Um Algoritmo para Extração de Conhecimento em Saúde Baseado em Regras de Associação

André Magno C. de Araújo[1], Marcos Júnior Lopes[2], Wermeson Lopes Trindade[2]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
50.670-901– Recife – PE – Brasil

[2]Sistema de Informação – Inst. Tocantinense Pres. Antônio Carlos (ITPAC)
77.803-010– Araguaína – TO – Brasil

amca@cin.ufpe.br, marcosjr.m12@gmail.com, wermesonlt@gmail.com

Abstract. *This paper specifies an algorithm capable of extracting knowledge in health from clinical data repositories. Called ApHealth, and based on the concept of association rules for the work proposed here extends the Apriori algorithm by inserting a new input parameter named dimension. The definition of the new parameter will evaluate the association patterns found by means of a new perspective, for example, the frequency of the clinical tests were conducted on the observed size of patient diagnosis. Additionally, a prototype was developed and applied in a repository of clinical data from a hospital located in the northern region of the country.*

Resumo. *Este trabalho especifica um algoritmo capaz de extrair conhecimento em saúde a partir de repositórios de dados clínicos. Chamado de ApHealth, e baseado no conceito de regras por associação, o trabalho aqui proposto estende o algoritmo Apriori com a inserção de um novo parâmetro de entrada denominado dimensão. A definição do novo parâmetro permitirá avaliar os padrões associativos encontrados por meio de uma nova perspectiva, como por exemplo, a frequência com que os exames de análises clínicas foram realizados, observados sobre a dimensão do diagnóstico do paciente. Além disso, um protótipo de ApHealth foi desenvolvido e aplicado em um repositório de dados clínicos de um hospital situado na região norte do país.*

1. Introdução

Ao longo dos anos, a complexidade encontrada na descoberta de conhecimento em grandes repositórios de dados vem aumentando consideravelmente. Em suma, isso ocorre em virtude do grande poder de processamento dos Sistemas de Informação (SI) nos processos de negócio de uma empresa. O acirramento da competição e a necessidade de informações para a tomada de decisão em um curto espaço de tempo norteiam o cenário atual das organizações (Schuch et al, 2009). Cada vez mais abrangentes, os sistemas computacionais desempenham a importante tarefa de criar mecanismos para a manipulação (i.e. inserção, exclusão, atualização e consulta) dos dados. Entretanto, o grande volume de dados armazenados expõe a atual dificuldade encontrada pelas organizações, a falta de ferramentas que apoie a descoberta de conhecimento, padrões e comportamentos.

O Cenário em que as instituições hospitalares estão inseridas carece de informações oportunas e conhecimento personalizado que as auxilie nos processos decisórios (Copetti et al, 2008). A falta de ferramentas que forneçam indicadores gerenciais para a melhor obtenção de recursos junto a operadoras de saúde, secretarias

estaduais e órgãos federais de saúde, contextualiza as dificuldades encontradas na gestão hospitalar. A contratualização de novos serviços e a negociação financeira sobre os serviços prestados tem seu poder visivelmente aumentado quando se tem uma série de informações extraídas a partir dos dados organizacionais. Além disso, outro aspecto relevante a ser observado, consiste na dificuldade em se identificar padrões e comportamentos associados aos dados clínicos (e.g. cirurgias, exames, procedimentos, diagnósticos) para a realização de campanhas de saúde, prevenção de doenças e mapeamento de atendimentos.

Este trabalho especifica um algoritmo baseado em regras de associação que visa extrair conhecimento de dados clínicos a partir do Registro Eletrônico em Saúde (RES). O RES consiste em um repositório de dados clínicos e demográficos que abrange toda a vida do paciente e que visa suportar continuamente o tratamento, evolução e histórico dos dados. Chamado de ApHealth, o algoritmo proposto neste trabalho foi implementado em um protótipo de interface gráfica e aplicado no RES de pacientes do Sistema Único de Saúde (SUS) de uma organização hospitalar situada na região norte do país

Este artigo está organizado como segue. A seção 2 contextualiza os conceitos básicos utilizados ao longo deste artigo. A seção 3 apresenta a especificação do algoritmo *ApHealth* juntamente com o protótipo desenvolvido, destacando-se as principais contribuições alcançadas, enquanto que a Seção 4 descreve as considerações finais deste artigo.

2. Referencial Teórico

Esta seção descreve os principais conceitos utilizados para o desenvolvimento do trabalho aqui proposto.

2.1. Data Mining

De acordo com Elmasri e Navathe (2011), *Data Mining* ou Mineração de Dados pode ser definida como a descoberta de novas informações em termos de padrões ou regras com base em uma grande quantidade de dados. A mineração de dados ajuda na extração de novos padrões significativos que não podem ser necessariamente encontrados apenas ao consultar ou processar dados ou metadados em um repositório, podendo ser aplicada tanto em banco de dados históricos (i.e. data warehouse) como também em banco de dados operacionais com transações individuais.

O *Data Mining* é uma etapa do processo de KDD (*Knowledge Discovery in Databases* – Descoberta de conhecimento em bases de dados), nela ocorre a busca por conhecimentos novos e úteis a partir dos dados preparados nas etapas anteriores (Goldschmidt e Passos, 2005). A mineração de dados baseia-se em técnicas computacionais como aprendizagem de máquina, redes neurais e algoritmos genéticos. Os algoritmos de mineração percorrem e investigam relações de similaridade ou discordância entre os dados, para então encontrar padrões e modelos em grandes volumes de dados. Padrões são caracterizados como eventos temporais que ocorrem com frequência, enquanto que modelos correspondem à estrutura que descreve de forma resumida estes dados. Os padrões e modelos são a base para a avaliação de ocorrências, ou seja, irão refletir diretamente nos resultados, mas para que a mineração de dados, de fato seja eficiente, os padrões e modelos obtidos devem ser úteis, precisos, trazendo novidade e informações interessantes para qualquer organização. O resultado da

mineração de dados pode ser demonstrado em diversos formatos, como lista de dados, gráficos e tabelas.

2.2 Regras por Associação

A técnica de regras por associação procura itens que ocorram de forma simultânea e frequente em um conjunto de dados. O objetivo da mineração de dados com regras de associação é gerar as regras que atendam aos patamares mínimos de suporte e confiança estabelecidos (Silva e Ratke, 2011).

As regras de associação, como visto na Figura 1, tem a forma de: *Left-Hand Side* => *Right-Hand Side*, sendo estes denominados, LHS e RHS, e a união entre estes conjuntos de itens (i.e. LHS U RHS) formam o chamado *itemset*. Um *itemset* com k elementos denomina-se de *k-itemset*, para referir-se a quantidade de itens por conjunto.

Para que as regras de associação sejam geradas, os parâmetros de suporte e confiança devem ser fornecidos. Define-se o suporte como o percentual de ocorrências do *itemset* no conjunto de transações, tendo em vista a evidência de que LHS U RHS ocorram juntos. Dessa forma, calcula-se o suporte, dividindo o total de transações em que o itemset ocorre, pelo total geral de transações encontradas no banco de dados. O parâmetro confiança determina a validade das regras levando em consideração a probabilidade dos itens de RHS ocorrerem em relação aos itens de LHS. Dessa forma, calcula-se a confiança, dividindo o suporte de (LHS U RHS), pelo suporte de LHS.

Algoritmos baseados em regras de associação utilizam a propriedade de antimonotonicidade para reduzir o espaço de busca por soluções possíveis, com ela um *itemset* somente será frequente (i.e. suporte maior ou igual ao estabelecido) se todos os seus itens também forem frequentes (Schuch et al, 2010).

Figura 26. Exemplo de Regra por Associação.

2.3 Algoritmo Apriori

O Apriori é um dos mais conhecidos algoritmos na aplicação da tarefa de regras de associação. Serviu de base para outros algoritmos existentes como: DHP (Direct Hashing and Pruning), Partition, DIC (Dynamic Itemset Counting), Eclat, MaxEclat, Clique e MaxClique (RIBEIRO; VIEIRA; TRAINA, 2005).

O algoritmo pode ser dividido, com relação a sua execução, em três etapas. A primeira etapa realiza uma varredura na amostra de dados identificando os 1-*itemsets* frequentes (L1). Na segunda etapa, os *itemsets* são formados de acordo com o suporte definido, enquanto que na terceira etapa as regras são geradas com base no parâmetro da confiança.

Para exemplificar o funcionamento de Apriori, ilustra-se na Figura 2 um conjunto de transações (Pedidos de Exames) obtidos a partir do RES dos pacientes do SUS. Define-se inicialmente um valor para o suporte mínimo, supMin = 0,3 e confiança mínima, confMin = 0,3.

Conforme mostrado na Figura 2, o algoritmo realiza uma varredura no conjunto de dados para encontrar os *1-itemset* frequentes candidatos (chamado de C1) com seus respectivos valores de suporte. Então, a partir dos itens de C1, encontram-se os *1-itemset* frequentes L1, nesse caso, os itens que possuem o suporte maior ou igual ao suporte mínimo.

Na segunda etapa, há um laço de repetição que gera os demais conjuntos, isto é, L2, L3,...,Lk. Baseado no conjunto de itens frequentes anteriores, geram-se os *2-itemsets* frequentes candidatos (i.e. C2). Ao final da primeira iteração do *loop* têm-se os *2-itemsets* frequentes L2.

Obtem-se o conjunto L2 gerado na iteração anterior, e a partir dele geram-se os *3-itemsets* frequentes candidatos (i.e. C3), filtra-se os *itemsets* pelo suporte, ao final têm-se os *3-itemsets* frequentes L3.

Na terceira iteração do laço de repetição, nenhum conjunto de itens é gerado, isso se deve em razão da quantidade de elementos insuficientes em L3 para gerar os *itemsets* para C4, finalizando assim o *loop*. Dessa forma, termina-se a segunda etapa do algoritmo e os *itemsets* gerados conforme o suporte estabelecido.

Os *itemsets* gerados na etapa anterior (i.e. L2 e L3) do algoritmo são combinados, e para cada regra é calculado o valor da confiança. Na terceira etapa são selecionadas as regras com a confiança maior ou igual ao parâmetro informado, retornando ao analista os resultados com as associações que foram geradas, como visto na Figura 2.

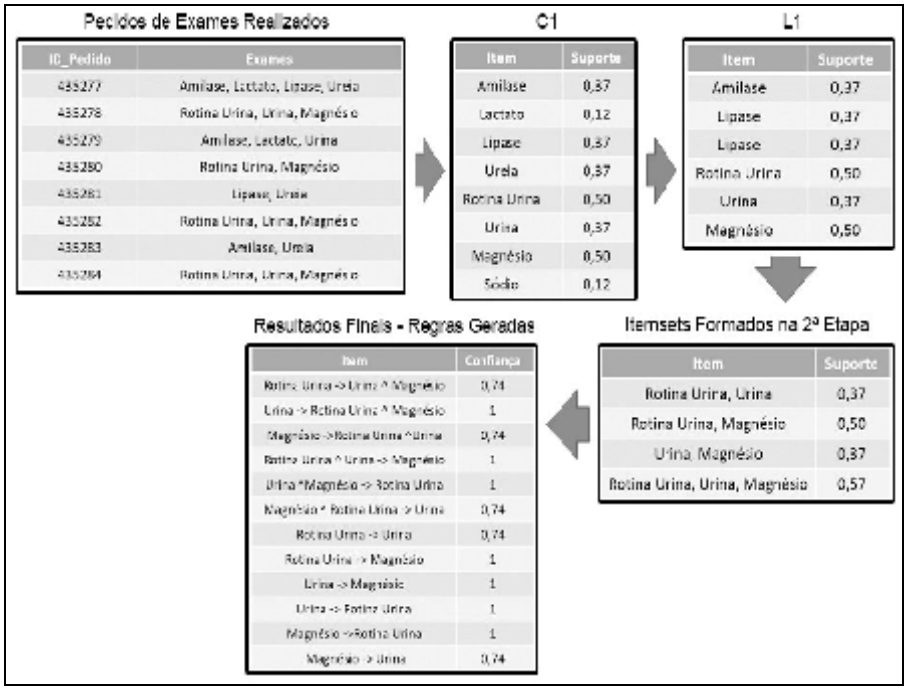

Figura 27. Dados gerados pelo Algoritmo Apriori.

3. O Algoritmo ApHealth

ApHealth baseia-se no algoritmo Apriori e no conceito de regras por associação. O objetivo do algoritmo é permitir que padrões e comportamentos sejam identificados a partir da leitura de um repositório de dados em saúde. Uma das características de ApHealth é a inserção de um parâmetro de entrada denominado dimensão. A definição do novo parâmetro permitirá avaliar os padrões associativos encontrados por meio de

uma nova perspectiva, como por exemplo, a frequência com que os exames de análises clínicas foram realizados, observados sobre a dimensão do diagnóstico do paciente. O pseudocódigo do algoritmo é ilustrado na Figura 3 e sua explicação é dada a seguir.

```
1  Início
2  Leia(sup_min);Leia(conf_min);Leia(dimensao);
3  Gere a partir de todas as transações os 1-itemsets frequentes (L1):
4  Lk = L1;
5  K=2;
6  termina – false;
7  repita
8        Obter a lista de itens (Li), distintos, contidos em LK;
9        Gerar as combinações para formar cada itemset, sem repetição de itens, baseado em
         Li, formando cada itemset i em conjuntos de K itens;
10       Calcular o suporte de cada itemset;
11       Remover as regras com suporte < sup_min;
12       Limpar a lista Lk;
13       Armazenar os itemsets restantes em Lk;
14       Armazenar os itemsets restantes na lista de todos itemsets já gerados, Lt.
15       k=k+1.
16       Se Lk estiver vazio então termina = true;
17 até que termina;
18 contLt  = quantidade total de itens presentes em Lt;
19 Para i=1 até contLt faça
20       Gerar as combinações lógicas (regras) com os itens distintos presentes no itemset i;
21       Calcule o suporte e a confiança para cada regra;
22 fim para;
23 Remover as regras que estejam com a confiança < conf_min.
24 Retornar a lista de regras encontradas através de técnicas de associação, geradas conforme
o suporte, confiança e dimensão estabelecidos.
25 Fim;
```

Figura 28. Pseudocódigo do Algoritmo ApHealth

Informado os parâmetros obrigatórios (sup_min, conf_min e dimensao), o algoritmo inicia a varredura nas transações buscando identificar os 1-itemsets frequentes (*L1*) e obedecendo aos seguintes critérios: itens com suporte maior ou igual ao parâmetro suporte mínimo, calculado sobre a dimensão informada. Os valores encontrados para L1 são atribuídos a uma lista auxiliar denominada L*k*, conforme mostra a linha 4 da Figura 3. L*k* tem a função de armazenar os *itemsets* frequentes de cada iteração. A variável *k* é utilizada para controlar a quantidade de itens por *itemset*, sendo seu valor iniciado com o valor 2 em função da primeira execução do loop que busca os demais i*temsets* frequentes, começando com 2-*itemset*, como mostrado na linha 5 da Figura 3.

Identificada a primeira lista de itens frequentes (i.e. L1), o algoritmo gera os demais conjuntos (i.e. L2, L3,..,Ln), obedecendo sempre aos critérios definidos nos parâmetros de entrada. Para isso, um loop inicia a tarefa de encontrar os itens distintos presentes em Lk (i.e Li) e assim, as combinações de k-itemsets são formadas. Com os itens gerados, faz-se o cálculo do suporte e eliminam-se as regras com o suporte abaixo do mínimo informado. A partir desse instante, têm-se os itemsets de acordo com os parâmetros informados. Os itemsets atuais presentes em Lk (i.e. formados na iteração anterior) são removidos para dar lugar aos itens frequentes atuais da iteração. Os itemsets presentes em Lk são atribuídos a uma nova lista auxiliar denominada Lt. Lt é responsável por armazenar todos os itens frequentes gerados por Lk. Para finalizar a iteração, o valor de k é incrementado e verificado se há itens gerados na passagem atual, caso não haja, o loop é finalizado. O trecho de código que representa a geração de todos os conjuntos é ilustrado na Figura 3, das linhas 7 a17.

O trecho de código exibido entre as linhas 19 a 22 da Figura 3 percorre todos os *itemsets* presentes em Lt e, para cada elemento encontrado, as regras são formadas, e o cálculo do suporte e confiança é realizado para cada uma delas. Encerrada essa rotina, as regras com a confiança abaixo do valor estabelecido são removidas, e as restantes são retornadas ao analista.

Na versão original do algoritmo Apriori, as combinações lógicas (i.e. regras) são geradas na forma de implicação, como por exemplo, A→B, B→A, A^B→C e C→ A^B. Nota-se que na definição do algoritmo não há nenhum mecanismo que controle a posição do item em relação á implicação na regra, sendo assim, as regras tornam-se distintas em relação ao cálculo da confiança.

ApHealth possui três características que o diferencia do algoritmo Apriori. Primeiro, o cálculo do suporte para gerar os itens frequentes (i.e. L1, L2, Ln) adota o parâmetro dimensão como filtro para a contagem da quantidade de ocorrências do itemset no conjunto de transações. Segundo, a dimensão faz parte da composição da regra, ou seja, pode-se analisar a implicação do item em relação à dimensão, por fim, o cálculo de confiança no que se refere ao suporte de (LHS U RHS) considera também a nova dimensão informada.

3.1. Resultados e Discussão

Para validar ApHealth, utiliza-se neste trabalho o RES dos pacientes atendidos pelo SUS, na modalidade ambulatorial. Nessa modalidade de atendimento são ofertadas consultas médicas e exames especializados. Foram coletados para a análise deste trabalho, os atendimentos dos pacientes (transações), exames solicitados por pedido (itens) e diagnósticos (dimensões), totalizando assim uma amostra de 1000 Registros.

Neste cenário, as instituições carecem de informações oportunas que as ajudem no processo de tomada de decisão. Prova disso, é a dificuldade encontrada em se mapear a frequência com que determinados exames são realizados dado o diagnóstico do paciente. Nesse sentido, A Figura 4 ilustra a interface do aplicativo desenvolvido para dar suporte ao algoritmo ApHealth. No exemplo mostrado na Figura 4, a amostra de dados refere-se aos pedidos de exames realizados nos atendimentos do SUS. Com base nesta amostra, deseja-se investigar qual a probabilidade de ocorrência dos exames realizados, em relação ao diagnóstico (i.e. dimensão) do paciente. Esta informação é de suma importância, por exemplo, para se determinar quais exames tem demandado mais insumos para sua realização e também, quais exames são sempre realizados para um determinado diagnóstico.

Depois de inseridos os valores para o suporte mínimo, confiança mínima e dimensão, que nesse caso representa o diagnóstico sobre o qual o analista deseja investigar a frequência dos exames, é possível verificar os resultados gerados por ApHealth. Para a dimensão: Infecções Agudas Não Especificadas das Vias Aéreas, constata-se que, em 100% dos casos será realizado o exame de Proteína C Ultra Sensível, isso em virtude do parâmetro da confiança ter retornado o valor igual a 1. Outro dado interessante de ser observado é que, somente em 50% dos casos foram solicitados juntos os exames de Cálcio Iônico e Hemograma Completo, enquanto que Cálcio Iônico mais Proteína C Ultra Sensível aparecem em 100% dos pedidos.

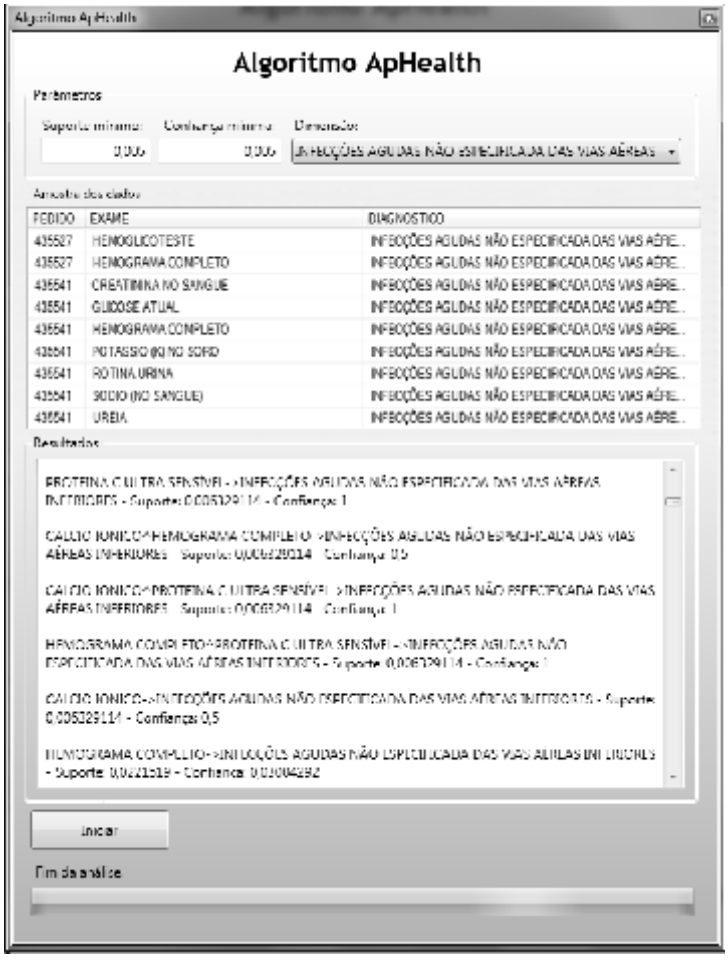

Figura 29. Interface Gráfica de ApHealth.

4. Conclusão

A extração de conhecimento em saúde representa um importante passo para que as organizações e unidades hospitalares tomem decisões alinhadas a sua estratégia organizacional. Nesse sentido, a aplicação de técnicas de mineração de dados permite que padrões e comportamentos associados ao RES sejam identificados e utilizados na identificação prévia de surtos de doenças, no mapeamento de atendimento por perfil de usuário, entre outras vantagens competitivas.

Neste trabalho apresentamos um algoritmo denominado ApHealth que a partir da leitura de um repositório de dados clínicos identifica padrões com base no conceito de regras por associação. Como contribuição, ApHealth define um novo parâmetro de entrada no qual é possível avaliar os padrões associativos encontrados por meio de nova perspectiva definida pelo analista. A aplicação do algoritmo em um

cenário real permitiu identificar, por exemplo, a probabilidade de ocorrência dos exames de análises clínicas, observados a partir do diagnóstico do paciente. Outras informações em saúde como dados de partos, cirurgias e consultas, também podem ser aplicadas em ApHealth.

Referências

Elmasri, R. e Navathe, B. (2011) Sistemas de Banco de Dados, Addison Wesley, 6ª edição.

Goldschmidt, R. e Passos, E. (2005) Data Mining: um guia prático, Elsevier.

Ribeiro, M. X., Vieira, M.T.P. e Traina, A.J.M. (2005) "Mineração de Regras de Associação Usando Agrupamentos", http://www.lbd.dcc.ufmg.br/colecoes/wamd/2005/WAMD_2.pdf, Março.

Schuch, et al. (2010) "Mineração de dados em uma subestação de energia elétrica", http://www.sbmac.org.br/dincon/trabalhos/PDF/energy/68015.pdf, Março.

Schuch, et al. (2009) "Mineração de Regras de Associação Aplicada a Dados do Tratamento Quimioterápico", ftp://200.143.198.48/nsi/CONFENIS2010/ 2.track%20Regular/confenis2010_submission_40.pdf, Julho.

Silva, M. e Ratke, C. (2011) "Gestão da informação em biblioteca universitária: uma proposta utilizando regras de associação na disseminação das informações de novas aquisições bibliográficas", http://www.inf.furb.br/seminco/2011/pdfs/ seminco_artigo3.pdf, Julho.