APLICANDO ALGORITMOS DE MINERAÇÃO NÃO SUPERVISIONADOS PARA INFERIR O PERFIL DE PACIENTES INTERNADOS

RESUMO

A melhoria da qualidade das informações e da tomada de decisão incentiva a exploração de novas perspectivas sobre as informações hospitalares. Neste contexto não é suficiente apenas ter um BI, é necessário conhecer a realidade e desta forma se preparar para o futuro. A união das ferramentas de BI com as de mineração de dados possibilitam esta melhoria. A ferramenta de data mining Weka foi utilizada para determinar o perfil dos pacientes internados em 2013 no Hospital Universitário da UFGD/EBSERH para desta forma identificar os maiores grupos de pacientes do hospital, conseguindo assim melhorar as práticas administrativas e assistenciais no atendimento a estes pacientes, tendo estes grupos como prioridade.

PALAVRAS-CHAVES: Mineração de Dados; Algoritmo; Segmentação.

APPLYING MINING ALGORITHM NOT SUPERVISED TO INFER THE PROFILE OF HOSPITALIZED PATIENTS

ABSTRACT

Improving the quality of information and decision-making stimulates exploitation of new perspectives on hospital information. In this context it is not enough just to have an BI, you must know the reality and thus prepare for the future. The union of the BI tools with of the data mining allow this improvement. The data mining tool Weka was used to determine the profile of patients hospitalized in 2013 at the University Hospital of UFGD / Ebserh to thereby identify the major groups of hospital patients, thereby achieving improve administrative and care practices in the care of these patients, having these groups as priority.

KEYWORDS: Data Mining; Algorithm; Clustering.

*Revista Brasileira de
Administração Científica,
Aquidabã, v.5, n.2, Out 2014.*

ISSN 2179-684X

SECTION: Articles

TOPIC: *Sistemas e Tecnologia da
Informação*

*Anais do Simpósio Brasileiro de
Tecnologia da Informação (SBTI 2014)*

DOI: 10.6008/SPC2179-684X.2014.002.0017

Luiz Fernando Stopa Arcenio

Universidade Federal da Grande Dourados, Brasil
<http://lattes.cnpq.br/7686411158143328>
luizarcenio@gmail.com

Received: 31/08/2014

Approved: 15/10/2014

Reviewed anonymously in the process of blind peer.

Referencing this:

ARCENIO, L. F. S.. Aplicando algoritmos de mineração não supervisionados para inferir o perfil de pacientes internados. *Revista Brasileira de Administração Científica*, Aquidabã, v.5, n.2, p.226-234, 2014. DOI: <http://dx.doi.org/10.6008/SPC2179-684X.2014.002.0017>

INTRODUÇÃO

O Hospital Universitário da UFGD foi incorporado pela Universidade Federal da Grande Dourados em 2009, em um processo de reestruturação tecnológica, financeira e administrativa realizado em 2013 o hospital foi incorporado pela EBSEH - Empresa Brasileira de Serviços Hospitalares. Neste intervalo de 2009 a 2013, houve a implantação do Aplicativo de Gestão dos Hospitais Universitários – AGHU. Este aplicativo é composto por diversos módulos que atendem às áreas administrativas e assistenciais, contando para isso de relatórios sintéticos e analíticos, e de um módulo de *Business Intelligence* - BI. Apesar de todas as funcionalidades apresentadas pelo sistema, há necessidades ainda não contempladas, como por exemplo, conhecer o perfil do paciente internado no hospital, ou seja, conhecer realmente qual o perfil da maioria dos pacientes atendidos.

Neste artigo são apresentados os algoritmos de mineração de dados baseados em agrupamentos disponíveis na ferramenta *Weka* e os resultados obtidos desta mineração. Através da execução destes algoritmos e na análise dos resultados, indicar qual algoritmo melhor representará o perfil dos pacientes internados.

Por este artigo ser tecnológico, busca a ligação da prática à luz do referencial teórico, de forma a melhorar os processos administrativos pelo resultado da aplicação do referencial teórico. Como também de caráter exploratório, objetivando a melhoria e agregação de novas técnicas aos processos já atualmente utilizados [Alves et al 2008]. Este projeto foi desenvolvido para demonstrar esta nova forma de analisar dados e assim adquirir novos conhecimentos sobre os dados já atualmente utilizados.

Este artigo apresenta os conceitos sobre *data mining*, os algoritmos e sua utilização na ferramenta *Weka* sobre os dados dos pacientes do Hospital Universitário da UFGD/EBSEH. As seções ficaram distribuídas da seguinte forma: a seção 2 apresenta os fatores motivadores para a mineração de dados no hospital. A seção 3 apresenta os conceitos sobre *Data Mining* e a descoberta do conhecimento. A seção 4 apresenta os algoritmos de agrupamentos - *clustering*. A seção 5 apresenta a execução dos algoritmos e seus resultados. A seção 6 apresenta as conclusões finais e o que ainda pode ser analisado nesta temática.

REVISÃO TEÓRICA

A utilização das ferramentas de BI tem se tornado grandes instrumentos para a melhoria da tomada de decisão. Segundo Barbieri (2011) as técnicas OLAP¹ objetivam a consolidação dos fatos e dimensões em diversos níveis de análise, mas as técnicas de *data mining*, segundo Tan et

¹ *On-Line Analytical Processing*, técnica responsável por manipular e analisar grandes volumes de dados em suas múltiplas dimensionalidades.

al. (2009), buscam mais que interpretações, buscam nos fatos as suas correlações em grandes volumes de dados.

Serão analisados os perfis dos pacientes internados durante o ano de 2013 em todas as unidades de internação. Este montante resultou em um arquivo inicial para análise contendo 27.651 registros, contendo o nome do atributo e os valores possíveis de atribuição aos mesmos.

Para o atributo Faixa Etária foram definidos os valores conforme o Estatuto da Criança e do Adolescente (BRASIL 1990), definindo como 'Infantil' as idades menores que 12 anos e 'Adolescente' para as idades acima de 11 e menores que 19 anos. Também foi utilizado o Estatuto do Idoso em (BRASIL 2003), definindo como "Idoso" as idades acima de 59 anos e nas demais idades restantes foram definidos como "Adultos". Para os atributos Sexo e Cor os valores foram baseados nos dados armazenados no banco de dados do sistema AGHU. E finalmente para o atributo Cidade, os valores foram alterados para refletir apenas a informação se o paciente é morador (Y) ou não (N) da cidade de Dourados, para os cadastros não informados a cidade ficou representado pelo sinal de interrogação (?) (WITTEN, 2011).

Será realizada a definição de qual algoritmo será utilizado e a quantidade ideal de *clusters* que representarão estes agrupamentos, e dentro destes clusters qual representará o perfil dos pacientes internados. Com base nestas informações será possível compreender e identificar melhor os principais grupos de pacientes atendidos no hospital.

DATA MINING e a Descoberta do Conhecimento em Bancos de Dados

Como visto anteriormente, o *Data Mining* vem da consolidação de grandes massas de dados, onde segundo Carvalho (2005) o *data warehouse* é a memória dos dados e o *data mining* é a inteligência da empresa sobre estes dados. *Data mining* utilizam técnicas automáticas para a exploração em grandes massas de dados para a descoberta de novos padrões e relações que o ser humano sozinho não seria capaz.

A mineração de dados faz parte da Descoberta de Conhecimento em Bancos de Dados (DCBD) ou *Knowledge Discovery in Databases* (KDD), onde segundo Tan et al (2009) descreve o KDD como o processo de transformação de dados brutos em informações úteis. Este processo é realizado através dos passos de pré-processamento dos dados até o pós-processamento dos resultados da mineração, conforme a Figura 1.

No pré-processamento é realizada a transformação dos dados brutos em um formato apropriado para a análise, através da: integração, limpeza, seleção de registros e características relevantes à mineração. Esta etapa é a que mais demanda trabalho e tempo dentro do processo geral, onde é gerada uma amostra dos dados para mineração.

No pós-processamento é assegurado que apenas resultados válidos e úteis sejam utilizados na tomada de decisão, permitindo aos tomadores de decisão explorem e analisem os resultados

utilizando diversos pontos de vista. Podem ser utilizados medições estatísticas ou teste de hipóteses durante o pós-processamento para retirar resultados não legítimos.

Figura 1: Processo de DCBD segundo. **Fonte:** Salvador et al. (2009).

A metodologia a ser utilizado na mineração depende de quais padrões tem-se interesse, ou seja, nas especificações do que se procura nos dados e desta forma buscar novos padrões. Os algoritmos são as formas aos quais com a utilização de métodos, é possível encontrar novos padrões (SALVADOR et al., 2009).

Para a metodologia de mineração têm-se duas formas de aprendizagem, Salvador et al. (2009) apresentam estas metodologias como sendo: (a) Supervisionada: que tem como característica a predição, ao qual determina com base em padrões uma previsão ou tendência em informações ausentes ou desconhecidas. São tarefas desta metodologia: a.1) Classificação: realiza o mapeamento com base em um atributo-meta ou classe, através da comparação entre este atributo-meta ou classe com os demais dados; a.2) Predição: similar à classificação, mas ao invés de utilizar categorias, utiliza valores numéricos para indicar tendências futuras; (b) Não-Supervisionada: neste tipo de tarefa não são utilizados classes ou pré-determinações de como formarão os grupos, tem-se no máximo a quantidade de grupos que poderão ser formadas. São tarefas desta metodologia: b.1) Dependência: permite a descrição da dependência entre atributos, podendo ser estrutural, ou seja, a dependência entre os atributos localmente; ou quantitativa, que expressa a numericamente o grau da dependência entre os atributos; b.2) Segmentação: realiza o agrupamento em classes de forma natural dos dados de acordo com as similaridades dos itens intra-classes e nas diferenças entre as classes; b.3) Utilizando as informações anteriores e comparando com os dados de entrada para a mineração, tem-se como metodologia de mineração a ser utilizada a aprendizagem Não-Supervisionada, pelo fato de não haver uma classe pré-determinada para realizar a comparação com os dados de entrada que serão utilizados. Na metodologia não-supervisionada será utilizada a tarefa de Segmentação, pois não há dependência entre os atributos localmente e a mineração será realizada apenas pelo simples agrupamento das similaridades entre os valores dos atributos.

As técnicas existentes para Segmentação são conhecidas por *Clustering* ou Agrupamentos, ao qual ressaltam as similaridades entre os objetos e as diferenças entre as classes. Nesta técnica encontram-se diversos algoritmos disponíveis na ferramenta *Weka*, como: *CLOPE*, *CobWeb*, *DBScan*, *EM*, *FarthestFirst*, *FilteredClusterer*, *HierarchicalClusterer*, *MakeDensityBasedClusterer*, *OPTICS* e *SimpleKMeans*.

A escolha da ferramenta de mineração teve como base as ferramentas *R Data Miner*, *RapidMiner* e *Weka*. A ferramenta *R Data Miner* apresentou apenas 04 algoritmos de *cluster*, sendo eles: *KMeans*, *Ewkm*, *Hierarchical* e *BiCluster*. A ferramenta *RapidMiner*, necessita de domínio na construção fluxos para o processo de mineração, ou seja, é necessário primeiro dominar a ferramenta para a montagem de todos os fluxos dos algoritmos. A ferramenta *Weka* possui 10 algoritmos de *cluster* e a mineração pode ser realizada com a carga do arquivo de entrada e a escolha do algoritmo. Desta forma a ferramenta que melhor contempla à necessidade é a *Weka*. A versão utilizada será a 3.6.10 nas simulações e os resultados serão salvos para a análise posterior à definição do algoritmo que será utilizado. A simulação será realizada, através da execução dos algoritmos no modo *Use training set*, sendo ao final visualizado e salvo o *Cluster Assignments* para análise detalhada.

Algoritmos de Clustering no Weka

As técnicas de *Clustering* são utilizadas na inexistência de uma classe já prevista, mas permitem desenhar fortes semelhanças entre si nos dados de entrada, como também a maior diferença entre eles, e desta forma, tornar mais distintos os agrupamentos. Os resultados de um *clustering* podem ser representados de diversas formas, como: exclusiva, onde cada instância pertence a apenas um grupo; sobrepostas, cada instância pode pertencer a mais de um grupo; probabilísticas, cada instância pode pertencer a cada grupo em certa probabilidade; e hierárquicas, onde são formados grupos de nível superior e que estes grupos podem ser refinados até casos individuais (WITTEN, 2011).

As escolhas entre as formas apresentadas dependem da natureza dos dados em relação aos resultados desejados do *clustering*, principalmente pelo fato de que estes grupos não serem conhecidos. Justamente são estes grupos é o que está sendo procurado, e que muitas vezes para sua definição, dependemos da percepção visual humana e seu conhecimento sobre o assunto para defini-los.

O *Weka* implementa os seguintes algoritmos de clustering: (1) *CLOPE*: utilizado para *cluster* em dados transacionais, com a utilização de histogramas para representar a contagem das ocorrências dos itens em *cluster*; (2) *CobWeb*: implementa o algoritmo *CobWeb* para atributos nominais e *Classit* para atributos numéricos, em uma clusterização hierárquica conceitual; (3) *DBScan*: utiliza a métrica da distância Euclidiana para determinar quais instâncias pertencem a um ou outro *cluster*; (4) *EM (Expectation Maximization)*: realiza o cálculo das probabilidades do *cluster*

e com base nos parâmetros da distribuição faz a reestimativa dos parâmetros (maximização); (5) *FarthestFirst*: utiliza o algoritmo *FarthestFirst*, onde o cluster modelado é próximo ao modelado pelo algoritmo *k-means*; (6) *FilteredClusterer*: os dados são passados por um filtro antes de atingir um *cluster*, o *cluster* e o filtro podem ser configurados pelo usuário; (7) *HierarchicalClusterer*: implementa aglomerados do tipo *bottum-up* (base ao topo) para geração de agrupamentos hierárquicos; (8) *MakeDensityBasedClusterer*: é um metaclusterer que retorna a distribuição de probabilidade e densidade; (9) *OPTICS*: é uma extensão do *DBScan*, ao qual impões uma ordenação, com visualização bidimensional e expõe a estrutura hierárquica do *cluster*; (10) *SimpleKMeans*: permite a utilização das médias Euclidiana e Manhattan como métrica para calcular a distância entre os centróides do *cluster*.

Weka é uma coleção de algoritmos para aprendizado de máquina e pré-processamento de dados. É possível realizar experimentos, preparação dos dados de entrada, avaliação dos sistemas de aprendizagem, visualizar os dados de entrada e resultados (WITTEN et al., 2011). *Weka* foi desenvolvido pela Universidade Waikato, na Nova Zelândia, que significa *Waikato Environment for Knowledge Analysis*. O sistema foi escrito em Java e licenciado pela GNU *General Public License*. Possui uma interface gráfica chamado de Explorer, que permite a entrada de dados através de um arquivo no formato ARFF, names, data, csv, libsvm, dat, bsi e XRFF.

Será utilizado o padrão ARFF para representação do conjunto de dados para o *Weka*. Segundo Witten et al. (2011), o arquivo possui o nome da relação, um bloco para definição dos atributos e o conjunto de valores possíveis e ao final os dados, conforme a Tabela 1.

Quadro 1: Atributos e valores do arquivo de entrada do *Weka*.

Atributo	Valores atribuídos
Faixa Etária	Infantil, Adolescente, Adulto e Idoso
Sexo	Feminino, Masculino, Ignorado
Cor	Amarela, Branca, Indígena, Parda e Preta
Clinica	ANES, CCARD, CARD, CGE, CGI, CIRCP, CME, CPE, CPL, COLOPROC, CVA, CCIR, ENDOMET, GASTROE, GOB, HHE, INFEC, MIN, NCI, NEFROL, NEUROCIR, NEUROL, NEURPED, NEONAT, OFTALM, ORTOP, OTOR, PEDIAT, PNEUM, PSIC, REUM, URO
Cidade	Y, ?, N
Mês	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
GrauInstrução	NENHUM, IGNORADO, SUPERIOR, 1GRAUINCOMPLETO, 2GRAUINCOMPLETO, 2GRAUCOMPLETO, 1GRAUCOMPLETO, SUPERIORINCOMPLETO

RESULTADOS E DISCUSSÃO

Na Tabela 2, podemos visualizar os resultados alcançados nas primeiras simulações. Será utilizado como regra para definição do algoritmo definitivo para a definição do perfil, o algoritmo que tiver o menor tempo de execução. O algoritmo que resultou no menor tempo foi o *SimpleKMeans* com 0,08 segundos em relação aos demais algoritmos.

Tabela 1: Resultados obtidos do arquivo de entrada no *Weka*.

Algoritmo	Tempo (seg.)
CLOPE	0,85

CobWeb	2,33
DBScan	0,17
EM	101,73
FarthestFirst	0,10
FilteredClusterer	0,09
HierarchicalClusterer	-
MakeDensityBasedClusterer	0,23
OPTICS	4,44
SimpleKMeans	0,08

O algoritmo *SimpleKMeans*, consiste na utilização de centróides para representar o ponto médio de localização em *K clusters* desejados. E desta forma aumentar a coesão interna entre as similaridades dos *K clusters*, e ao mesmo tempo diminuir a coesão externa entre estes *K clusters* (TAN et al., 2009).

Tabela 2: Resultados obtidos da informação *Sum of squared errors*.

Número de Clusters	Total SSE	Diferença do total SSE entre os clusters ²
2	83.384	-
3	77.347	6.037
4	73.709	3.638
5	71.745	1.964
6	70.111	1.634
7	67.260	2.851
8	66.848	412
9	66.294	554
10	64.710	1.584
11	63.558	1.152

A utilização deste algoritmo possui uma vantagem pela utilização da informação *sum of squared errors* (SSE), ou seja, a soma dos quadrados dos erros. Esta soma é medida para calcular a diferença entre os valores das médias aritméticas, e desta forma aprimorar a precisão da média (TAN et al., 2009). A utilização deste dado servirá para definir qual o melhor número de *cluster* sobre os dados utilizados e com este, identificar o perfil dos pacientes internados. A Tabela 3 apresenta o número de *clusters* utilizados nas simulações, resultados obtidos por número de *cluster* e a variação entre SSE em relação ao número de *cluster* utilizados.

Pelos resultados obtidos a melhor configuração para números de *cluster* será 11 *clusters*, por apresentar o menor número do SSE e também por estar em uma faixa com menor variação do SSE entre a quantidade de *clusters* utilizados na simulação (*clusters* 8 e 11 da coluna 3 na Tabela 3). Como resultado, utilizar 11 *clusters* torna-se uma boa opção para que não haja um número excessivo de *clusters* na prática.

Na execução do algoritmo *SimpleKMeans* com 11 *clusters*, temos como resultados a Tabela 4 contendo os valores de cada atributo em cada número de *cluster* utilizado. Para demonstrar mais facilmente os dados da Tabela 4, a Figura 2 apresenta o gráfico dos *clusters*, sendo no eixo x os *clusters* e no eixo y o total das instâncias por *cluster*.

Observando o gráfico da Figura 2, nota-se como maior *cluster* o de número 4 para o conjunto de atributos e valores com total de 5.755. Este *cluster* apresentou os atributos com valores faixa

² Diferença calculada pela subtração do SSE resultante do *cluster* anterior menos o atual.

etários na idade Adulta, o sexo sendo Feminina, a cor como Parda, a Clínica Médica para Ginecologia e Obstetrícia, residência em Dourados, no Mês de Abril e com escolaridade de 1º grau incompleto. Desta forma este é o maior grupo de pacientes internados no hospital.

Figura 2: Gráficos dos Clusters obtidos com *SimpleKMeans*.

Tabela 3: Clusters obtidos com *SimpleKMeans*.

Cluster	Faixa etária	Sexo	Cor	Clínica	Cidade	Mês	Grau Instrução	Instâncias
0	Infantil	M	Branca	PEDIAT	N	6	Nenhum	4.856
1	Adolescente	F	Parda	GOB	Y	12	1GrauIncompleto	3.335
2	Adulto	F	Branca	CME	Y	5	1GrauIncompleto	3.798
3	Infantil	M	Parda	PEDIAT	Y	6	Ignorado	3.470
4	Adulto	F	Parda	GOB	Y	4	1GrauIncompleto	5.755
5	Infantil	M	Indígena	Pediat	Y	4	Nenhum	846
6	Adulto	F	Parda	CME	Y	12	Ignorado	1.816
7	Adolescente	F	Indígena	GOB	N	1	1GrauIncompleto	337
8	Adulto	F	Branca	CME	Y	2	1GrauIncompleto	467
9	Adulto	F	Branca	GOB	Y	2	2GrauCompleto	2.033
10	Adulto	F	Parda	GOB	N	4	2GrauCompleto	938

CONCLUSÕES

O algoritmo *SimpleKMeans*, foi a melhor opção por apresentar o menor tempo de execução, e a utilização do *sum of square errors* permitiu balizar o número de *clusters* necessários para ser utilizado na simulação. Esta limitação na quantidade de *clusters* se faz necessário para não se ter números excessivos de *clusters*, o que não resolveria o problema de definir o perfil dos pacientes internados. O excesso dos *clusters* reduziria o *sum of squared errors* a zero e desta forma chegaria ao ponto de ter um registro para cada *cluster*.

Utilizando como fonte de dados para a mineração o *data warehouse* sugere-se a inclusão de outros atributos como, o código internacional de doenças e o procedimento realizado. Desta forma o *clustering* não atenderia apenas à área administrativa, poderia estar colaborando com a área assistencial, indicando fatos como a incidência de determinada doença, o procedimento realizado ou a sazonalidade destas enfermidades.

Desta mesma forma o estudo do perfil epidemiológico pode ser expandido para as consultas ambulatoriais. Este estudo pode contribuir igualmente para acompanhar as incidências de determinadas doenças e a sazonalidade das mesmas.

REFERÊNCIAS

ALVES, R. M.; PÁDUA, C. I. P. S.; ZAMBALDE, A. L.. **O documento científico em ciência da computação e sistemas de informação**. Lavras: DCC/UFLA, 2008.

BARBIERI, C.. **Business Intelligence**: modelagem & qualidade. Rio de Janeiro: Editora Elsevier, 2011.

BRASIL. **Lei nº 8.069 de 13 de julho de 1990**. Dispõe sobre o Estatuto da Criança e do Adolescente e dá outras providências. Brasília, 1990.

BRASIL. **Lei nº 10.741 de 01 de outubro de 2003**. Dispõe sobre o Estatuto do Idoso e dá outras providências. Brasília, 2003.

CARVALHO, L. A. V.. **Data Mining**: A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. Rio de Janeiro: Ciência Moderna, 2005.

SALVADOR, H. G.; CUNHA, A. M.; CORRÊA, C. S.. *Vedalogic*: um método de Verificação de Dados Climatológicos Apoiados em Modelos Minerados. **Revista Brasileira de Meteorologia**, São Paulo, v.24, n.4, 2009.

TAN, P. N.; STEINBACH, M.; KUMAR, V.. **Introdução ao Data Mining**: Mineração de Dados. Rio de Janeiro: Moderna, 2009.

WITTEN, I. H.; FRANK, E.; HALL, M. A.. **Data Mining: Practical Machine Learning Tools and Techniques**. 3 ed. Morgan Kaufmann, 2011.