Fusão Bayesiana aplicada a doentes com síndrome da artéria coronária

1st Duarte Emanuel Ramos Meneses
Departamento de Engenharia Informática
Universidade de Coimbra
Coimbra, Portugal
duartemeneses@student.dei.uc.pt

2nd Patrícia Beatriz Silva Costa
Departamento de Engenharia Informática
Universidade de Coimbra
Coimbra, Portugal
patriciacosta@student.dei.uc.pt

***Abstract*—A síndrome da artéria coronária é uma doença grave que afeta milhares de pessoas por todo o mundo e que pode levar ao enfarte do miocárdio. Deste modo, é de extrema importância a sua monitorização e acompanhamento por parte dos médicos para evitar o pior cenário.**

Foi com isso em vista que realizamos este trabalho: através da fusão Bayesiana de vários tipos de informação, tentamos prever a probabilidade de ocorrência de enfarte do miocárdio num espaço temporal de 30 dias.

***Index Terms*—fusão de informação, Bayes, fusão Bayesiana, síndrome da artéria coronária, doença, enfarte do miocárdio**

I. Introdução

A síndrome da artéria coronária é uma doença cardiovascular comum e grave que afeta milhares de pessoas pelo mundo, levando ao enfarte do miocárdio. A precisão do diagnóstico e prognóstico nesta doença é de extrema importância para garantir intervenções médicas adequadas.

Existem inúmeros fatores que podem ajudar a prever a ocorrência de um episódio de enfarte do miocárdio. Combinando toda essa informação, é possível controlar o episódio e melhorar a qualidade de vida dos pacientes.

Dos vários tipos de métodos de fusão de informação, a Bayesiana tem-se mostrado uma abordagem promissora para combinar dados de diferentes fontes e obter estimativas mais confiáveis e precisas. Esta situação pode-se dever ao facto de a regra de Bayes permitir a combinação de dados já conhecidos (*a priori*) com novas medições, quer a informação seja contínua ou discreta.

De um modo geral, a fusão Bayesiana é um método de fusão de dados que utiliza os princípios do teorema de Bayes para juntar informações de múltiplas fontes e gerar uma estimativa *a posteriori* mais robusta. No caso do enfarte do miocárdio, a fusão bayesiana pode ser aplicada para combinar dados clínicos, exames e o historial do doente para obter um diagnóstico mais preciso e proporcionar um prognóstico mais fiável.

A fusão bayesiana oferece a oportunidade de abordar a incerteza dos dados médicos e de obter estimativas mais fiáveis para auxiliar os médicos na tomada de decisões clínicas. Esta abordagem avançada pode contribuir significativamente para a classificação do risco do doente, a adaptação do tratamento e a melhoria dos resultados globais na gestão da síndrome da artéria coronária.

Neste trabalho, iremos explorar a aplicação da fusão bayesiana na síndrome da artéria coronária tendo em conta os dados fornecidos de doentes que tenham sido admitidos na unidade de emergência médica com um episódio de enfarte do miocárdio. O objetivo passa por conseguir prever um novo evento de enfarte nos 30 dias seguintes. Nas próximas secções iremos apresentar um pouco da teoria por detrás deste método, analisar os dados que utilizamos, explicar a implementação, expor os resultados e revelar possíveis tarefas futuras para aprimorar a nossa abordagem.

II. Fundamentação Teórica

A fusão bayesiana é uma abordagem de fusão de dados que se baseia nos princípios do teorema de Bayes que determina a probabilidade de um evento acontecer através de um conhecimento prévio.

Em fusão de dados, o teorema de Bayes, consegue combinar dois tipos de informações distintas.

- O que se sabe sobre um determinado problema (passado, histórico);
- Uma nova observação relacionada com o problema.

Os princípios fundamentais da teoria Bayesiana incluem a definição de uma distribuição de probabilidade inicial, designada por *priori*, que representa o conhecimento prévio ou as hipóteses iniciais sobre o acontecimento em causa. Estas convicções são depois atualizadas através da incorporação de novas evidências, representadas pela verosimilhança dos dados observados. Através do teorema de Bayes, é possível obter a distribuição de probabilidade posterior, que combina a informação inicial com as novas evidências.

O teorema de Bayes pode ser aplicado diretamente para fundir dados provenientes de diferentes fontes e pode combinar informação heterogénea (discreta/contínua). Para tal acontecer, é necessário ter em conta que os diferentes tipos de informação são tratados de diferentes formas.

A. Informação Discreta

No caso da informação discreta, a fusão bayesiana envolve a atualização das probabilidades *a priori* com base nas novas observações. As novas probabilidades podem ser obtidas a partir de dados históricos como estatísticas relevantes.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

- P(A) - o que se sabe antes da observação (*a prior*)
- P(B) - nova observação sobre o problema
- P(B|A) - qual o grau de certeza de B em relação a A (qualidade do sensor)
- P(A|B) - atualizar A dado que B foi observado (*posterior*)

B. Informação Contínua

No caso de dados contínuos, a fusão bayesiana envolve as distribuições de probabilidade das variáveis em questão.

A distribuição gaussiana, sendo uma das mais utilizadas, foi escolhida para o nosso trabalho devido à sua importância no contexto abordado nas aulas.

$$P(B|A) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(x-\mu)^2}{2\sigma^2} \quad (2)$$

Depois disto, a probabilidade *posterior* é calculada da mesma forma que na informação discreta.

III. Dados

De modo a realizar este trabalho, utilizamos um *dataset* disponibilizado no contexto da cadeira de Fusão de Informação em Análise de Dados do Mestrado em Engenharia e Ciência dos Dados da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Esse conjunto de dados apresenta 10 atributos com 457 entradas cada. Não tem nenhum valor em falta e todos os valores são do tipo *float64*. Os atributos são os seguintes:

TABLE I
ATRIBUTOS DO *dataset*

Atributo	Descrição	Tipo
Gender	Género do paciente	Discreto (F=0, M=1)
Age	Idade do paciente	Contínuo (33... 91)
Risk factors	Fatores de risco do paciente tendo em conta o seu histórico clínico, família,..	Discreto (noRisk=0, risk=1)
Systolic Blood pressure	Pressão arterial sistólica	Contínuo (60... 221)
Heart rate (1)	Frequência cardíaca medida através de aparelho BP	Contínuo (40... 152)
ST deviation	Elevação ST (ECG)	Discreto (no=0, yes=1)
Heart rate (2)	Frequência cardíaca medida através de ECG	Contínuo (44... 153)
Creatinine	Níveis de creatinina do paciente	Contínuo (0,6... 11,5)
Killip class	Classe Killip - Exames físicos da capacidade funcional	Discreto (noSigns=1, moderateSigns=2, pulmonaryEdema=3, cardiogenicChock=4
Event	Evento - Ocorrer ou não um enfarte	Discreto (noEvent=0, event=1)

Fica assim evidente que o nosso conjunto de dados está dividido essencialmente em três partes:

- Dados de histórico: género, idade, fatores de risco;
- Dados de medição: pressão arterial sistólica, frequência cardíaca, elevação ST;
- Dados de exames/diagnósticos: creatinina, classe *Killip*.

Depois da leitura dos dados a partir do ficheiro, foi observado que a *feature Heart Rate* é medida por dois sensores.

No entanto, tínhamos que $\sigma HRBP = 2$ e $\sigma HRECG = 0.5$

Logo, para fundir os dados dos dois sensores, foi utilizada a seguinte fórmula:

$$Hr = \frac{\sigma HRECG}{(\sigma HRBP)^2 + \sigma HRECG)^2} * HRBP + \frac{\sigma HRBP}{(\sigma HRBP)^2 + \sigma HRECG)^2} * HRECG$$

Foi também adicionada uma nova *feature "Guidelines"* tendo em conta os valores provenientes de orientações clínicas da creatinina, do segmento ST e do *heart rate*.

- Se $CT \geq 1.3$ e $ST = 1$, guidelines = 1
- Se $KL \geq 2$, guidelines = 1

IV. Implementação

A. Dados discretos

Começamos por implementar fusão Bayesiana nos dados discretos. A fórmula para calcular a probabilidade de ocorrência de enfarte do miocárdio é a seguinte:

$$P(B|A) = \frac{\textit{\# B quando A toma determinado valor}}{\textit{\# A com determinado valor}}$$

Deste modo, fomos vendo a probabilidade de cada valor dos atributos tendo em conta a ocorrência evento. Os resultados foram os seguintes:

TABLE II
PROBABILIDADES DE CADA ATRIBUTO TENDO EM CONTA A OCORRÊNCIA DO EVENTO (A PRIORI)

	Event=0	Event=1
Gender=0	0.2122	0.2123
Gender=1	0.7878	0.7877
RF=0	0.8453	0.8492
RF=1	0.1547	0.1508
ST=0	0.6619	0.1676
ST=1	0.3381	0.8324
KIL=1	0.9748	0.6816
KIL=2	0.0180	0.1453
KIL=3	0.0072	0.1732
GUID=0	0.8669	0.3631
GUID=1	0.1331	0.6369

Analisando os resultados acima, fica claro que existem mais homens no *dataset* que mulheres, uma vez que é o género que apresenta maiores valores em ambos os casos de existir ou não evento.

É também evidente pela tabela II que quando não existe previsão de ocorrência de enfarte, existe maior probabilidade de as *guidelines* estarem a 0. Já se o evento estiver a 1, é mais provável que as orientações clínicas estejam a 1.

Com isto, ficou relativamente simples prever o evento tendo em conta cada atributo. Os resultados foram os seguintes:

TABLE III
PROBABILIDADES DE OCORRÊNCIA DO EVENTO TENDO EM CONTA CADA ATRIBUTO (A POSTERIORI)

	Event=0	**Event=1**
Gender=0	0.1291	0.0832
Gender=1	0.4792	0.3085
RF=0	0.5142	0.3326
RF=1	0.0941	0.0591
ST=0	0.4026	0.0656
ST=1	0.2057	0.3260
KIL=1	0.5930	0.2670
KIL=2	0.0109	0.0569
KIL=3	0.0044	0.0678
GUID=0	0.5273	0.1422
GUID=1	0.0810	0.2495

Tal como era expectável, a probabilidade nos piores cenários em cada atributo é maior no caso de ocorrer enfarte. Inversamente, a probabilidade nos casos menos graves é maior quando não ocorre o evento.

B. Dados contínuos

No caso dos dados contínuos, optamos por utilizar a distribuição gaussiana para calcular a probabilidade *a priori*, tal como já referimos acima. De seguida, calculamos a *posterior* da mesma forma que calculamos nos dados discretos.

Deste modo, apresentamos abaixo os resultados das probabilidades *a posteriori* da ocorrência do evento, tendo em conta cada atributo.

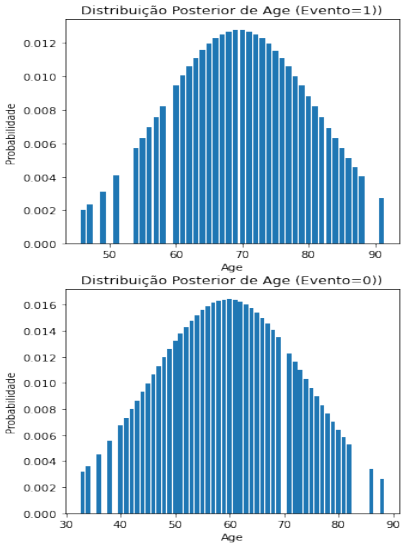

Fig. 1. Probabilidade de ocorrer ou não enfarte tendo em conta a idade

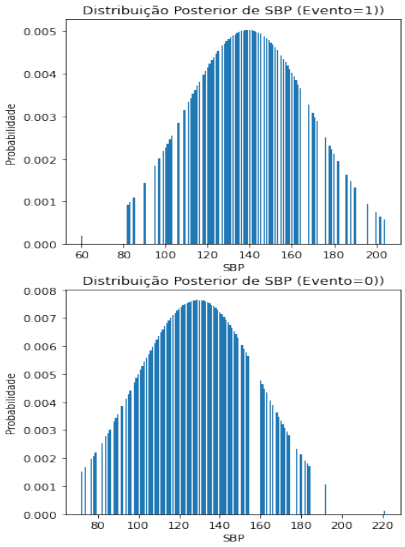

Fig. 2. Probabilidade de ocorrer ou não enfarte tendo em conta a pressão arterial sistólica

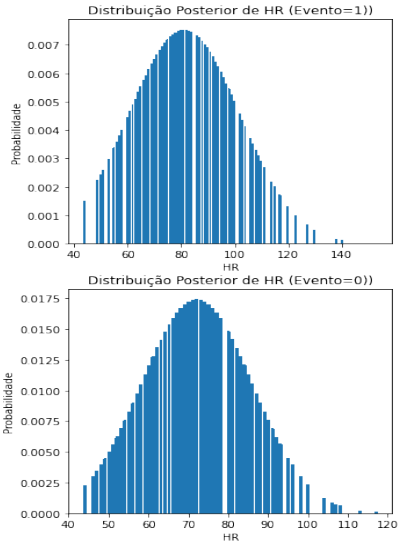

Fig. 3. Probabilidade de ocorrer ou não enfarte tendo em conta a frequência cardíaca

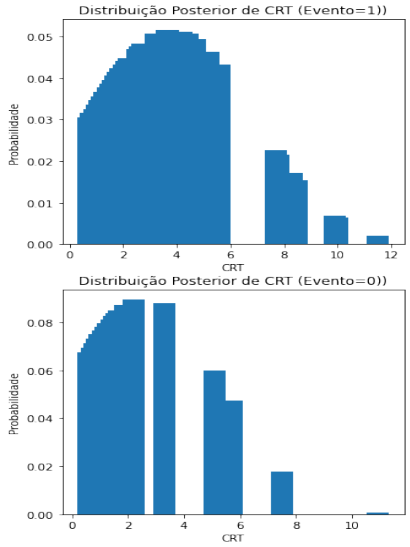

Fig. 4. Probabilidade de ocorrer ou não enfarte tendo em conta os níveis de creatinina

Tal como nos dados discretos, é possível verificar em que valores é mais provável o paciente ter ou não um enfarte do miocárdio.

Por exemplo, na *feature* da pressão arterial sistólica, podemos observar que quem tem os valores perto de 140 tem muito mais probabilidade de ter um enfarte.

C. Fusão final

Para juntar todas as probabilidades calculadas de modo a obter os resultados finais, foi necessário multiplicar todas as Probabilidades P(X|Event), ou seja, as probabilidades *a priori*, e, de seguida, multiplicar por P(Event=0) ou P(Event=1).

$$P(Event|X^n) = P(Event)\prod_{i=1}^{n} P(X|Event) \qquad (3)$$

Os resultados encontram-se na secção seguinte.

V. Resultados

Considerando todos os atributos, conseguimos ter uma melhor previsão da ocorrência ou não de enfarte do miocárdio. Deste modo, calculamos a probabilidade de existir a patologia para todos os valores disponíveis. Como, no total, existem 457 combinações, apresentamos aqui apenas alguns exemplos dos resultados que obtivemos.

TABLE IV
Probabilidades de ocorrência do evento tendo em conta todos os atributos

User	User 01	User 02	User 03	User 04
Gender	1	1	1	0
Age	33	69	63	79
RF	0	0	0	0
SBP	132	147	142	147
HR1	91.5792	52.7859	40.6928	106.4208
ST	1	0	1	1
HR2	90	52	44	110
CRT	0.8	1.4	1.1	0.9
KIL	1	1	3	1
Prob Event=0	1.3366e-08	3.1665e-07	1.0850e-10	8.5978e-10
Prob Event=1	8.9088e-12	1.1240e-08	8.4184e-09	1.0789e-08

Para percebermos se um indivíduo pode sofrer num espaço temporal de 30 dias de enfarte do miocárdio, basta comparar as probabilidades da ocorrência ou não do evento. Por exemplo, para o segundo indivíduo apresentado:

$$3.1665e^{-07}(Ev = 0) > 1.1240e^{-08}(Ev = 1) \qquad (4)$$

Deste modo, é mais provável que não ocorra o evento patológico. No entanto, a probabilidade de ocorrência não é nula, apenas é menor. Mediante o seu valor, os médicos decidem o tipo de tratamento e cuidados necessários a ter.

VI. Trabalho Futuro

Este trabalho foi elaborado no contexto da cadeira de Fusão de Informação em Análise de Dados. Com isto, foi desenvolvido num curto espaço de tempo, pelo que não tivemos tempo para calcular todas as combinações de *features* possíveis.

Desta feita, como trabalho futuro gostaríamos de determinar todas as combinações para estudar melhor os casos que levam a um enfarte do miocárdio.

VII. Conclusão

A síndrome da artéria coronária é uma doença grave que afeta milhares de pessoas por todo o mundo e que pode levar ao enfarte do miocárdio. Deste modo, é de extrema importância a sua monitorização e acompanhamento por parte dos médicos para evitar o pior cenário.

Foi com isso em vista que decidimos realizar este trabalho. Utilizando fusão Bayesiana, combinamos vários tipos de informação para chegar a um prognóstico mais eficaz. Este revelou-se um método extremamente útil para o caso uma vez que permite a fusão de dados de várias fontes, discretos e contínuos, e possibilita a combinação de dados já conhecidos (*a priori*) com novas observações.

Em suma, pensamos ter cumprido o que nos comprometemos ao início: ajudar os pacientes com síndrome da artéria coronária a ter melhor qualidade de vida ao conseguir atuar perante a previsão de ocorrência de enfarte do miocárdio no espaço temporal de 30 dias.

REFERENCES

[1] "Naive Bayes Classifier From Scratch in Python - MachineLearningMastery.com", MachineLearningMastery.com. Disponível: https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/. [Acedido: 28-Maio-2023]

[2] "O que são variáveis categóricas, discretas e contínuas?", Support — Minitab. Disponível: https://support.minitab.com/pt-br/minitab/21/help-and-how-to/statistical-modeling/regression/supporting-topics/basics/what-are-categorical-discrete-and-continuous-variables/. [Acedido: 28-Maio-2023]

[3] "Teorema de Bayes: entenda o que é e de que forma calcular", Blog CAE Treinamentos. Disponível: https://caetreinamentos.com.br/blog/processos/teorema-de-bayes-o-que-e-aplicacoes-e-como-usar-i-cae-treinamentos. [Acedido: 28-Maio-2023]

[4] Saheki A. H. (2005), " Construção de uma rede Bayesiana aplicada ao diagnóstico de doenças cardíacas" [Universidade de São Paulo]. Disponível: https://www.teses.usp.br/teses/disponiveis/3/3132/tde-06042005-203820/publico/ANDRESAHEKI.pdf. [Acedido: 28-Maio-2023]

[5] Material fornecido pelo docente da cadeira,https://ucstudent.uc.pt/bucket/zl360mgr. [Acedido: 28-Maio-2023]