Fayyad et al. (1996), define como "processo não trivial de identificar em dados alguns padrões válidos, novos, potencialmente úteis e compreensíveis". O autor também destaca a versatilidade da mineração de dados em diferentes segmentos e setores (FAYYAD et al., 1996).

O processo do KDD é um processo iterativo e interativo, composto por uma série de etapas, o qual envolve a prepação dos dados, procura por padrões, avaliação e refinamento (FAYYAD et al., 1996). Essas etapas serão especificadas a seguir.

2.1.1 Fases principais do KDD

Para Fayyad et al. (1996), as fases principais do processo de KDD são: seleção, pré-processamento e limpeza, transformação, *data mining*, intepretação e avaliação. A Figura 1 representa o processo.

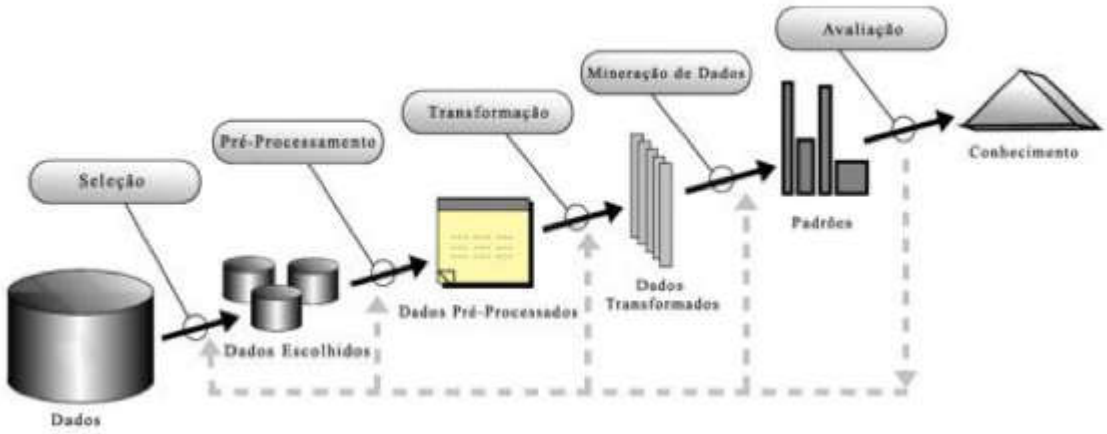

Figura 1- Processo do KDD

Fonte: Fayyad et al, 1996

Esse ciclo corresponde ao processo que o dado percorre até se transformar em conhecimento útil. As fases do processo ocorrem logicamente a partir de uma base de dados, e essas fases têm objetivos específicos para cada tipo de base, porém as finalidades são as mesmas. Seguidamente, serão apresentadas todas as etapas do processo, tendo como enfoque suas características e finalidades.

2.1.1.1 Seleção

Etapa que tem como propósito identificar os dados significantes a serem trabalhados e que podem ser encontrados em vários formatos. Fase na qual é mapeado o conjunto de informações a serem utilizadas.

É realizada uma identificação de quais informações serão trabalhadas na base de dados (GOLDSCHIMIDT; PASSOS, 2005, p.26).

Um aspecto importante referente à escolha das informações significantes é a busca por dados que satisfaçam seus objetivos. Por exemplo, na área médica, pode-se tentar encontrar as possíveis causas da doença de um paciente. Uma escolha incerta pode levar a informações errôneas, prejudicando, assim, a tomada de decisão (BATISTA, 2003).

De acordo com Diniz (2000), "as variáveis selecionadas para a mineração de dados são denominadas variáveis ativas, uma vez que são usadas para distinguir segmentos, fazer predições ou desenvolver outras operações específicas de mineração de dados".

2.1.1.2 Pré-processamento e Limpeza

Nessa fase ocorre o tratamento dos dados, ou seja, são realizadas tarefas envolvendo a limpeza de dados, objetivando a inconsistência, redundância e ausência entre os mesmos.

A qualidade é crucial para obter dados confiáveis, uma vez que dados limpos e compreensíveis são requisitos básicos para o sucesso da mineração (DINIZ, 2000).

Por último, os dados são selecionados e analisados para serem transformados.

2.1.1.3 Transformação

Após o pré-processamento dos dados, os mesmos são transformados para um formato adequado, para que possa ser trabalhado da forma apropriada para mineração. Por exemplo, para ser trabalhado com a ferramenta Weka, os dados precisam ser convertidos para o formato ARFF (*Attribute-Relation File Format*), que é trabalhado pelo programa. É comum empresas de grande porte terem diversos computadores, muitos deles utilizando diferentes sistemas operacionais, o que faz com que os dados necessitem ser armazenados e formatados corretamente, para que o programa possa ser aplicado.

Segundo Melo (2010), além de localizar características úteis para representar os dados, essa etapa é também responsável pela escolha dos melhores atributos presentes no conjunto de dados.

2.1.1.4 *Data Mining*

Após a etapa de transformação dos dados, é a vez do *Data Mining* desempenhar seu papel, que é central na extração do conhecimento.

De acordo Fayyad et al. (1996), a mineração de dados é um método que busca o conhecimento em grandes bases de dados, fazendo com que as informações importantes ajudem no processo da tomada de decisão. Além disso, essa etapa visa à escolha dos algoritmos que mais combinam com o objetivo proposto, aquele que se quer extrair da base de dados escolhida.

Conforme Berry e Linoff (1997), um dos objetivos do *Data Mining* é a descoberta de regras e padrões, que é feita através da exploração e análise, de forma automática ou semi-automática, das grandes bases de dados.

De acordo com Berson e Smith (1997) e Ramos (1999), a mineração de dados pode ser realizada através de algumas tarefas, como: agrupamento, associação e classificação. No item 2.2 será apresentada a tarefa de agrupamento, que será aplicada no desenvolvimento do trabalho.

Para cada tipo de problema existe alguma regra e um algoritmo em específico que melhor se encaixam nas situações exigidas.

Por fim, tem-se a interpretação do resultado e a avaliação para determinar a eficácia do conhecimento extraído.

2.1.1.5 Interpretação e Avaliação

A última fase do processo do KDD é a avaliação e interpretação, que verifica se o conhecimento adquirido a partir das etapas realizadas anteriormente atingiu o objetivo principal.

A fase de pós-processamento, como é conhecido, também avalia os dados explorados a partir dos algoritmos da mineração de dados. Caso aconteça de o resultado não for o esperado, o processo pode ser modificado a partir das etapas anteriores, e corrigidas posteriormente. Uma atitude comum é a troca de algoritmo.

Além da visualização dos padrões adquiridos, são medidas também informações que correspondem a erro médio, erro quadrático, taxas de falsos positivos, matriz confusão, precisão, entre outras.

Ao final, ter-se-ão os padrões extraídos com a finalidade de acatar os objetivos inicialmente propostos.

2.2 Regras de associação

As regras de associação determinam quais conjuntos ocorrem de forma simultânea em um mesmo evento, sendo assim, são resultados da relação entre os dados processados, associações de condição e resultado (BRUSSO, 2000).

As relações existentes entre os valores do conjunto da base de dados são pertinentes na regra de associação, pois contribuem no processo de tomada de decisão. É importante destacar a importância das regras existentes no *Data Mining* como um todo, ou seja, escolhendo a regra certa, provavelmente as relações ajudarão a se ter respostas mais precisas, extraindo o máximo

dos dados ofertados.

Um dos determinantes de grande importância na regra de associação é o de suporte (frequência que os padrões ocorrem) e confiança, pois apenas as regras de associação com um alto valor de confiança e de suporte são as regras de associação vistas para uma determinada pesquisa. Para as regras de associação serem efetivadas, é preciso algoritmos que executem as tarefas programadas. Um algoritmo que é comum e eficaz em regras de associação é o algoritmo *Apriori*. A seguir, será feito um leve aprofundamento sobre os conceitos deste algoritmo.

2.3 Algoritmo *Apriori*

Algoritmo usado para realizar tarefas de associação, que busca em um banco de dados conjuntos frequentes, que satisfazem uma regra mínima estabelecida. Esse algoritmo identifica relações e dependências significativas entre os atributos (DIAS et al., 2015). Ele trabalha com as medidas de confiança e suporte para classificar as melhores regras (CORRÊA, 2009).

O algoritmo soluciona problemas envolvendo mineração de dados com conjuntos de itens frequentes, atributos que se inter-relacionam, além de realizar diversas buscas no banco de dados.

O primeiro passo deste algoritmo é encontrar conjuntos de itens frequentes. Primeiramente, o usuário dá um limite mínimo para o apoio e o algoritmo pesquisa todos os conjuntos de itens que aparecem com um apoio superior a esse limite (PASTA, 2011), ou seja, ao especificar o suporte mínimo, que condiz com os conjuntos de itens que ocorrem em pelo menos uma dada percentagem, o algoritmo seleciona as transações acima desse suporte pré-estabelecido. No próximo passo, são construídas as regras dos itens selecionados anteriormente, e o algoritmo calcula a confiabilidade de cada regra e as mantêm de acordo com a confiança limiar definida pelo usuário, ou seja, apenas mantém as regras em que a confiança é maior do que o limite dado pelo usuário. A tarefa de associação identifica e descreve as associações entre as variáveis no mesmo item ou associações entre os itens diferentes que ocorrem concomitantemente, de uma forma frequente em bases de dados (PASTA, 2011).

O algoritmo *Apriori* realiza buscas sucessivas em toda a base de dados, mantendo um ótimo desempenho em termos de tempo de processamento (AGRAWAL; SRIKANT, 1994).

Após serem passados os conceitos envolvendo as etapas do KDD e as regras que serão utilizadas no trabalho, o tópico seguinte visa fazer uma breve explicação sobre a base de dados escolhida.

2.4 Base de dados

A base de dados denominada de Breast Cancer foi originada através do Centro Médico Universitário de Oncologia, de Ljubljana, Iugoslávia. O conjunto possui 201 instâncias de pacientes com câncer sem recorrência e 85 pacientes nos quais o câncer retornou em até um ano após o diagnóstico e tratamentos feitos inicialmente, além das instâncias serem descritas por nove atributos.

O Quadro 1 apresenta as informações a respeito dos atributos, descrições e o tipo de dado.

Atributo	Descrição	Tipo de dado
Age	Idade do indivíduo	Integer (10-19 até 90-99)
Menopause	Indica se o paciente é pré ou pós-menopausa no momento do diagnóstico	Varchar (lt40 (antes dos 40), ge40 (depois dos 40), premeno (pré-menopausa))
Tumor-size	Diâmetro do tumor (mm)	Integer(0-4 até 55-59)

Inv-nodes	Número de linfonodos axilares	Integer (0-2 até 36-39)
Node-caps	Penetração do tumor na cápsula do linfonodo	Varchar (Yes, No)
Deg-malig	Grau de malignidade do tumor	Integer (1,2,3)
Breast	Mama em que o câncer pode ocorrer	Varchar (Left, Right)
Breast-quad	Quadrante da mama afetado considerando o bico como o centro	Varchar (Left-up, Left-low, Right-up, Right-low, Central)
Irradiat	Histórico de radioterapia	Varchar (Yes, No)
Class	Recorrência ou não após o tratamento	Varchar (no-recurrence-events, recurrence-events)

Quadro 1– Descrição da base de dados Breast Cancer

A partir das informações referentes à base de dados, o andamento do trabalho será apresentando nas próximas seções, contemplando-se o desenvolvimento, resultado e conclusão do estudo.

3. Desenvolvimento

Com a evolução da tecnologia, a análise correta dos dados se tornou uma variante importante para várias organizações. Um dos aspectos responsáveis por essa importância é a otimização de dados, que busca, além da confiabilidade, informações que são importantes e úteis e que ajudam na busca por mais qualidade e produtividade dentro da empresa, além de competitividade em seu segmento.

Através da correta análise de dados, as informações pertinentes são cabíveis para tomar decisões. Com isso, o ramo da medicina aderiu a conceitos que visam a essa otimização, em virtude da grande quantidade de informações e registros, que nem sempre são realmente importantes para o diagnóstico. Surge, assim, a necessidade de analisar e entender esses dados a partir das técnicas de Data Mining.

Logo, a proposta deste trabalho juntamente com o processo de KDD visa à utilização das técnicas de Data Mining, especificamente a regra de associação. A base de dados, como referida anteriormente, será a base Breast Cancer e que se encontra na biblioteca do Weka, software a ser detalhado no próximo tópico.

3.1 Weka

O Weka (Waikato Enviroment for Knowledge Analysis) é um software livre desenvolvido em linguagem Java, pela Universidade de Waikato, Nova Zelândia, o qual objetiva a técnica de mineração de dados. É formado por um conjunto de implementações de algoritmos de diversas técnicas de Minerações de Dados (UNIVERSITY OF WAIKATO, 2010). Essas implementações correspondem a um conjunto de algoritmos de aprendizado de máquina, que possibilita a extração do conhecimento (WEKA, 1997).

É importante destacar sua utilidade em diversos trabalhos acadêmicos, artigos científicos, dissertações e teses, envolvendo processo de mineração de dados, em virtude de o software

disponibilizar um grande conjunto de recursos para execução dos processos de KDD.

Uma de suas características é a portabilidade, pois pode ser instalada em qualquer computador e em diferentes sistemas operacionais. O software pode ser disponibilizado pelo site http://www.cs.waikato.ac.nz/ml/weka.

O formato para se utilizar o Weka é o arquivo ARFF (Attribute-Relation File Format). A seguir, serão apresentadas explicações sobre este arquivo.

3.2 Arquivo ARFF

Arquivo ARFF (Attribute-Relation File Format) é o formato utilizado pelo software Weka. Este arquivo precisa conter informações a respeito de domínio de atributos, valores que os atributos podem representar e o atributo classe. O arquivo é dividido, primeiramente, por uma lista de todos os atributos de seu estudo, em que se deve escolher o tipo do atributo e/ou valores desejados. Os valores escolhidos precisam estar entre chaves e separados por vírgulas. A segunda parte é formada por instâncias que estão presentes nos dados escolhidos. Os atributos precisam estar separados por vírgula.

Na Figura 2 têm-se as informações da base de dados em questão e a sua formatação no arquivo ARFF.

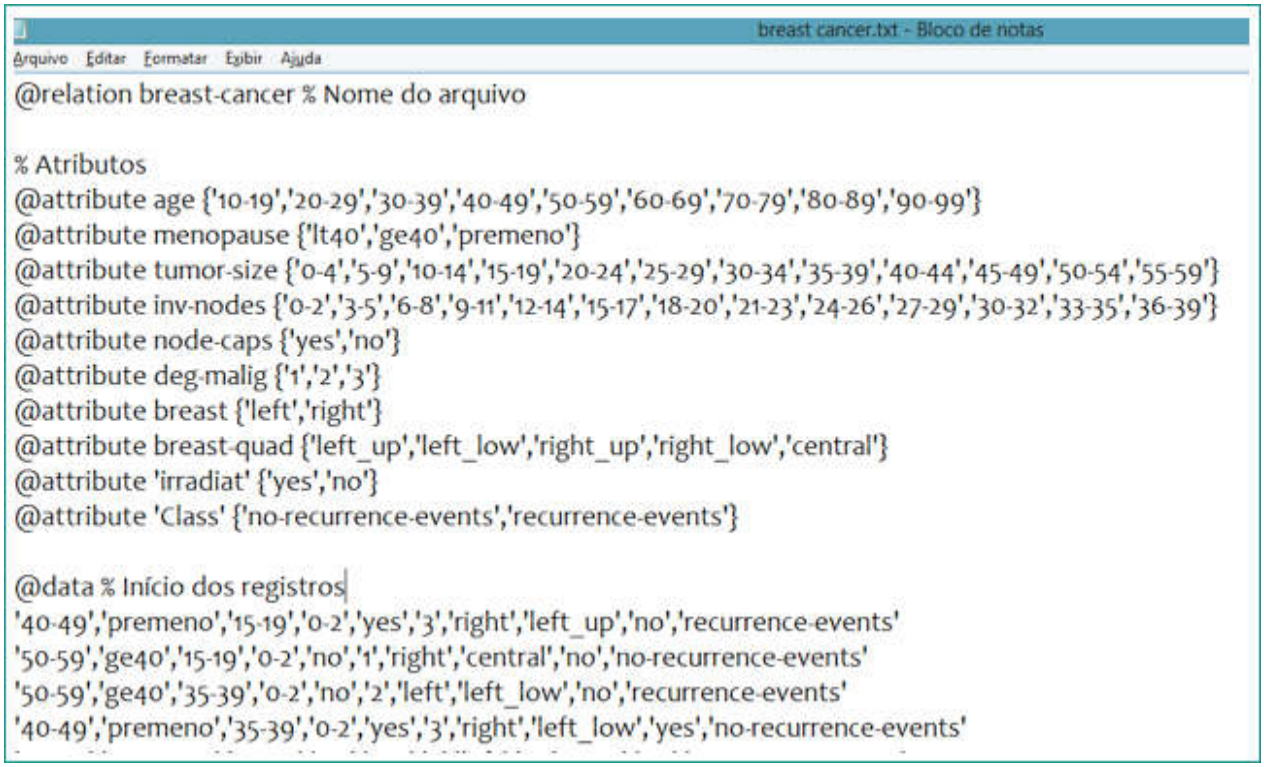

Figura 2 – Arquivo ARFF

Posteriormente à transformação dos dados, o arquivo pode ser executado pelo software Weka. Para utilizar o algoritmo Apriori, seleciona-se primeiramente a aba Associate e escolhe-se o pacote weka.associations.Apriori.

Visando à busca das melhores regras para seu estudo, é necessário selecionar alguns parâmetros do algoritmo. Para isso, é preciso dar um duplo clique no algoritmo, que abrirá uma janela com algumas informações a respeito do mesmo. O Quadro 2 apresenta três parâmetros significativos na aplicação do algoritmo Apriori.

Opção	Função

N° de Regras (*numRules*)	Especifica a quantidade regras desejada. (*default* =12)
Confiança (*minMetric*)	Apresenta a confiança ou a precisão mínima exigida (*default* = 0,9 = 90%)
Suporte Mínimo (*lowerboundminsupport*)	Suporte Mínimo desejado. (*default* = 0,1 = 10%)

Quadro 2– Parâmetros do algoritmo Apriori

Nessa perspectiva, é possível observar que, quando os valores estiverem mais perto de 1.0, ou seja, corresponde a 100%, maior será a confiabilidade dos dados, ou seja, maiores serão a confiança e o suporte da regra.

Nesta seção serão apresentados os resultados obtidos, resultantes das fases já concluídas anteriormente. Neste estudo, o suporte mínimo foi de 10%, confiança de 90% e número de regras igual a 12, aplicados na base de dados *Breast Cancer*. Na Figura 3, são apresentadas as regras obtidas a partir do algoritmo *Apriori*.

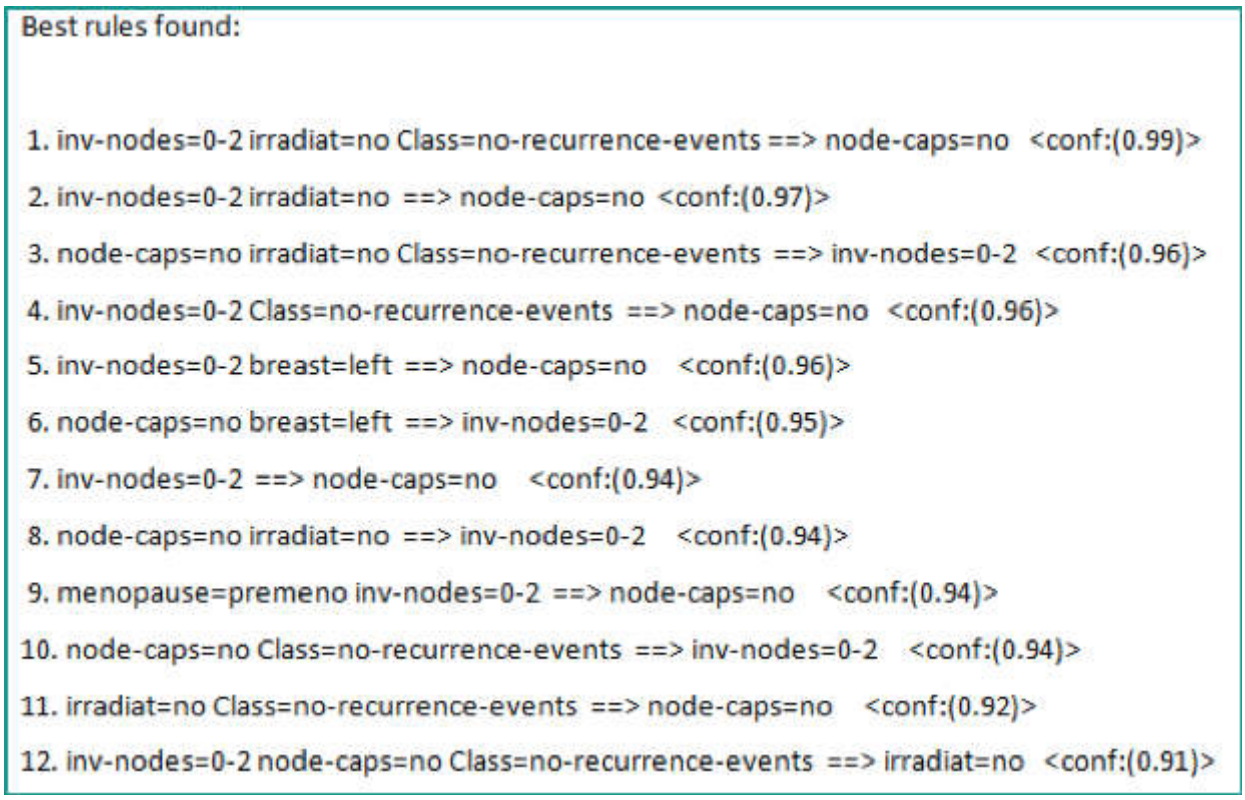

Figura 3 – Resultados gerados pelo algoritmo Apriori

Através da Figura 3, obtêm-se as seguintes informações a respeito das 12 regras encontradas:

1. Se o paciente possui nódulo entre 0-2 mm, não tem histórico de radioterapia e não teve recorrências de sintomas após o tratamento, então o paciente não possui penetração do tumor na cápsula do linfonodo. 99% de confiança;
2. Se o paciente possui nódulo entre 0-2 mm, não tem histórico de radioterapia, então o paciente não possui penetração do tumor na cápsula do linfonodo. 97% de confiança;
3. Se o paciente não possui penetração do tumor na cápsula do linfonodo, não tem histórico de radioterapia e não teve recorrências de sintomas após o tratamento, então o paciente possui nódulo entre 0-2 mm. 96% de confiança.
4. Se o paciente possui nódulo entre 0-2 mm e não teve recorrência após o tratamento,

então não possui penetração do tumor na cápsula do linfonodo. 96% de confiança;

5. Se o paciente possui nódulo entre 0-2 mm na mama esquerda, então não tem penetração do tumor na cápsula do linfonodo. 96% de confiança;
6. Se o paciente não tem penetração do tumor na cápsula do linfonodo na mama esquerda, então o paciente possui nódulo entre 0-2 mm. 95% de confiança;
7. Se o paciente possui nódulo entre 0-2 mm, então não tem penetração na cápsula do linfonodo. 94% de confiança;
8. Se o paciente não tem penetração na cápsula do linfonodo e não tem histórico de radioterapia, então o paciente possui nódulo entre 0-2 mm. 94% de confiança;
9. Se o paciente tiver nódulo entre 0-2 mm e na pré-menopausa, então não possui penetração do tumor na cápsula do linfonodo. 94% de confiança;
10. Se o paciente não possui penetração do tumor na cápsula do linfonodo e não teve recorrência de sintomas após o tratamento, então o paciente possui nódulo entre 0-2 mm. 94 % de confiança;
11. Se o paciente não tem histórico de radioterapia e não teve recorrência de sintomas após o tratamento, então o paciente não possui penetração do tumor na cápsula do linfonodo. 92 % de confiança;
12. Se o paciente tiver nódulo entre 0-2 mm, não possuir penetração do tumor na cápsula fibrosa e não teve recorrência de sintomas após o tratamento, então o paciente não possui histórico de radioterapia. 91 % de confiança.

Com o entendimento das regras acima, é possível obter informações privilegiadas a respeito da base de dados em estudo, ou seja, é possível determinar a relação entre os sintomas prévios e os diagnósticos realizados, sabendo que a regra utilizada é uma das tarefas que pode ser realizada para a descoberta do conhecimento útil; no entanto, existem outras técnicas que podem ser utilizadas com o mesmo propósito. Após a extração do conhecimento, serão apresentadas a seguir as principais conclusões do trabalho.

4. Conclusão

Conclui-se com o estudo que a adoção das técnicas de KDD apresentou-se como uma estratégia eficaz para a descoberta de conhecimento em base de dados, extraindo informações ocultas e gerando dados privilegiados.

Foi possível compreender, com este trabalho, as etapas pertinentes à mineração de dados, especificamente a regra de associação. Esta regra visava à determinação de conjuntos frequentes na base de dados para, posteriormente, encontrar padrões para determinar o perfil e comportamento das informações contidas na base de dados.

A principal contribuição deste trabalho, além de apresentar a aplicabilidade da tarefa de mineração de dados, é mostrar não só os fundamentos teóricos referentes à descoberta de conhecimento, mas também ressaltar a importância e os benefícios gerados a partir dessas técnicas. É importante mencionar sua importância à comunidade acadêmica, pois contribui para estudos iniciais envolvendo *Data Mining*, regras de associação e etapas do KDD.

Sugere-se, para trabalhos futuros, estudar outras técnicas de *Data Mining* para a mesma base de dados, além de estudar qual é a técnica que mais contribuiu para a descoberta de conhecimento. Também pode ser feito o aprofundamento da base de dados, visando relatar questões específicas sobre o mesmo.

5. Referências

ABDEL-AAL, R.E.; AL-GARNI, Z. (1997); **Forecasting Monthly Electric Energy Consumption in eastern Saudi Arabia using Univariate Time-Series Analysis**. Energy Vol. 22, n.11, p.1059-1069.

ABRAHAM, B.; LEDOLTER, J. (1983); **Statistical Methods for Forecasting**. New York: John Wiley & Sons.

AGRAWAL, R.; SRIKANT, R. (1994); Fast algorithms for mining association rules in large databases. **Proceedings of the International Conference on Very Large Databases**, Santiago, Chile.

AMO, S. (2010); Curso de Data Mining – Aula 2 – Mineracao de Regras de Associacao – O algoritmo APRIORI.

BATISTA, G.E.A. P. A. B. (2003); Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. Tese de Doutorado (Doutorado em Ciência da Computação e Matemática Computação). USP – São Carlos. Disponível em: <http://www.teses.usp.br/teses/disponiveis /55/55134/tde-06102003-160219/publico/TeseDoutorado.pdf> Acesso em: 19 ago. 2015.

BERRY, M. J. A.; LINOFF, G.. (1997); **Data Mining Techniques**: For Marketing, Sales, and Customer Support. New York: Wiley Computer Publishing,

BERSON, A.; SMITH, S. J. (1997); **Data Warehousing, Data Mining and OLAP.** McGraw-Hill, Estados Unidos.

BRUSSO, M.J. (2000); Access Miner: uma proposta para a extração de regras de associação aplicada à mineração do uso da Web. Dissertação de Mestrado (Mestrado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul. Rio Grande do Sul. Disponível em: <http://upf.tche.br/~brusso/pub/dissertacao.pdf>. Acesso em: 23 ago. 2015.

CORRÊA, K. S. (2009); Processo de mineração de dados no estudo de fenômenos solares e geomagnéticos. 28 f. Trabalho Acadêmico (mestre) – INEP, São José dos Campos.

COSTA, A. F.; TRAINA, A. J. M. (2012); Mineração de Imagens Médicas Utilizando Características de Forma..

DIAS, T.; BARBOSA, P.; DIAS, P.; MOITA, G.. (2015); **Identificação de Padrões em Registros de Doenças com Técnicas de Mineração de Dados**. CONTECSI, BrasiI. Disponível em: <http://www.contecsi.fea.usp.br/envio/index.php/contecsi/12CONTECSI/paper/view/3234 /2449>. Acesso em: 18 ago. 2015.

DINIZ, C. A. R.; LOUZADA NETO, F. (2000); **Data mining:** uma introdução. São Paulo: ABE, 123p.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. (1996); From Data Mining to Knowledge Discovery in Databases. **American Association for Artificial Intelligence**.

GOLDSCHIMIDT, R.; PASSOS, E. (2005); **Data mining:** um guia prático. Rio de Janeiro: Campus.

LAVRAC, N.; KERAVNOU, E.; ZUPAN, B. (2000); Intelligent data analysis in medicine. **Encyclopedia of computer science and technology**, 42(9), 113-157.

LIM, C.; McALEER, M. (2001); **Time Series Forecasts of International Travel Demand for Australia**. Tourism Management.

MAKRIDAKIS, S.; WHEELWRIGHT, S.; HYNDMAN, R.J. (1998); **Forecasting Methods and Applications**. 3. ed. New York: John Wiley & Sons.

MELO, M.D. (2010); **Introdução à Mineração de Dados usando o Weka**. V CONNEPI.

PASTA, A. (2011); Aplicação da técnica de Data Mining na base de dados do ambiente de gestão educacional: Um estudo de caso de uma instituição de ensino superior de Blumenau-SC. Dissertação de Mestrado (Mestrado em Computação Aplicada) – Universidade do Vale do Itajaí, São José. Disponível em: < http://www.uniedu.sed.sc.gov.br/wp-content/uploads/2013/10 /Arquelau-Pasta.pdf >. Acesso em: 21 ago. 2015.

PELLEGRINI, F.R.; FOGLIATTO, F. (2000); Estudo comparativo entre modelos de Winters e de Box-Jenkins para a previsão de demanda sazonal. **Revista Produto & Produção**. Vol. 4, número especial, p.72-85.

PRAHALAD, C. K.; KRISHNAN, M. S. (2008); A nova era da inovação: a inovação focada no relacionamento com o cliente. Rio de Janeiro: Elsevier.

PRASS, F. S.Uma visão geral sobre as fases do Knowledge Discovery in Databases (KDD). Disponível em: <http://fp2.com.br/blog/wp-content/uploads/2012/07 /KDD_Uma_visao_geral_do_processo.pdf. Acesso em 18 ago. 2015.

RAMOS, P.G.. (1999); Uma Investigação das Redes Neuro-Fuzzy aplicadas à Mineração de Dados. Monografia de Graduação (Graduação em Ciência da Computação) – Universidade Federal de Recife, Recife. Disponível em: <http://www.di.ufpe.br/~tg/1999-1/pgr.doc>. Acesso em: 21 ago. 2015.

UNIVERSITY OF WAIKATO. (2010); Weka 3 – **Machine Learning Software in Java**. Disponível no site da University of Waikato. URL: http://www.cs.waikato.ac.nz/ml/weka

VIANA, R. P. R. Data Mining: Auxiliando na tomada de decisões estratégicas nas empresas. 2013. 51 f. Monografia (Curso de Ciência da Computação) – Universidade Fundação Mineira de Educação e Cultura (FUMEC), Belo Horizonte, 2013.

WEKA. (1997); **Machine Learning**. McGraw-Hill. Disponível em: <http://www.cs.waikato.ac.nz/~ml/WEKA/index.html>. Acesso em: 27 ago. 2015.

---

1. Mestrando em Engenharia de Produção pela Universidade Tecnológica Federal do Parará (UTFPR). Email: jovanisouza5@gmail.com;
2. Professor e Coordenador do Programa de Pós-Graduação em Engenharia de Produção da Universidade Tecnológica Federal do Parará (UTFPR). Email: acfrancisco@utfpr.edu.br;
3. Professor Titular da Universidade Tecnológica Federal do Parará (UTFPR). Email: kovaleski@utfpr.edu.br;
4. Cursando especialização em Engenharia de Produção pela Uninter. Email: brunoaparecidooliveira@gmail.com;
5. Professor da Universidade Tecnológica Federal do Parará (UTFPR). Email: alamobatista@utfpr.edu.br.
