Detecção de anomalia através da comparação de modelos representativos

Giovana Jaskulski Gelatti[1], **Pedro Pereira Rodrigues** [2], **André Carlos P. L. F. de Carvalho**[1]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
Av. Trab. São Carlense, 400 - Centro, 13566-590, São Carlos - SP - BR

[2]Faculdade de Medicina da Universidade do Porto (FMUP)
R. Dr. Plácido da Costa, 4200-450 Porto - PT

gi.jask.gelatti@gmail.com, pprodrigues@med.up.pt , andre@icmc.usp.br

Resumo. *Existem barreiras burocráticas e de ideais que tornam a comparação de departamentos e identificação de padrões, em geral, tarefas difíceis. A exploração dos dados coletados, juntamente com um modelo descritivo induzido a partir desses dados, podem ajudar a identificar modelos destoantes e promover a comparação das instituições. O estudo propõe a criação de modelos de redes Bayesianas capazes de representar e extrair conhecimentos novos e significativos a partir dos dados nas variáveis utilizadas no estudo de caso. São selecionadas variáveis sobre secções obstétricas de hospitais de Portugal para a criação do modelo de cada entidade/secção. As funções do pacote R "bnlearn"foram usadas para manipular e recriar dados no modelo. O desempenho deste modelo foi validado quanto à capacidade de recriar dados. Para construir uma matriz de distâncias entre modelos para identificação de entidades destoantes, a distância de Hamming. As anomalias detectadas pela comparação dos modelos criados foram validadas por especialista de acordo com a escala Likert. Os dados foram descritos e recriados através de redes Bayesianas com imputação de dados, com referência significativa aos dados reais. A comparação dos modelos sobre as secções de obstetrícia identificou padrões e anomalias. A comparação permitiu diferenciar os setores com diferentes taxas de cesárea e distribuição nos grupos de Robson, de acordo com as variáveis selecionadas, preservando o acesso aos dados reais das instituições.*

1. Introdução

A detecção de anomalias e/ou *outliers* é definida como detecção de observações anormais baseada em modelos predefinidos normais [Wu and Banzhaf 2010], entre outros nomes [Tan et al. 2006]. A anotação sobre os dados e modelos como "anômalos ou não"requer alto custo de mão de obra especializada e tempo para análise. Por esta razão, técnicas não supervisionadas de detecção de anomalia são escolhidas para o estudo desta identificação.

Considerando uma entidade como um conjunto de dados que representa algo, ao comparar as entidades, interessa-nos descobrir as influências dos atributos nas possíveis decisões e a relação das entidades entre elas. Porém, o processo de requerer, analisar, processar e compartilhar dados e resultados entre entidades não é uma tarefa simples, primeiramente pelas leis que regem o acesso aos dados.

Conforme o RGPD (Regulamento Geral sobre a Proteção de Dados) da UE (União Europeia) 2016/679 [reg 2016], os dados só podem ser compartilhados com o consentimento das entidades as quais os dados representam. Porém, de acordo com a lei sob o Artigo 3º da Lei portuguesa nº 12/2005, dados de saúde podem ser compartilhados sem o consentimento dos titulares dos dados para fins de investigação científica, desde que anonimizados e com autorização do responsável pelo acesso à informação. Este processo ocorre juntamente com a autorização pelo comitê de ética de cada instituição seguindo Lei Geral de Proteção de Dados Pessoais (Lei n.º 58/2019 [Diário da República n.º 151/2019]). No Brasil, similar ao Regulamento na UE, a LGPD (Lei Geral de Proteção de Dados Pessoais [Diário Oficial da União 2018]) Lei nº 13.709, permite o tratamento de dados pessoais "para a realização de estudos por órgão de pesquisa, garantida, sempre que possível, a anonimização dos dados pessoais".

E, para além da barreira jurídica e burocrática, muitos estudos apontam a insegurança de investigadores na avaliação dos seus resultados e o uso de resultados sobre bases de dados em colheita ou ainda em desenvolvimento [Wicherts et al. 2011, Walport and Brest 2011, Hutson 2018]. Essas instituições acreditam em uma falsa segurança ao não compartilhar suas informações enquanto há perda de informação e de conhecimento que poderiam ser obtidos a partir destes dados. Um destes benefícios, no contexto desta pesquisa, é a possibilidade de, pela sabedoria de multidão, agregar o conhecimento de sessões obstétricas.

Em um cenário ideal, os partos por cesarianas são realizados apenas em situações em que a grávida ou o bebê estão em situação de risco de vida. No entanto, o cenário real parte da utilização em partos de alto risco para uma possível banalização deste tipo de parto e preocupação de seus elevados índices. Para fins de entendimento dos índices de cesáreas e classificação do tipo de parto, a OMS [Organization et al. 2015] indica o uso da classificação de Robson [Robson 2001] como forma de classificar as grávidas para entendimento da do cenário obstétrico.

No entanto, a distribuição de tipos de parto pelos grupos de Robson atualmente não condiz com o esperado indicado como um padrão a ser seguido [Robson et al. 2013]. Interessa-nos identificar se outras variáveis sejam estatisticamente relevantes para incluir ao estudo. Pelas razões apresentadas, é proposta a criação de redes Bayesianas a partir dos dados reais localmente nos hospitais, assim conservando os dados reais. A partir das redes Bayesianas é possível comparar secções obstétricas sem o acesso e a partir delas recriar dados que correspondam aos dados reais para mais análises.

2. Material e métodos

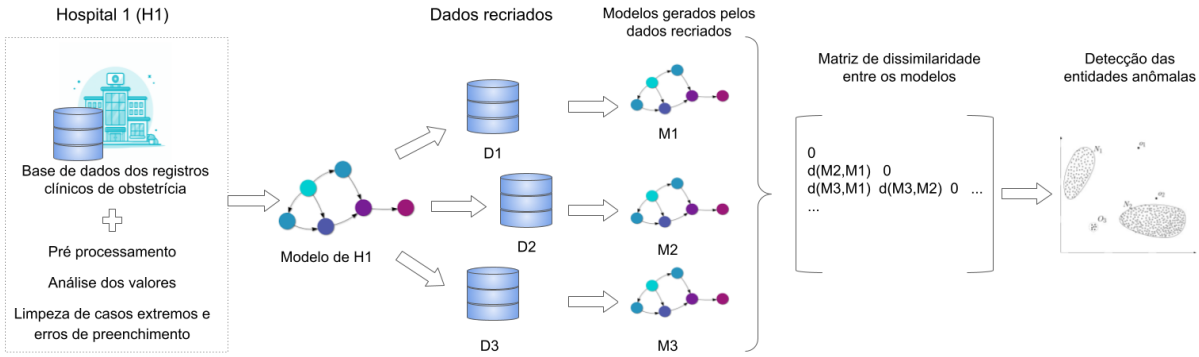
Foram mapeadas as variáveis do sistema Obscare da empresa VirtualCare[1] juntamente com especialista e identificadas as variáveis estatisticamente relevantes para a definição do tipo de parto segundo $p - value$. O Obscare [Cruz-Correia and Amorim 2008] é uma ferramenta de armazenamento e processamento de dados, entregando relatórios de qualidade de dados e índices obstétricos e atualmente é utilizada por 12 hospitais de Portugal.

Os modelos de redes Bayesianas (RB) de cada hospital ou secção obstétrica de hospitais gerados a partir dos dados reais, são utilizados para a amostragem de dados

[1] Saber mais em http://virtualcare.pt/

recriando dados a partir do modelo. A partir desses dados recriados, são gerados modelos representativos. Estes modelos criados pelos dados recriados são comparados pela distância Hamming. O processo é ilustrado na Figura 1.

Figura 1. Processo de criação de modelos e amostragem de dados deles em uma instituição exemplo H1



Para a avaliação de dados e informação médica, os dados foram cedidos pela empresa VirtualCare após o acesso autorizado pelos comitês de ética. Os hospitais participantes deste caso de estudo, respectivamente com as autorizações e pareceres dos comitês de ética: HSO (85/2020), CHEDV (CA-371/2020), CHVNGE (192/2020). Cada registro corresponde a uma gestação e parto e as informações são sobre antecedentes clínicos, gestação entre outras informações até o momento do parto. Em caso de gêmeos, o tipo de parto do segundo bebê é desconsiderado para manter um registro apenas pela gestação. Para confidencialidade dos hospitais, eles serão referenciados como H1, H2 e H3 nomeados em ordem aleatória.

O caso de estudo também agrega a análise de *missings* e sua relação entre as variáveis. Em primeiro lugar, analisamos os valores ausentes realizando o preenchimento de dados determinístico, em que a própria informação consta na base de dados. Após, variáveis com mais de 90% de dados faltantes foram retiradas da amostra e os valores faltantes reais são preenchidos por técnicas de preenchimento de dados. Para escolher a técnica de preenchimento, são analisados os mecanismos, proposto por [RUBIN 1976], com os quais os valores faltantes foram gerados. A técnica de preenchimento de dados que utiliza os conceitos dos mecanismos é o algoritmo MICE [Van Buuren and Oudshoorn 1999]. Outra técnica de preenchimento utilizada é o preenchimento por moda ou média geral de cada variável ou por grupo.

2.1. Redes Bayesianas

Uma rede Bayesiana é uma representação compacta de uma distribuição de probabilidade [Darwiche 2010], possuindo uma maior explicabilidade do que outros tipos de modelos representativos. A partir delas pode-se também sintetizar um grande volume de informação. Neste contexto, redes Bayesianas foram eleitas para gerar e representar o modelo de comportamento de cada entidade.

Para criar os modelos, será utilizada a ferramenta R e o pacote *bnlearn* [Scutari et al. 2020]. Os modelos são construídos pelos dados totais recebidos das

instituições, o que representa uma entidade neste estudo, produzindo um modelo para cada. A partir dos dados, são aprendidas as probabilidades e relações entre cada variável.

2.2. Detecção de anomalia

Para comparação entre as entidades, foi utilizada a distância Hamming que quantifica a dissemelhança entre redes Bayesianas.

Após a construção da matriz de dissemelhança desses modelos, dois tipos de algoritmos baseados em distância (KNN) e densidade (LOF) são utilizados para a detecção das exceções. O KNN, algoritmo de detecção de anomalia que utiliza o kNN [Ramaswamy et al. 2000],Ele utiliza a distância dos kNN para atribuir pontuações as observações. Dado um conjunto de dados com N observações, cada observação é agrupada aos seus k vizinhos mais próximos. Observações não atribuídas são consideradas anomalias. Já o algoritmo Local Outlier Detection (LOF), proposto por [Breunig et al. 2000], compara a densidade local, atribuindo um fator de anormalidade local de cada observação por meio de um conjunto de kNN do ponto de teste com pontos dentro da vizinhança de cada membro do conjunto kNN. Sendo k um parâmetro a ser definido. Quanto maior a distância k do vizinho, maior a pontuação de anormalidade.

3. Resultados

3.1. Preenchimento de dados

Os dados preenchidos e dados reais foram comparados quanto a distribuição nos grupos de Robson. Na comparação, não houve diferença considerada significativa na distribuição de partos nos grupos de Robson entre os dados reais, os dados preenchidos pela média e moda e os preenchidos pela técnica MICE para cada variável do conjunto de dados.

3.2. Modelos gerados

Foram selecionados os registros com os valores completos das variáveis selecionadas e criados modelos a partir deles. Ao gerar o modelo a partir dos dados reais completos (antes do preenchimento de dados), o modelo representativo do H3 não foram adicionados os nodos de algumas variáveis. A partir destes modelos, foram recriados dados. Foram comparadas as distribuições dos partos nos grupos de Robson nos dados reais, na seleção dos dados completos e nos dados recriados pelos modelos. Ao comparar a distribuição dos partos pelos grupos nos diferentes conjuntos, há grande diferença dos conjuntos reais para a seleção dos dados completos. Esta diferença é refletida quando gerado um modelo a partir da seleção dados completos e recriados dados deste modelo.

A estrutura dos modelos gerados pelos conjunto de dados inserindo os valores faltantes pelas duas técnicas de preenchimento foram iguais. Ao preencher os dados, os nodos omitidos ao selecionar os dados completos voltam a ser representadas nos modelos gerados. Os dados que geraram estes modelos produziram uma distribuição semelhante a dos dados reais independente do tipo de preenchimento. Por este motivo, os dados preenchidos por média e moda foram utilizados para a comparação e detecção de anomalia.

Com o fim de aumentar a amostra, foram realizadas subamostragens de cada hospital com o mesmo volume de dados dos dados originais. Para uma comparação mais próxima ao cenário real e sabendo que havia 33 são hospitais públicos na região Norte

de Portugal em 2019 [Portugal 2020], para simular esta quantidade de hospitais públicos, foram gerados 11 conjuntos de dados por modelo (total das 33 entidades a serem comparadas) a partir dos três modelos de hospitais criados. Sobre os 33 conjunto de dados novos gerados pelos 3 modelos são construídos 33 modelos a serem comparados.

3.3. Detecção de anomalia

O modelo gerado pela seleção dos registros completos dos dados reais do H3 apresentou nodos excluídos da análise foi considerado uma anomalia pois sua estrutura e relação com os dados é tão diferente da maioria dos modelos criados que não pode ser comparado aos demais pela distância calculada. A construção da matriz de dissimilaridade com os outros modelos resultou em uma matriz 33x33 nula. O que sugere que as estruturas e relações geradas pelos dados preenchidos nos diferentes hospitais são idênticas.

Foram incluídos 22 modelos criados a partir dos dados reais completos para encontrar as dissemelhanças de estruturas. Todas comparações foram nulas, exceto com o modelo criado a partir da seleção dos registros completos do hospital H2, que apresentou $distânciaHamming = 16$ a todos outros. Ao executar os algoritmos de detecção de anomalia, este modelo foi classificado como anomalia.

Foi aplicado um questionário apresentando as tabelas sobre a distribuição dos partos nos grupos de Robson nos dados que geraram os modelos considerados anômalos.A especialista concorda em parte que ambas entidades são anomalias.

4. Conclusões e contribuições

O trabalho apresentado corresponde a uma alternativa que apresentou-se favorável para suprir a limitação do compartilhamento dos dados entre instituições e com a comunidade acadêmica construindo localmente modelos que representam as entidades a serem compartilhados com a possibilidade de recriar dados com base nos dados reais a partir deles.

Sobre o contexto do caso de estudo e momento histórico, o trabalho teve limitações de amostra e validação das anomalias encontradas. Apesar da intenção de mais hospitais e especialistas integrarem o estudo, a amostra dos hospitais corresponde a uma parcela dos hospitais do norte de Portugal. Devido a pandemia vivida atualmente, a pesquisa de campo não foi realizada em sua totalidade e por isso a validação foi limitada por esta situação. Apesar das limitações, foi possível contribuir com a construção do modelo representativo para cada hospital integrante ao estudo. Adotando os métodos e *guidelines* realizados nesta pesquisa, pode-se também comparar instituições de diversos contextos e em problemas similares, o que é extremamente encorajado pela autora.

A principal contribuição do trabalho é a possibilidade de realizar análises distribuídas de detecção de anomalia entre entidades não comunicantes sem a utilização dos dados reais. A detecção de anomalia por comparação de redes Bayesianas permitiu realizar a comparação entre entidades sem a partilha de dados entre elas.

Ao conviver e participar da coleta de dados, soube da dificuldade dos hospitais em manterem uma motivação de preencher os dados. A rotina clínica não permite pausas para um preenchimento detalhado e muitas validações podem ser feitas a nível do sistema que coleta os dados. As verificações e validações foram encaminhadas à empresa que realiza a coleta de dados e foi proposto para que as imputações determinísticas tornem-se automáticas no preenchimento do sistema.

Referências

(2016). *Regulamento Geral sobre a Proteção de Dados*. Regulamento (UE) 2016/679 do parlamento europeu e do conselho de 27 de abril de 2016 relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados e que revoga a Diretiva 95/46/CE.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104.

Cruz-Correia, R. and Amorim, E. (2008). Obs.care: aplicacao de cuidados intensivos de obstetricia. *Porto: Universidade do Porto/Faculdade de Medicina/Departamento Ciencias da Informacao e da Decisao em Saude*.

Darwiche, A. (2010). Bayesian networks. *Commun. ACM*, 53(12):80–90.

Diário Oficial da União, Seção 1, p. . (2018). Lei nº 13.709, de 14 de agosto de 2018.

Diário da República n.º 151/2019, S. I. d. .-.-. Lei n.º 58/2019.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis.

Organization, W. H. et al. (2015). Who statement on caesarean section rates. Technical report, World Health Organization.

Portugal, E. (2020). *Instituto Nacional de Estatística - Estatísticas da Saúde : 2018.*

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438.

Robson, M., Hartigan, L., and Murphy, M. (2013). Methods of achieving and maintaining an appropriate caesarean section rate. *Best practice & research Clinical obstetrics & gynaecology*, 27(2):297–308.

Robson, M. S. (2001). Classification of caesarean sections. *Fetal and maternal medicine review*, 12(1):23–39.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Scutari, M., Scutari, M. M., and MMPC, H.-P. (2020). Package 'bnlearn'.

Tan, P.-N., Steinbach, M., Kumar, V., et al. (2006). *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston.

Van Buuren, S. and Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden: TNO.

Walport, M. and Brest, P. (2011). Sharing research data to improve public health. *The Lancet*, 377(9765):537–539.

Wicherts, J. M., Bakker, M., and Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS one*, 6(11):e26828.

Wu, S. X. and Banzhaf, W. (2010). The use of computational intelligence in intrusion detection systems: A review. *Applied Soft Computing*, 10(1):1–35.