Análise de Perfis de Doenças com Base em Técnicas de Descoberta de Conhecimento em Bases de Dados

Kaique Augusto Morais da Silva[1], Rodrigo Feuser[2], Richardson Ribeiro[2], Dalcimar Casanova[2], Marcelo Teixeira[2], André Pinz Borges[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR), Câmpus Ponta Grossa, Departamento de Informática, Ponta Grossa, Brasil

[2]Universidade Tecnológica Federal do Paraná (UTFPR), Câmpus Pato Branco, Departamento de Informática, Pato Branco, Brasil

{kaiqmo, dalcimar}@gmail.com, rjfeuser@hotmail.com, {marceloteixeira, richardsonr, apborges}@utfpr.edu.br

Abstract. *In this work, disease profiles are analyzed based on techniques for discovering knowledge of databases of the electronic medical records of the Unified Health System (SUS) patients. The analysis focused on two disease groups: neoplasms and trauma. The process used the C4.5, Bagging and Boosting algorithms to create rules that help identify user profiles in health facilities. Compared to previous work, our approach is superior in terms of interpretation of data by health professionals.*

Resumo. *Neste trabalho são analisados perfis de doenças com base em técnicas de descoberta do conhecimento de bases de dados do prontuário eletrônico dos pacientes do Sistema Único de Saúde (SUS). A análise concentrou-se em dois grupos de doenças: neoplasias e traumatismos. O processo utilizou os algoritmos C4.5, Bagging e Boosting para criar regras que auxiliem na identificação de perfis de usuários em unidades de saúde. Em comparação a trabalhos anteriores, nossa abordagem é superior em termos de interpretação dos dados por profissionais da saúde.*

Keywords: Knowledge Discovery; Electronic Health Record.

Palavras-chave: Descoberta de Conhecimento; Prontuário Eletrônico do Paciente.

1. Introdução

Atividades clínicas, como consultas, exames laboratoriais, prescrições médicas diagnósticos, vacinações, entre outras, são realizadas constantemente pelos profissionais da área da saúde. Consequentemente, tais atividades geram uma quantidade significativa de dados, documentados geralmente quando ocorrem consultas médicas ou exames laboratoriais. Na área da saúde, esses dados possuem um modo específico para armazenamento, denominado Prontuário Eletrônico do Paciente (PEP) [Witten 2011].

O PEP normalmente armazena dados como nome, idade, profissão, problemas de saúde do paciente, entre outros. O uso destes tem ajudado cada vez mais as entidades de saúde a gerenciar dados, como informações dos pacientes, profissionais, médicos, etc. [Krysztof 2002]. Apesar dos investimentos em sistemas de informação para melhorar a gestão dos PEPs, esforços são necessários para a geração de conhecimento com os dados armazenados.

Um dos meios para explorar dados de PEP é por meio da Descoberta de Conhecimento em Banco de Dados (em inglês, *Knowledge Discovery in Databases* - KDD). Por KDD entende-se o processo, normalmente não trivial, de obter informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados [Fayyad, Piatetsky-Shapiro and Smyth 1996]. KDD permite obter novas informações após uma série de etapas, como: seleção dos dados, pré-processamento e limpeza, transformação, mineração de dados e interpretação/avaliação dos resultados obtidos.

Neste trabalho foram aplicadas as etapas do KDD em uma base de dados de PEPs, com dados de pacientes da cidade de Pato Branco, Paraná. Inicialmente, foram levantadas as doenças contidas na base e, após um estudo, foi decidido trabalhar com dois grupos de doenças: neoplasias e traumatismos. Tais doenças foram escolhidas por apresentarem diferentes sub-grupos de doenças e por serem aquelas com maior registros na base. Na etapa de mineração de dados foi aplicado os algoritmos C4.5 [Quinlan 1993], *Bagging* [Breiman 1996] e *Boosting* [Freund and Schapire 1996], com objetivo de criar regras que auxiliem na identificação de perfis de usuários que contenham estas doenças.

2. Referencial

A Mineração de Dados é uma das etapas do KDD que, em seu processo, utiliza conceitos de base de dados, métodos estatísticos, ferramentas para visualização e técnicas de Inteligência Artificial (IA). O processo divide-se nas etapas de Seleção, Pré-processamento, Transformação, Mineração de Dados (MD) e Interpretação de dados [Fayyad, *et al.* 1996]. Na etapa de seleção é realizada a escolha da base e dos dados de interesse para o KDD. Na sequência é possível iniciar o pré-processamento, onde dados ruidosos e incompletos são removidos. Na etapa de formatação é feita a transformação dos dados para aplicação dos modelos de MD. E com a MD, algoritmos são aplicados de modo a classificar os dados gerando padrões, os quais podem ser avaliados para a descoberta de conhecimento sobre os dados de interesse. Durante o processo de interpretação é possível voltar as etapas anteriores para que seja melhorado a saída e obtenha-se melhores resultados. Ao completar o processo obtemos conhecimento sobre o banco de dados que foi usado para este processo.

Como dito anteriormente, o processo de KDD é aplicado neste trabalho em PEPs. Os PEPs possuem como vantagens: a agilidade no preenchimento dos documentos; segurança dos dados; atualização em tempo real e a portabilidade. Na literatura, é possível encontrar diversas pesquisas a partir de dados oriundos de PEP. Vilarinho R. (2017) utilizou algoritmos de mineração para obtenção de informações úteis relativas a casos de Dengue nos munícipios brasileiros. Trindade (2012) aplicou o KDD para a identificação de padrões de comportamento das Hepatites Virais nas bases de dados do

SINAN (Sistema de Informação de Agravos e Notificações) do Sistema Único de Saúde - Governo Federal do Brasil, objetivando subsidiar ações de controle e prevenção da doença. Martins e Lima (2014) apresentam um estudo com as vantagens e desvantagens de usar um PEP, onde, apesar do custo de implantação, os PEPs apresentam grandes vantagens na sua implantação para gestão hospitalar. Destaca-se o trabalho de Feuser R. (2017), que aplicou os processos do KDD para PEP oriundo do SUS, utilizando algoritmo de associação *a priori* na etapa da mineração de dados, encontrando regras com elevado fator de confiança.

Há ainda na literatura diversos outros trabalhos que fizeram uso de dados oriundos de PEP para: análise de custo-efetividade de vacinas [De Soarez 2009], problemas de fratura ortopédica [Zorman *et al.* 2000], tomadas de decisões clínicas [Bae J-M 2014], tratamento de feridas crônicas [Letourneau and Jensen 2008], etc.

Contudo, os trabalhos já desenvolvidos não abordaram as doenças aqui analisadas: neoplasias e traumatismos. Além disso, os trabalhos não utilizaram de técnicas capazes de gerar conhecimento facilmente interpretáveis por profissionais da saúde, como as árvores geradas pelos algoritmos C4.5, *Bagging* e *Boosting*.

3. Desenvolvimento

Os dados utilizados nesta pesquisa foram obtidos por meio de um projeto com a participação das secretárias de Saúde, Ciência e Tecnologia do município de Pato Branco, bem como a participação da Universidade Tecnológica Federal do Paraná (UTFPR) Campus Pato Branco. Os dados foram fornecidos pela empresa responsável por desenvolver o sistema de PEP utilizado no município de Pato Branco. O conjunto de dados contém 43.879 pacientes e 2.296.626 registros de atendimentos. Dados pessoais, como nome, RG, CPF, telefone, e outros, não fazem parte do escopo desse projeto. Portanto, em momento algum, os pesquisadores desse projeto souberam a identificação de pacientes ou de seus responsáveis.

Como etapa de pré-processamento, foram executados os seguintes procedimentos: (i) identificação da faixa etária dos pacientes; (ii) utilização de dados do censo (e-SUS) tais como: bairro, altura, peso, frequência escolar, se frequenta benzedeira, se possui plano de saúde, se é fumante, se possui diabetes, se é gestante, se possui asma, se é alcoólatra, se já teve infarto, AVC ou derrame; (iii) remoção de registros com dados ausentes ou com ruídos foram eliminados, como por exemplo, pacientes que não possuem identificação de doenças ou que possuem apenas informações de exames; (iv) remoção de dados discrepantes relacionados à altura e peso dos pacientes, utilizando para isto valores mínimos e máximos.

A base de dados foi enriquecida com dados do grupo de Classificação Internacional de Doenças e Problemas Relacionados à Saúde (CID) das doenças [OMS, 2019]. Todos estes dados derivados foram também utilizados na geração das árvores de decisão na etapa de MD, cujo objetivo (classe alvo) era o grupo CID.

Uma análise dos grupos de CIDs foi realizada com objetivo de tratar apenas os CIDs referentes às doenças estudadas neste trabalho: neoplasias e traumatismos. As duas doenças foram escolhidas por apresentarem, no conjunto de dados, um maior equilíbrio no número de instâncias entre os diferentes CIDs. Esta filtragem ocorreu devido à falta

de recursos computacionais para processar todo o conjunto de dados. O grupo de traumatismos apresentou um total de 236 instâncias, e o grupo Neoplasias um total de 119 instâncias para a MD (conforme Tabelas 1 e 2).

Tabela 1: Quantidade de ocorrências de Traumatismos.

Tipo de Traumatismo	Qtde
Outras causas externas de traumatismos acidentais	16
Sequelas de traumatismos, intoxicações e de outras consequências das causas externas	5
Cabeça	18
Abdome, dorso, coluna lombar e da pelve	4
De localização não especificada do tronco, membro ou outra região do corpo	13
Cotovelo e do antebraço	11
Joelho e da perna	51
Ombro e do braço	7
Pescoço	3
Punho e da mão	49
Quadril e da coxa	7
Tórax	33
Tornozelo e do pé	40
Total Geral	**257**

Tabela 2: Quantidade de ocorrências de Neoplasias.

Tipo de Neoplasia	Qtde
Melanoma e outras(os) neoplasias malignas da pele	2
Neoplasias (tumores) benignas(os)	65
Neoplasias (tumores) de comportamento incerto ou desconhecido	3
Neoplasias (tumores) in situ	20
Neoplasias (tumores) malignas(os)	3
Neoplasias (tumores) malignas(os), declaradas ou presumidas como primárias, dos tecidos linfático, hematopoético e tecidos correlatos	7
Neoplasias malignas da mama	2
Neoplasias malignas do aparelho respiratório e dos órgãos intratorácicos	7
Neoplasias malignas dos órgãos digestivos	4
Neoplasias malignas dos órgãos genitais masculinos	6
Total Geral	**119**

4. Resultados

Para a etapa de mineração de dados foi utilizada a ferramenta Weka[1] 3.8, onde foram executados os algoritmos C4.5 (algoritmo J48), *Bagging* e *Boosting* (algoritmo *AdaBoostM1*), todos utilizando as configurações padrão do Weka.

Para a execução no WEKA também é escolhido o método de testes de validação cruzada com 10 *folds* para a execução dos três algoritmos. Os resultados obtidos visam analisar os percentuais de acerto dos algoritmos de classificação. A Tabela 3 apresenta os percentuais de acerto, verdadeiros e falsos positivos obtidos dos classificadores C4.5, *Bagging* e *Boosting* referentes aos dois grupos de doenças estudados.

É possível observar que o grupo de neoplasias apresentou uma melhor precisão em todos os casos, sendo o algoritmo de *Boosting* aquele que melhor classificou o conjunto de dados, alcançando 87% de acerto. Ainda em neoplasias, o meta-classificador *Bagging* não foi capaz de melhorar a taxa de acerto do C4.5, algo que aconteceu no grupo de doenças relacionadas aos traumatismos. É interessante observar que nenhum falso positivo foi obtido para Traumatismos, apesar da menor precisão dos classificadores.

[1] Weka: https://www.cs.waikato.ac.nz/ml/weka/

Tabela 3 – Dados comparativos dos algoritmos utilizados.

	Neoplasias			Traumatismos		
	J48	Bagging	Boosting	J48	Bagging	Boosting
Verdadeiros positivos	84%	85%	87%	8%	7%	7%
Falso positivos	9%	10%	8%	0%	0%	0%
Precisão	70%	70%	87%	60%	61%	65%

No que diz respeito ao conteúdo das árvores obtidas, pôde-se observar, para ambos os conjuntos de doenças, que o principal fator na identificação das doenças foram os atributos bairro, indicação de frequência escolar e faixa etária. Todos os demais atributos utilizados apareceram em regras específicas e em menor quantidade nas árvores geradas.

5. Conclusões

A aplicação das etapas do KDD é uma alternativa para a geração do conhecimento em base de dados da área da saúde. É possível perceber o aumento de aplicações que fazem uso do KDD na área da saúde, evidenciando sua necessidade de utilização devido ao aumento exponencialmente de atendimentos clínicos, exames, etc. Uma das dificuldades é o acesso a pesquisa destes bancos de dados. Ainda não se dispõe de legislações sobre a disponibilidade de prontuários eletrônicos padronizados, ou centralizados, o que facilitaria o acesso ao histórico de tratamento do paciente por outras unidades de saúde e médicos.

A relação de atendimentos de unidades de pronto atendimento, são muito peculiares devido ao fato de ser em geral procurado pelos usuários para tratamento de dores agudas, acidentes ou males repentinos. Neste caso as correlações de padrões encontradas pelos algoritmos podem ser afetadas por vários fatores, mas podem inevitavelmente demonstrar fatores ainda não detectados em outras formas de visualizações.

Ao finalizar a aplicação das etapas do KDD para análise de perfis de doenças e dividir a amostra em dois grupos de doenças, um relacionado a neoplasias e o outro relacionado a traumatismos, foi possível determinar os principais atributos a serem analisados para identificação das doenças.

Para trabalhos futuros é possível a validação dos resultados obtidos por profissionais da saúde para definir a veracidade das regras encontradas. Isso poderia expandir este trabalho para outros locais de atendimentos de emergência, como: hospitais, ambulatórios e ou clínicas médicas. Outros trabalhos possíveis envolvem a integração com dados do IBGE, comparação com dados de cidades com mesmas características, aplicações de novos parâmetros e algoritmos.

Agradecimentos

Está pesquisa é apoiada por Decit/SCTIE/MS, por intermédio do CNPq, com apoio da Fundação Araucária e da SESA-PR. Programa Pesquisa para o Sistema Único de Saúde: Gestão Compartilhada em Saúde - PPSUS.

Referências

Alves, L. (2018) “Prontuário Eletrônico x Prontuário no papel”. Acessado em: http://meuprontuario.net/prontuario-eletronico-x-prontuario-papel-qual-e-o-melhor/ Disponível em: 24/06/2018.

Bae, J-M. (2014) “The clinical decision analysis using decision tree”. In: Epidemiology and Health. Vol. 36, pg 1-7.

Breiman, L. (1996) "Bagging predictors". Machine Learning. 24 (2): 123–140.

Carvalho, R. (2018) “Prontuário e registro de enfermagem”. Acessado em: http://www.ebah.com.br/content/ABAAAASqAAG/prontuario-registro-enfermagem#. Disponível em: 24/06/2018.

De Soarez, P. C. (2009) “Use of decision analysis in the programs of vaccination against varicella”. São Paulo: “Faculdade de Medicina da Universidade de São Paulo”.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996) “From Data Mining to Knowledge Discovery in Databases”. AI Magazine, AAAI, Boston.

Feuser, R. (2017) “Mineração de Dados com Regras de Associação Aplicada em Dados de Unidade de Saúde de Pronto Atendimento”. UTFPR, Pato Branco (PR).

Freund Y. and Schapire R.E. (1996) “Experiments with a new boosting algorithm”. In Proceedings of the Thirteenth International Conference on International Conference on Machine Learning (ICML'96), Lorenza Saitta (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 148-156.

Krysztof, J.C. (2002) “Uniqueness of Medical Data Mining. Artificial Intelligence in Medicine”, mar.

Letourneau, S., Jensen, L. (2008) "Impact of a decision tree on chronic wound care". Journal Wound Ostomy Continence Nurs, vol. 25, pp. 240-247.

Martins, C. and Lima, S.M. (2014) “Vantagens e desvantagens do prontuário eletrônico para instituição de saúde”. Revista de Administração em Saúde (RAS), v. 16, n. 63.

Organização Mundial da Saúde (OMS). (2019) “Centro Colaborador da OMS para a Classificação de Doenças em Português (CBCD)”. Classificação estatística Internacional de Doenças e Problemas Relacionados a Saúde - CID - 10. 2008. Disponível em http://www.datasus.gov.br/cid10/V2008/cid10.htm. Acesso em 02/02/2019.

Quinlan, J.R. (1993) “C4.5: Programs for Machine Learning”. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Thomaz, M. (2018) “Tudo que você precisa saber sobre prontuários eletrônicos”. Disponível em: https://blog.iclinic.com.br/tudo-sobre-prontuario-eletronico/. Acessado em: 24/06/2018.

Trindade, C M et al. (2012) “Technology in health: knowledge discovery in public health databases: study of viral hepatitis in the state of Paraná”, Brazil. Iberoamerican Journal of Applied Computing, Ponta Grossa, v. 2, n. 2.

Vilarinho R.A. (2017) “Uso de Técnicas de Mineração de Dados para Classificação das Ocorrências de Casos de Dengue nos Municípios Brasileiros”. UFOP – Universidade Federal de Ouro Preto. Março.

Zorman, M., Podgorelec, V., Kokol, P., Peterson and M., Lane, J. (2000) “Decision tree's induction strategies evaluated on a hard real world problem”. In: Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems CBMS'2000, pp. 19-24.

Witten, I. H. (2011) “Data Mining Practical Machine Learning Tools and Techniques”. 3 ed.