Conteúdo disponível em: https://www.ifgoiano.edu.br/periodicos/multiscience

Multi-Science Journal

Website do periódico: https://www.ifgoiano.edu.br/periodicos/index.php/multiscience

Artigo Completo

Descoberta de regras de associação com algoritmo Tertius

Discovery of association rules with Tertius algorithm

Norton Coelho Guimarães[1], Rosângela Nunes[2]

[1]IF Goiano campus Morrinhos, BR 153 KM 633 Zona Rural, Morrinhos, Goiás, Brasil.

[2]Universidade Federal de Goiás, CERCOMP, Goiânia, Goiás, Brasil. rdsnunes@gmail.com

* Autor para correspondência: norton.guimaraes@ifgoiano.edu.br

I N F O A R T I G O

Histórico do artigo
Recebido: 24/09/2016
Aceito: 27/03/2017

Palavras chaves:
Mineração de Dados,
Regras de Associação,
Tertius.

R E S U M O

O trabalho descreve os experimentos do processo de definição das regras de associação com base nas regras do desafio PhysioNet de 2012. O desafio disponibilizou uma base com 4000 pacientes com informações das últimas 48 horas de permanência em uma UTI. Na descoberta das regras de associação, agrupamos os registros por paciente e calculamos a média aritmética de cada atributo temporal. Para realizar os treinamentos utilizamos somente as primeiras 24 horas agrupadas de 6 em 6 horas. Utilizamos o algoritmo Tertius para gerar automaticamente as regras em conjunto com a ferramenta Weka.

Keywords:
Data Mining,
Associate Rules,
Tertius.

A B S T R A C T

This paper describes experiments in the definition of association rules based on the rules of the challenge 2012 Physionet process. The challenge provided a base with 4000 patients with information of the last 48 hours of stay in the ICU. The discovery of association rules, we grouped the records by patient and calculated the arithmetic mean of each temporal attributes. To perform the training use only the first 24 hours grouped into six 6 hours. We use the algorithm Tertius to automatically generate rules together with the Weka tool.

1. Introdução

Uma Unidade de Tratamento Intensivo (UTI) é para pacientes que necessitam de tratamentos especiais, com uso de aparelhos, na maioria das vezes. Estes pacientes são monitorados frequentemente e são coletados diversos parâmetros fisiológicos para serem analisados posteriormente (XIA et. al., 2012).

Atualmente, grupos de saúde estão focados em pesquisar novas técnicas para melhorar a eficácia do tratamento para os pacientes em estado crítico na UTI (BERA e NAYAK, 2012). Apesar destes esforços, o potencial das informações digitais não é plenamente utilizado (POLLARD et. al., 2012).

Por décadas, o custo de permanência na UTI vem se elevando (XIA et. al., 2012). Em 2005, o custo médio de um paciente na UTI variava de 32 mil dólares à 42 mil dólares, para os que necessitavam de equipamentos para sobreviver, e para os pacientes que não necessitavam de aparelhos, constantemente, o custo médio se reduzia e ficava entre 12 mil dólares à 20 mil dólares (GOLDBERGER et. al., 2000; XIA et. al., 2012).

De um modo geral, para resolver o problema de mortalidade na UTI é utilizado a técnicas de regressão linear. Mas, podemos encontrar outros trabalhos que utilizem redes neurais, redes bayesianas, SVM e entre outras (XIA et. al., 2012).

Este trabalho, é uma resposta para o desafio Computing in Cardiology/Physionet Challenge 2012 "Predicting Mortality of ICU Patients". O foco do desafio é desenvolver um método de predição da mortalidade hospitalar na UTI (GOLDBERGER et. al., 2000).

Várias pesquisas têm aplicado diversas técnicas para encontrar soluções que possam resolver o problema de

mortalidade em UTI. Dentre estas técnicas, pode-se cita: Clusterização (XU et. al., 2012), Lógica de Regressão (BERA e NAYAK, 2012; HAMILTON e HAMILTON, 2012), Redes Neurais (XIA et. al., 2012), Redes Bayesianas (MACAS, 2012), SVM (BOSNJAK e MONTILLA, 2012), Motif Discovery (MCMILLAN, 2012) e entre outras.

Alguns dos estudos analisados propuseram que o próprio algoritmo manipule a falta de informação (JOHNSON et. al., 2012). Outros, propuseram um método para manipular a falta de informação através da atribuição de valores. Em Lee (LEE et. al., 2012), os valores faltantes foram preenchidos pela média considerando a idade e o gênero. Em Citi (CITI e BARBIERI, 2012), supôs que a falta de informação indicava uma situação estacionária, e propôs a replicação dos valores medidos anteriormente.

O objetivo, deste trabalho, é desenvolver um método para realizar a descoberta das regras de associação a partir dos registros disponibilizados.

Este trabalho, está organizado da seguinte forma. Na seção 3, descrevemos o desafio e suas etapas. Na seção 2, descrevemos os trabalhos relacionados. Na seção 4, descrevemos o método proposto para a descoberta das regras de associações. Na seção 5, descrevemos os experimentos e os resultados realizados durante a pesquisa. Por último, a conclusão é apresentada.

2. Material e Método

O PhysionNEt/Computing Challenge é um desafio anual, iniciado em 1999, pelo National Institutes of Health e mantido pelo National Institute of Biomedical Imaging and Bioengineering (NIBIB) e National Institute of General Medical Sciences (NIGMS) com o intuito de desenvolver métodos para solução de temas específicos na área médica. Na edição de 2012, o foco foi o desenvolvimento de métodos para a predição de mortalidade de pacientes na UTI. Os dados usados para o desafio, consiste em 5 descrições gerais (veja detalhes na tabela 1) e 37 series temporais (veja detalhes na Tabela 3). Dados estes que foram coletados para às primeiras 48 horas iniciais desde a entrada do paciente na UTI e sua saída (veja detalhes na tabela 2). Foram disponibilizados registros de 12 mil pacientes adultos do banco de dados MIMIC II (SAEED et. al., 2012).

As medições, das séries temporais, são registradas em ordem cronológica e podem ser registradas em intervalos regulares, variando de hora em hora, diariamente ou em intervalos irregulares, conforme necessário. Nem todas as séries temporais foram coletadas em todos os casos (SAEED et. al., 2012).

Tabela 1: Tabela das descrições gerais (GOLDBERGER et. al., 2000)

Parâmetro	Descrição
RecordID	Identificação do paciente
Age	Idade do paciente
Gender	0: mulher ou 1 homem.
Height	Altura do paciente em cm
ICUType	Tipo da Série Temporal.
Weight	Peso do paciente em kg

Tabela 2: Tabela das descrições de saída do paciente (GOLDBERGER et. al., 2000)

Parâmetro	Descrição
RecordID	Identificação do paciente
SAPS-I score	Referência do estado psíquico
SOFA score	Referência da morte
Length of stay	Quantidade de dias de permanência no hospital
Survival	Quantidade dias de permanência na UTI
In-hospital death	0: Sobreviveu; 1: Morreu

Tabela 3: Tabela das séries temporais (GOLDBERGER et. al., 2000)

Tipo	Descrição	Medida
Albumin		(g/dL)
ALP	*Alkaline phosphatase*	(IU/L)
ALT	*Alanine transaminase*	(IU/L)
AST	*Aspartate transaminase*	(IU/L)
Bilirubin		(mg/dL)
BUN	*Blood urea nitrogen*	(mg/dL)
Cholesterol		(mg/dL)
Creatinine	*Serum creatinine*	(mg/dL)
DiasABP	*Invasive diastolic arterial blood pressure*	(mmHg)
FiO2	*Fractional inspired* O2	(0-1)
GCS	*Glasgow Coma Score*	(3-15)
Glucose	*Serum glucose*	(mg/dL)
HCO3	*Serum bicarbonate*	(mmol/L)
HCT	*Hematocrit*	(%)
HR	*Heart rate*	(bpm)
K	*Serum potassium*	(mEq/L)
Lactate		(mmol/L)
Mg	*Serum magnesium*	(mmol/L)
MAP	*Invasive mean arterial blood pressure*	(mmHg)
MechVent	*Mechanical ventilation respiration*	(0:false, or 1:true)
Na	*Serum sodium*	(mEq/L)
NIDiasABP	*Non-invasive diastolic arterial blood pressure*	(mmHg)
NIMAP	*Non-invasive mean arterial blood pressure*	(mmHg)
NISysABP	*Non-invasive systolic arterial blood pressure*	(mmHg)
PaCO2	*Partial pressure of arterial* CO2	(mmHg)
PaO2	*Partial pressure of arterial* O2	(mmHg)
pH	*Arterial pH*	(0-14)
Platelets		(cells/nL)
RespRate	*Respiration rate*	(bpm)
SaO2	O2 *saturation in hemoglobin*	(%)
SysABP	*Invasive systolic arterial blood pressure*	(mmHg)
Temp	*Temperature*	(°C)

TropI	*Troponin-I*	(mg/L)
TropT	*Troponin-T*	(mg/L)
Urine	*Urine output*	(mL)
WBC	*White blood cell count*	(cells/nL)
Weight		(kg)

O desafio de 2012, foi organizado em duas etapas: 1a - medição da performance do classificador binário; 2a - medição da performance do risco estimado. A pontuação da 1a etapa foi o valor baixo sensitivo e positiva para predição. A pontuação da 2a etapa foi o tamanho normalizado da estatística de Hosmer-Lemeshow (SAEED et. al., 2012).

Os dados utilizados na ICU 2012 foram extraídos do bando de dados clínico MIMIC II na versão 2.6 (SAEED et. al., 2012).

Foram selecionados 12 mil registros aleatoriamente de 12.753 registros com idade superior a 16 anos e com valores de entrada na UTI menor que 48 horas. Foram disponibilizados os registros em 3 agrupamentos (A, B e C) de 4000 pacientes (SAEED et. al., 2012).

No desafio de 2012, disponibilizaram o agrupamento A (treinamento) com as informações de estadia na UTI (treinamento), o agrupamento B (teste) para implementação do método de predição da mortalidade na UTI sem as informações de estadia na UTI e o agrupamento C somente para validar os resultados do método proposto, caso, a solução fosse selecionada para o final do desafio (SAEED et. al., 2012).

Utilizamos o conjunto de treinamento A. Este conjunto, consiste nos registros médicos de 4.000 pacientes, durante a estada na UTI. Para cada paciente foram disponibilizados um arquivo do tipo ".txt" com as informações coletadas, conforme descrito na tabela 3, das primeiras 48 horas do paciente na UTI e as informações gerais do paciente, descrito nas tabelas 1 e 2. O nome do arquivo ".txt" é o identificador do paciente.

Após a análise dos arquivos, verificamos que não seria possível utiliza-los, na forma original, e escolhemos migrar os dados de cada arquivo em 03 tabelas (Paciente, Temporais e Outcomes) com relacionamentos, entre elas, através do ID do paciente.

Um dos problemas relatados no conjunto de dados disponibilizado foi a falta de informação em alguns dados dos pacientes. Para resolver esse problema atribuímos o valor nulo que posteriormente será discretizado como "?".

Para minimizar o ruído nos dados, resolvemos agrupar os dados pelo "id do paciente" e calcular a média aritmética de todas os atributos temporais.

Após a migração dos registros verificamos que a variável "id do paciente" não fornecia nenhuma informação relevante neste estudo, então, preferimos descarta-la. A ideia, inicial, é usar todas as variáveis restantes neste estudo e, deixar o algoritmo de associação tratar as possíveis redundâncias e/ou as variáveis não informativas.

Para definir as regras de associação utilizamos os registros das últimas 24 horas sendo agrupados de 6 em 6 horas e, geramos 4 grupos (G1, G2, G3 e G4) e 2 duas classes (F_sim = Faleceu, F_nao = Sobreviveu).

Para utilizar o algoritmo selecionado é necessário discretizar os dados (FLACH e LACHICHE, 1999). O ideal seria ter o apoio de um especialista nesta tarefa. Entretanto, por não termos um especialista participando deste trabalho, e devido ao tempo disponível para sua conclusão, optamos por utilizar somente os valores referenciados no material disponível no desafio para realizar a tarefa de discretização. Então, cada um dos valores associados aos registros dos pacientes foi mapeado para os valores discretos, conforme informações apresentadas nas regras do desafio. Vejamos os detalhes na tabela 4

Tabela 4: Tabela das discretizações

Parâmetro	Descrição
Albumin, ALP, ALT, AST, Bilirubin, bun, Cholesterol, Creat, FiO2, Glucose, HCO, HCT, HR, K, Lactate, Mg, MAP, Na, NIMAP, PaCO2, PaO2, ph, Patelets, RespRate, SaO2, Temperatura, TroponinT, Urina, WBC, Weight	B: Baixo; N: Normal; A: Alto
DiasABP, NIDiasABP, NISysABP, SysABP	H; D; P; Stage_1; Stage_2; E
GCS	*Severe; Moderate; Minor*
MechVent	*true; false*
TroponinI	N: Normal; aN: Anormal

Por fim, utilizamos a ferramenta Weka (HALL et. al., 2009) juntamente com o algoritmo Tertius (FLACH e LACHICHE, 1999) na geração automática das regras de associação.

O algoritmo Tertius, faz o uso de uma abordagem heurística. Através de um algoritmo tipo "A*", é realizada uma busca no espaço de possibilidades das regras de associação. A heurística utiliza uma distribuição Qui quadrado X2 que fornece uma regra $A \Rightarrow B(\alpha,\beta)$ (FLACH e LACHICHE, 1999).

O objetivo do algoritmo consiste em aplicar um best-first-search, com a abordagem "A*", encontrando as k hipóteses mais confirmadas, procurando regras com múltiplas condições, como o Apriori, e também inclui um refinamento de operadores não redundantes para anular as buscas desnecessárias. Os resultados são as regras estimadas, seguidas pelas observadas e por fim as confirmadas (FLACH e LACHICHE, 1999).

É com base nessas informações que um analista especializado pode realizar suas considerações e efetivar suas próximas estratégias. Ao final, o Tertius descreve a quantidade de hipóteses consideradas e a quantidade explorada (ANDRADE, 2007).

O algoritmo Tertius, demonstrou melhores resultados do que o algoritmo Predictive Apriori (SCHEFFER, 2005).

3 – Resultados e Discussão

Após a análise dos arquivos disponibilizados tivemos que migrar para um banco de dados relacional para facilitar a visualização das informações. Para auxiliar nessa migração foi desenvolvido uma aplicação Web[i] que gerou um arquivo ".arff"[ii] . Este arquivo, contêm 4 grupos com os 37 atributos temporais, o atributo sexo e a classe (Faleceu, Sobreviveu). Vejamos no código 1 o exemplo dos cabeçalhos.

Código 1 Exemplo do cabeçalho do arquivo ARFF

```
1  @ATTRIBUTE Albumin_1 {B,N,A}
2  @ATTRIBUTE ALP_1 {B,N,A}
3  @ATTRIBUTE ALT_1 {B,N,A}
4  @ATTRIBUTE AST_1 {B,N,A}
```