AMAGODIS: Algoritmos de Mineração para Apoio à Gerência de Ocorrências de Dengue a partir de Informações presentes na base dados do SINAN

Marcelo Silva Santos[1], **José Craveiro da Costa Neto**[1]

[1]Departamento de Ciências Exatas e Tecnológicas – Universidade Estadual de Santa Cruz (UESC)
Caixa Postal 45 662 900 – Ilhéus – BA – Brasil

marcelossjj@gmail.com,jccneto@uesc.br

Abstract. *Dengue is now one of the most serious public health problems in Brazil, so there is a need for tools to help health managers in their control. This paper describes the development of a tool for application of algorithms for classification and discovery of association rules in the database System Diseases Information and Notification (SINAN) obtaining decision trees and association rules regarding the occurrence of dengue called of AMAGODIS. The AMAGODIS was developed considering the end user. Anyway, this has led to significant results with the tool that is described throughout the article.*

Resumo. *A dengue hoje é um dos mais sérios problemas de saúde pública do Brasil, por isso, existe a necessidade de ferramentas que auxiliem gestores de saúde no seu controle. Este trabalho descreve o desenvolvimento de uma ferramenta para aplicação de algoritmos de classificação e descoberta de regras de associação na base de dados do Sistema de Informações de Agravos e Notificação (SINAN) obtendo árvores de decisão e regras de associação a respeito das ocorrências de dengue chamado de AMAGODIS. O AMAGODIS foi desenvolvido pensando no usuário final. Enfim, geraram-se resultados significativos com a ferramenta que é descrito ao decorrer do artigo.*

1. Introdução

A dengue vem sendo considerada como uma das doenças com maior incidência no Brasil, independente da classe social. O aumento dos casos de dengue que vêm ocorrendo no Brasil traz preocupação para a sociedade e para as autoridades sanitárias em razão das dificuldades para o seu controle.

As características clínicas e epidemiológicas da dengue no Brasil têm despertado o interesse de pesquisadores e organismos nacionais e internacionais de saúde pública, tendo em vista a importância da identificação dos fatores que determinam as distintas formas de expressão individual e coletiva dessas infecções para o aperfeiçoamento do seu tratamento e controle, pois, em termos de número de casos, representa a segunda mais importante doença transmitida por vetor no mundo.

O número de casos de dengue no Brasil quase triplicou entre 2009 e 2010 segundo dados do Ministério da Saúde. A partir de dados do Ministério da Saúde, mais de um milhão de pessoas podem ter contraído a doença, contra 323.876 do ano anterior. O número

de óbitos saltou de 298 para 592 no mesmo período. Para cada grupo de cem mil habitantes, o número de casos saltou de 170,8 para 489 pessoas.

Por esta razão, manifestou-se o interesse na aplicação de algoritmos de mineração de dados na base de dados do SINAN, referente às ocorrências de surtos de dengue, no intuito de extrair padrões e associações entre os dados que proporcionem informações valiosas, que servirão de auxílio aos gestores de saúde no manejo clínico dos pacientes com suspeita de dengue.

Esta aplicação será feita através do software denominado AMAGODIS (Algoritmos de Mineração para Apoio à Gerência de Ocorrências de Dengue a partir de Informações presentes na base dados do SINAN), onde o usuário vai ter total permissão para escolher o algoritmo e os atributos relativos ao paciente que lhe achar mais conveniente em sua busca.

O objetivo do trabalho é construir uma ferramenta para auxiliar os gestores de saúde onde disponibiliza a aplicação do algoritmo de mineração de dados com as características dos pacientes selecionadas pelo usuário na base de dados do SINAN para se obter regras de associação ou regras de classificação a respeito da dengue que podem ser valiosas para os gestores de saúde.

Com o processo de mineração de dados obteve-se informações valiosas a respeito dos surtos de dengue que aconteceram no município de Itabuna do estado da Bahia. Com esse tipo de informação os gestores de saúde podem utilizar como um auxílio no diagnóstico de pacientes.

2. Trabalhos Relacionados

Existem trabalhos que tratam do tema abordado. [Castro 2005] desenvolveu um trabalho que tinha como objetivo, predizer por meio de um modelo baseado em redes neurais artificiais, a ocorrência de surtos urbanos causados pelo mosquito da dengue, possibilitando aos gestores de saúde um suporte à decisão para o planejamento de combate a epidemias. Entretanto, este trabalho não procura desenvolver ferramentas que possam ser utilizadas por gestores de saúde.

O trabalho deste artigo segue uma área semelhante, mas um propósito diferente. O desenvolvimento de uma ferramenta de aplicação de algoritmos de mineração de dados em uma base de dados do SINAN, com o intuito de auxiliar os gestores de saúde no diagnóstico e comportamento da dengue.

3. Processo de Notificação das Ocorrências de Dengue

Segundo [Saúde 2003], a dengue é uma doença febril aguda, de etiologia viral e de evolução benigna na forma clássica, e grave quando se apresenta na forma hemorrágica. A transmissão do vírus da dengue ocorre pela picada do mosquito *Aedes aegypti* já infectado.

No diagnóstico da dengue, inicialmente, é feito um diagnóstico clínico para descartar possíveis doenças semelhantes como a febre amarela. Após esta etapa, são realizados alguns exames específicos se necessário, como a contagem de plaquetas que faz uma contagem quantitativa de plaquetas e a partir do número indicado consegue-se o resultado esperado.

As manifestações podem trazer complicações aos infectados, sendo elas: hemorrragia cutânea (petequias, púrpura, equimose), gengivorragia (sangramento gengival),

sangramento nasal (epistaxe), sangramento gastrointestinal (vômitos de sangue), metrorragia (sangramento da pele) e hematúria (sangue na urina).

Para um maior controle, o Ministério da Saúde criou um sistema chamado SINAN para fazer a notificação e investigação de casos de doenças e agravos que constam da lista nacional de doenças de notificação compulsória [Sinan 2003]. Entre estas doenças está a dengue, motivo deste estudo. Sua utilização efetiva permite a realização do diagnóstico dinâmico da ocorrência de um evento na população, fornecendo subsídios para explicar causas dos agravos de notificação compulsória.

4. Descoberta do Conhecimento em Bases de Dados

A mineração de dados é a etapa mais importante da metodologia para aquisição do conhecimento, onde possui várias etapas, por meio da qual se buscam padrões, associações, anomalias e estruturas significativas entre os dados obtendo informações valiosas em uma base de dados [Han e Kamber 2006]. Os resultados desta busca, em geral, manifestam-se por meio de regras de associação, agrupamentos e estruturas de classificação.

Segundo [Gonçalves 2005], o modelo típico para mineração de regras de associação em bases de dados consiste em encontrar todas as regras que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo (SupMin), que são chamados de conjuntos frequentes, e uma confiança mínima (ConfMin), cada conjunto frequente encontrado no SupMin.

O número de regras gerado costuma ser volumoso. Identificar as regras realmente interessantes é uma tarefa difícil utilizando o modelo suporte/confiança. Então, torna-se interessante a aplicação de medidas de correlação. Uma das medidas de correlação que solucionam esse problema é o *lift* que funciona da seguinte forma: a partir de uma regra de associação A implicando B, esta medida indica o quanto mais freqüente torna-se B quando existe a ocorrência de A [Han e Kamber 2006].

A classificação é o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados. O modelo construído baseia-se na análise prévia de um conjunto de dados de treinamento [Amo 2004].

O modelo de classificação mais usual tem sido a árvore de decisão. Para construí-la, usa-se frequentemente o algoritmo C4.5, trabalhando com o conceito de entropia. Sua característica principal é o sistema de podas, onde o algoritmo poda alguns ramos que não trazem um ganho de informação tão importante.

A medida de avaliação do algoritmo de classificação C4.5 é utilizada para validar o grau de acertos de uma árvore de decisão e identificar o quão perto as informações extraídas da árvore correspondem verdadeiramente aos dados reais. O *10 fold Cross-Validation* é uma das medidas de avaliação mais utilizadas em árvores de decisão, funcionando da seguinte forma: o conjunto de dados é aleatoriamente dividido em 10 subconjuntos (*folds*) mutuamente exclusivos de tamanho aproximadamente igual. O indutor treina com 10-1 *folds* e testa com o *fold* remanescente [Queiroga 2005].

Ocorre com bastante freqüência de os dados do mundo real estarem representados de forma incompleta, fora do padrão ou inconsistentes. As rotinas de limpeza de dados empreendem esforços no sentido de preencher os valores ausentes (*missing values*) e valores fora do padrão (*outliers*) [Ham e Kamber 2006]. Existem vários processos de limpeza destes dados; neste trabalho, trabalha-se com a substituição pela média, onde para

cada tipo de informação de um atributo foi realizado uma média de vezes em que ele aparece em relação com toda a massa de dados, com isso designou-se qual informação seria povoada, ao se encontrar informações com valores ausentes.

Uma das bibliotecas implementam a maioria dos algoritmos de mineração de dados é o Weka, oferecendo medidas de avaliação de classificadores e de regras de associação [Waikato 2008]. Além disto, o Weka oferece uma extensa coleção de técnicas de pré-processamento de dados e modelagem de dados, permitindo a criação das próprias fontes de dados, bem como a conexão direta com alguns bancos de dados e ainda disponibilizando uma API para o Java.

5. Projetando o Software

Com todo o conhecimento obtido em estudos relativo ao projeto, iniciou-se a projeção do software AMAGODIS. Como esta ferramenta possivelmente não será utilizada por especialista da computação foi necessário projetá-la pensando sempre no usuário final e suas limitações na possível utilização do software e em relação ao conhecimento provavelmente inexistente sobre mineração de dados.

Primeiramente foram analisadas quais informações podem contribuir aos gestores de saúde no controle e diagnóstico da dengue. Então foi identificado quais as informações que mais se adaptavam a mineração de dados para se obter um conhecimento rico.

Umas dessas informações são: classificar se um paciente tem dengue clássica ou hemorrágica a partir de algumas características pessoais; identificação da realização de um exame quando um paciente foi diagnosticado com algum tipo de dengue, pois existe um grande número de casos de dengue e a maior parte dos diagnósticos é realizada sem fazer nenhum tipo de exame, podendo acarretar problemas quando o paciente tem um tipo de dengue com complicações e é diagnosticado com dengue clássica; detectar manifestações hemorrágicas é de vital importância em pacientes com suspeita de dengue hemorrágica e algumas delas são identificadas por meio de exames. Com o grande número de casos em epidemias de dengue, a realização desses exames acaba não sendo realizada por causa do número de recursos limitados da saúde pública. Entre outras como classificar a zona do bairro do paciente.

Para aplicação do algoritmo de classificação a ferramenta foi projetada para que qualquer atributo possa ser classificado, procurando abranger vários conhecimentos para necessidades distintas dos gestores de saúde.

No processo de mineração a escolha dos atributos é de fundamental importância para se obter informações relevantes da base de dados. E por causa dessa importância foi necessária uma busca minuciosa dos atributos que obtinham informações mais ricas para o gestor de saúde e as que mais se adaptavam a este processo.

Na aplicação do algoritmo de classificação os atributos escolhidos com possível ganho de conhecimento foram: idade, sexo, raça, escolaridade, classificaçãoBairro, fezExame e diagnóstico. O atributo idade foi classificado em criança de 1 a 17 anos, adulto de 18 a 49 e idoso de 50 anos em diante. Cada atributo contém suas informações: idade (criança, adulto e idoso), sexo (masculino, feminino), raça (negra, branca), escolaridade (fundamental_1, fundamental_2 e médio), classificaçãoBairro (zona_pobre, zona_média e zona_rica), fezExame (sim, não) e diagnóstico (dengue clássica e dengue hemorrágica).

Observando-se as manifestações hemorrágicas, pode-se identificar a implicação ou não de uma em outra. Para se obter esse objetivo utilizando algoritmo de regras de

associação foram selecionados os seguintes atributos: epistaxe, gengivorragia, metrorragia, petéquias, hematúria, sangramento.

Finalizado o processo de escolha de atributos, procurou-se escolher qual linguagem de programação seria utilizada no desenvolvimento da ferramenta. Pelo fato da biblioteca weka ser escrito em Java e conter API (Interface de Programação de Aplicativos) para esta linguagem, foi escolhido o Java para o desenvolvimento da ferramenta.

5.1 Funcionamento da Ferramenta

A ferramenta opera da seguinte forma: faz uma chamada ao weka obrigando-o a fazer conexão com o SGBD PostgreSQL; através de uma consulta SQL informa todos os pacientes e atributos necessários para determinado algoritmo e atribuem as informações em uma variável; essas informações são levadas ao weka aplicando o algoritmo escolhido; gerado o conhecimento valioso com a aplicação do algoritmo nas informações, ele é tratado para uma visualização que o usuário final compreenda.

5.2 A interface do Ambiente

O AMAGODIS foi desenvolvido em Java, todo desenvolvimento da interface foi projetada para o usuário final pelo fato da mineração de dados ser um pouco confusa para pessoas que não são da área da ciência da computação. Para se obter esse resultado, foi necessário que alguns usuários finais utilizassem a ferramenta nas primeiras versões e depois dessem um retorno sobre as dificuldades e facilidades observadas na interface da ferramenta. A partir dessas informações retornadas dos usuários a interface foi sendo modificada até chegar ao estágio que é apresentada nesse artigo.

Logo na tela inicial da ferramenta, é possível a visualização de todas as opções que podem ser utilizadas, pois a interface tem apenas uma tela, isso foi feito para que o usuário não se perca entre várias janelas. A Figura 1 ilustra a tela do AMAGODIS.

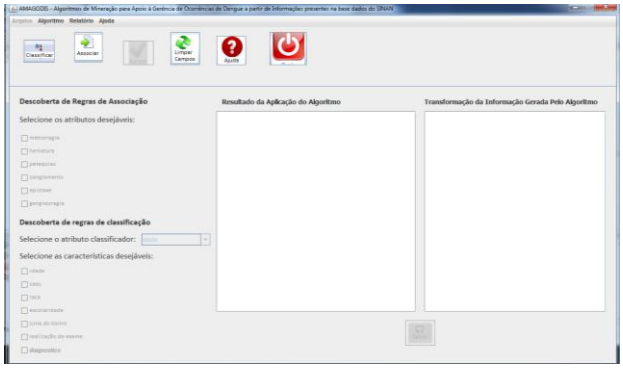

Figura 1. Apresentação da ferramenta AMAGODIS

Como é ilustrado na Figura 1, ao abrir a ferramenta as opções de seleção dos atributos para descoberta de regras de associação ou classificação ficam desabilitadas, ao selecionar a opção de classificar ou associar na barra de ícones os atributos são habilitados para escolha dependendo da seleção realizada. Outros botões como salvar e aplicar tem funcionalidades semelhantes, são desabilitados no início e apenas são habilitados a partir de uma determinada ação do usuário, ação essa simples de identificar.

Para fazer uma descoberta de regras de associação ou classificação as ações são semelhantes, apenas uma diferença marcante é que nas descobertas de regras de associação

é necessário selecionar o suporte, confiança e o número de regras (janela iniciada logo ao clicar no botão associar) que se deseja obter. Os passos de ações necessárias para fazer uma descoberta são simples, apenas deve-se clicar no botão referente ao tipo de descoberta, selecionar os atributos que irão se habilitar e clicar em aplicar. Serão ilustradas na primeira caixa de texto, as informações geradas pelos algoritmos, entretanto essas informações não serão compreendidas por um usuário comum, apenas por quem tem um conhecimento sobre mineração de dados e na segunda caixa de texto, as informações são tratadas para a compreensão do usuário comum.

Foi verificado indispensável o tratamento da informação gerada pelo algoritmo, quando os usuários que utilizarão a primeira versão do software não conseguiram compreender facilmente as informações obtidas na aplicação dos algoritmos. Então foi desenvolvido o tratamento dessas informações para se obter maior entendimento sobre o conhecimento que está sendo obtido. No Quadro 1 é ilustrado um exemplo da transformação da informação.

Quadro 1. Tradução da informação recebida do weka para o usuário final

```
Regras geradas pelo Weka
1 - hematúria=não sangramento=sim ==> metrorragia=não petéquias=não
2 - metrorragia=não sangramento=sim ==> petéquias=não

Regras tratadas para o usuário final
1 – Se não tem hematúria e tem sangramento, não tem metrorragia nem
petequias
2 – Se não tem metrorragia e tem sangramento, não tem petequias
```

As informações obtidas na descoberta de conhecimento podem ser exportadas para um arquivo no formato ".txt" que sempre vai ficar no diretório: "c:/RelatórioAmagodis/".

A mineração de dados é um assunto não abordado para pessoas fora dos cursos de ciência da computação, é indispensável um FAQ (Frequently Asked Questions). Para uma melhor compreensão, foi desenvolvido um FAQ com as perguntas que mais foram frequentes pelas pessoas que utilizaram o software para teste. Com esse FAQ procura-se ajudar o usuário a melhorar significativamente o entendimento sobre a ferramenta. Segue o exemplo de algumas perguntas: O que é mineração de dados? O que são algoritmos de classificação e de regras de associação? Como classificar um paciente com dengue clássica ou hemorrágica? O que é suporte e confiança? O que é um atributo classificador?

5.3 Resultados

Com esta ferramenta a descoberta de conhecimento na base de dados do SINAN é muito variada, pois, a ferramenta não limita o usuário a apenas um conhecimento, ela deixa que as opções sejam preenchidas pelo usuário para que ele obtenha um resultado que deseja.

A partir da seleção de um algoritmo e seus respectivos atributos a aplicação desse algoritmo irá gerar informações a respeito do que foram solicitadas, essas informações são exibidas nos Resultados da Aplicação e transformadas em uma linguagem mais acessível ao usuário, como já foi descrito neste artigo.

A Figura 2 ilustra o resultado obtido da aplicação do algoritmo de classificação para classificar se um paciente tem dengue clássica ou dengue hemorrágica. Para se obter esse resultado foram selecionados os seguintes atributos: idade, sexo, realização de exame, zona_bairro e diagnóstico como atributo classificador.

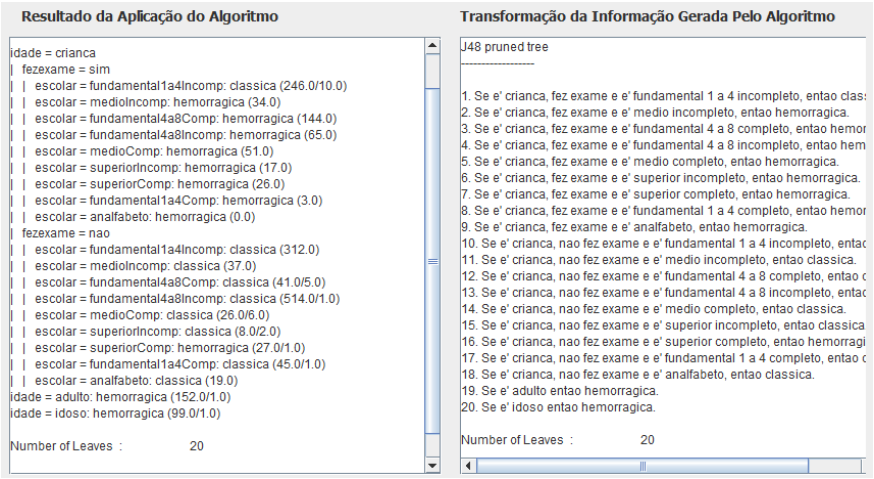

Figura 2. Resultado da Aplicação do Algoritmo de Classificação

Na Figura 3, pode-se analisar o resultado obtido da aplicação do algoritmo de associação quando o suporte é 0.65, confiança 0.6 e selecionado os seguintes atributos: epistaxe, petequias, sangramento e gengivorragia.

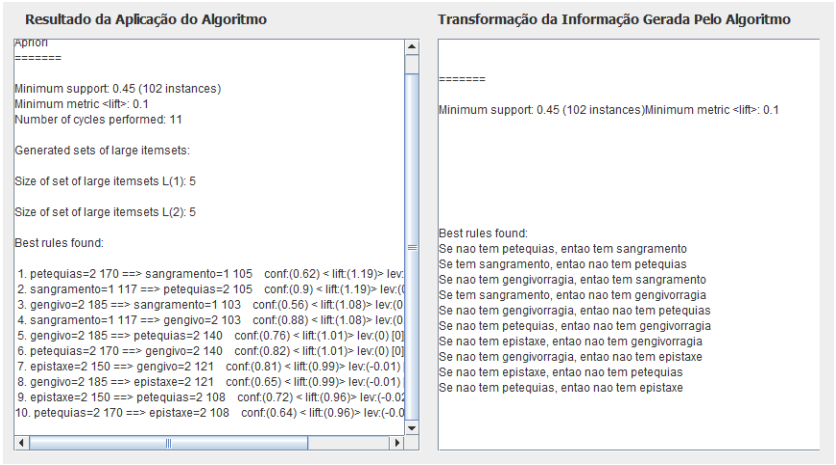

Figura 3. Resultado da Aplicação do Algoritmo de Associação

As regras de associações obtidas têm o intuito de informar a partir de associações das ocorrências de dengue hemorrágica se uma manifestação implica em outra. Obteve-se um número grande de regras de associação que foram validadas pela medida de correlação *lift* A Figura 3 ilustra as regras obtidas neste processo.

Como se pode analisar, a ferramenta deixa o usuário livre para fazer as combinações que deseja, e a partir disso analisar os resultados obtidos e identificar qual dos resultados traz mais contribuição para a sua necessidade.

6. Conclusões e Trabalhos Futuros

A busca de soluções para a prevenção da dengue é constante. Autoridades sanitárias trabalham com programas permanentes para o controle, desenvolvendo campanhas de mobilização e conscientização da população. É importante o fortalecimento da vigilância epidemiológica, a melhoria na qualidade do trabalho e o desenvolvimento de instrumentos eficazes no combate ao mosquito.

A medicina é um campo altamente promissor para utilização de técnicas de mineração de dados que procuram padrões, regras e anomalias entre os dados. Com o apoio da mineração de dados na área de saúde procurou-se aplicar os seus algoritmos para se obter um conhecimento sobre as ocorrências de dengue registradas na base de dados do SINAN para o auxilio do diagnóstico e outras informações sobre a dengue.

A ferramenta desenvolvida busca aplicar alguns dos algoritmos de mineração de dados gerando regras de associação ou de classificação relativa aos pacientes que teve dengue, permitindo ao usuário final a possibilidade de selecionar qual tarefa utilizar e quais atributos do banco de dados devem ser aplicados na tarefa selecionada, extraindo informações que passarão no processo de validação do conhecimento e retornaram informações úteis para os gestores de saúde.

As informações geradas pela ferramenta AMAGODIS têm o objetivo de auxiliar o gestor de saúde no diagnóstico de pacientes com suspeita de dengue. Entretanto é importante informar que essas informações são técnicas estatísticas que não substituem a presença do gestor de saúde.

Referências

Amo, S. (2004). "Técnicas de Mineração de Dados", In: XXIV Congresso da Sociedade Brasileira de Computação. Jornada de Atualização em Informática, Salvador.

Castro, G. G. (2005). "Suporte à Decisão para Vigilância Epidemiológica Baseado em Modelo Preditivo de Surtos de Dengue Utilizando Redes Neurais Artificiais". Dissertação (Mestrado), Brasília.

Figueredo, G. P. (2008). "Mecanismo Imuno-supressor para Seleção de Dados de Treinamento em Problemas de Classificação". Tese (Doutorado), Rio de Janeiro.

Gonçalves, E. C. (2005). "Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas". In: InfoComp, Jornal da Ciência da Computação, Lavras. p.26 – 35.

Han, J. e Kamber, M. (2006), Data Mining: Concepts and Techniques , 2th edition.

Queiroga, R. M. (2005). "Uso de Técnicas de Data Mining para Detecção de Fraudes em Energia Elétrica". Dissertação (Mestrado), Espírito Santo.

Saúde, M. (2003). Dengue – Aspectos Epidemiológicos, Diagnóstico e Tratamento, 1th edition.

SINAN (2003). "Sistema de Informação de Agravos de Notificação (SINAN)", http://portal.saude.gov.br/portal/saude/visualizar_texto.cfm?idtxt=21383, Janeiro.

Waikato (2008). "Weka 3: Data Mining Software in Java", http://www.cs.waikato.ac.nz/ml/Weka/, Janeiro.