COIMBRA BUSINESS SCHOOL
iscac 100
Politécnico de Coimbra

Método Híbrido de Imputação e Classificação Baseado em Informação Mútua e Reconhecimento dinâmico de semelhantes

Ricardo Filipe Mendes Belo Vicente

(2018073003, iscac17650@alumni.iscac.pt)

Plano de Trabalho para a UC Não Letiva

Mestrado em Análise de Dados e Sistemas de Apoio à Decisão

Orientador: Prof. Doutor Francisco José Nibau Antunes

Coimbra, junho de 2025

Índice

1. INTRODUÇÃO

Quando trabalhamos com dados reais, deparamo-nos frequentemente com dois desafios que parecem simples mas que, na verdade, escondem uma complexidade considerável: o que fazer quando faltam valores nos nossos dados e como classificar novas observações de forma correcta? Estes dois problemas são centrais nas ciências de dados e na aprendizagem automática, e a forma como os abordamos pode fazer toda a diferença nos resultados que obtemos.

Os métodos tradicionais para lidar com valores em falta (*missing values*), tais como a simples substituição pela média ou mediana, acabam por ignorar relações mais subtis entre as variáveis. Isto pode introduzir distorções nos dados ou levar à perda de informação valiosa sobre a sua variabilidade natural (Little & Rubin, 2002). Da mesma forma, algoritmos clássicos de classificação como o *k-Nearest Neighbors* (k-NN) trabalham com parâmetros fixos que nem sempre conseguem captar as nuances da estrutura local dos dados ou a importância relativa de cada variável (Cover & Hart, 1967).

É neste contexto que surge a proposta deste projeto: desenvolver um método híbrido capaz tanto de imputação de valores em falta como de classificação supervisionada, usando para isso conceitos da teoria da informação e desenvolvendo uma forma dinâmica de selecionar instâncias semelhantes. Em vez de "vizinhos" como no k-NN tradicional, preferimos chamar-lhes "amigos" - e há uma boa razão para isso.

Imagine-se um investigador que precisa de compreender um conceito específico de estatística (essa falta de conhecimento é o valor em falta). Seguindo a lógica tradicional do k-NN, primeiro decidiria quantos colegas consultar e depois iria simplesmente bater à porta desse número de colegas cujos gabinetes estão mais próximos, independentemente de serem especialistas em estatística ou não. O método aqui proposto funciona de forma diferente: primeiro identificam-se quais os colegas cuja área de conhecimento engloba estatística (através da informação mútua entre variáveis) e só depois se escolhe, considerando essa informação, aqueles que estão mais acessíveis - evitando, por exemplo, contactar um especialista que esteja noutro país. É por isso que falamos em "amigos": são escolhas intencionais de instâncias que realmente podem ajudar, garantindo resultados contextualmente mais adequados.

Objetivo Geral: Desenvolver e validar um método híbrido para imputação de valores em falta e classificação, baseado em em informação mútua e numa seleção dinâmica de instâncias semelhantes ("amigos").

Objetivos Específicos:

- Criar um método de imputação que use informação mútua para medir as dependências entre variáveis e que escolha dinamicamente os amigos mais adequados para estimar valores em falta;
- Adaptar este método para tarefas de classificação supervisionada, mantendo a lógica de ponderação e seleção dinâmica;
- Testar o desempenho do método tanto em dados simulados como em *benchmarks* na área, comparando-o com os métodos mais avançados atualmente disponíveis;
- Estudar as propriedades estatísticas e computacionais da abordagem proposta.

Este documento está organizado de forma a servir de guia através do projeto. Começamos com uma revisão da literatura e enquadramento teórico (Secção 2), onde fazemos uma brevíssima introdução aos métodos tradicionais de imputação e classificação.

Em seguida, apresentamos a metodologia que pretendemos aplicar (Secção 3), descrevemos os dados que vamos utilizar (Secção 4), detalhamos o cronograma de trabalho (Secção 5) e, por fim, incluímos as referências bibliográficas e um apêndice com detalhes sobre a pesquisa realizada.

2. REVISÃO DE LITERATURA

Quando estamos perante um *dataset* com valores em falta e temos por objectivo desenvolver um classificador, a classificação supervisionada e a imputação podem ser vistas como duas faces da mesma moeda. São tarefas fundamentais que muitas vezes se influenciam mutuamente - afinal, dados incompletos podem comprometer seriamente a qualidade dos nossos modelos preditivos. Neste capítulo, vamos percorrer o estado da arte nestas áreas, começando por uma visão geral e progredindo até aos detalhes da nossa proposta.

2.1 Imputação de Valores em Falta

Lidar com valores em falta é um dos problemas mais comuns - e por vezes frustrantes - na análise de dados. A forma como escolhemos lidar com eles pode ter um impacto tremendo na validade das nossas análises. Vejamos então as diferentes abordagens que foram desenvolvidas ao longo dos anos.

Métodos Estatísticos Simples

Os métodos mais básicos são também os mais intuitivos: substituir os valores em falta pela média, mediana ou moda. Little & Rubin (1987) sistematizaram estas abordagens no seu trabalho pioneiro sobre análise estatística com dados incompletos. Embora sejam computacionalmente rápidos, estes métodos assumem que os valores em falta acontecem completamente ao acaso (o que chamamos *Missing Completely At Random* - MCAR) e têm o problema de subestimar a variabilidade natural dos dados.

Ainda no campo de métodos simples, mas aplicados a séries temporais, é comum usar *forward fill* e *backward fill*, que simplesmente propagam o último valor conhecido ou o próximo valor disponível (Enders, 2010).

Métodos Baseados em Modelos de Regressão

Buck (1960) teve a ideia pioneira de usar regressão para imputar valores em falta, criando modelos preditivos baseados nas variáveis que conhecemos. Esta ideia evoluiu para métodos

mais sofisticados, como a imputação estocástica, que adiciona um toque de aleatoriedade às previsões para manter a variabilidade dos dados mais realista (Little & Rubin, 2002).

Métodos de Imputação Múltipla

Rubin (1987) revolucionou o campo quando propôs a imputação múltipla. A ideia assenta numa solução simples e eficaz: em vez de criar um único conjunto de dados completo, criamos vários, cada um com diferentes valores imputados que refletem a nossa incerteza. Esta abordagem permite capturar a variabilidade inerente ao processo de imputação, reconhecendo que não existe uma única "resposta correta" para os valores em falta. Ao analisar múltiplos conjuntos de dados imputados e depois combinar os resultados, obtemos estimativas mais robustas e intervalos de confiança que incorporam adequadamente a incerteza da imputação.

O método MICE (Multivariate Imputation by Chained Equations), desenvolvido por van Buuren & Groothuis-Oudshoorn (2011), leva esta ideia ainda mais longe com um processo iterativo onde cada variável com valores em falta é modelada tendo em conta todas as outras. O algoritmo funciona através de ciclos sucessivos, onde em cada iteração uma variável é imputada com base num modelo que usa as restantes variáveis como preditores, incluindo aquelas que foram imputadas em iterações anteriores. Este processo continua até convergir para uma solução estável, resultando num método extremamente flexível que pode lidar com diferentes tipos de variáveis e padrões complexos de dados em falta.

Métodos Baseados em Aprendizagem Automática

O k-NN também encontrou o seu lugar na imputação. Troyanskaya et al. (2001) formalizaram esta abordagem para dados de *microarrays*, estimando valores em falta através da média ponderada dos k vizinhos mais próximos.

Stekhoven & Bühlmann (2012) foram mais ambiciosos e propuseram o *missForest*, que usa Random Forests para imputação, mostrando bons resultados com dados mistos e relações não-lineares.

Mais recentemente, as redes neuronais entraram em cena com os *denoising autoencoders* (Vincent et al., 2008), adaptados para imputação por Gondara & Wang (2018).

Métodos Baseados em Decomposição Matricial

Para dados que se organizam naturalmente em matrizes, os métodos de factorização têm mostrado resultados promissores. A ideia fundamental é que muitas matrizes reais possuem uma estrutura subjacente de baixa dimensionalidade - ou seja, as suas linhas e colunas estão correlacionadas de formas que podem ser capturadas por um número reduzido de factores latentes. Candès & Recht (2009) demonstraram algo notável: estas matrizes de baixo posto podem ser recuperadas perfeitamente mesmo quando faltam muitos valores, desde que os dados em falta estejam distribuídos aleatoriamente. Esta descoberta abriu caminho para algoritmos como o Soft-Impute (Mazumder et al., 2010), que utiliza decomposição SVD (*Singular Value Decomposition*) iterativa para preencher os valores em falta explorando precisamente esta estrutura de baixa dimensionalidade.

2.2 Classificação em Ciência de Dados

A classificação supervisionada é talvez a tarefa mais emblemática da aprendizagem automática: queremos treinar um modelo a reconhecer padrões de várias categorias para ser capaz de colocar novas observações nas categorias certas, baseando-se em padrões que aprendeu de exemplos anteriores. Vamos explorar os diferentes paradigmas que surgiram ao longo do tempo.

Métodos Estatísticos Clássicos

Fisher (1936) lançou as bases com a análise discriminante linear (LDA), procurando projeções que maximizam a separação entre classes. Esta abordagem assume que os dados seguem distribuições normais multivariadas com matrizes de covariância iguais (Hastie et al., 2009) e procura o hiperplano que melhor distingue as diferentes categorias, sendo particularmente eficaz quando esta assunção se verifica.

Por sua vez, a regressão logística, desenvolvida por Cox (1958) e popularizada por Hosmer & Lemeshow (1989), adopta uma estratégia diferente ao modelar directamente a probabilidade de pertencer a uma classe usando a função logística. A sua grande vantagem reside na interpretabilidade dos coeficientes, que podem ser entendidos como log-odds ratios, razão pela qual continua a ser uma escolha popular em contextos médicos e sempre que é crucial explicar as decisões do modelo.

Numa abordagem conceptualmente distinta, o classificador Naive Bayes, formalizado por Maron (1961), baseia-se no teorema de Bayes assumindo independência condicional entre as variáveis preditoras. Apesar desta assunção "ingénua" raramente se verificar na prática, o método mantém-se surpreendentemente eficaz e computacionalmente eficiente, destacando-se particularmente na classificação de texto onde a assunção de independência é menos problemática.

Métodos Baseados em Instâncias

O k-Nearest Neighbors (k-NN) tem uma história interessante: foi introduzido por Fix & Hodges (1951) e mais tarde formalizado por Cover & Hart (1967). A ideia é simples - classificar uma observação com base na "votação" dos k vizinhos mais próximos. Mas esta simplicidade esconde questões importantes sobre como medir distâncias e escolher o valor de k, temas explorados por Weinberger & Saul (2009) no contexto da aprendizagem de métricas.

Métodos Baseados em Árvores

As árvores de decisão começaram com o algoritmo ID3 de Quinlan (1986), evoluindo para o C4.5 (Quinlan, 1993) e o CART (Breiman et al., 1984). Estas técnicas dividem o espaço de *features* através de regras que podemos facilmente interpretar. Mas foi Breiman (2001) quem realmente mudou o jogo ao propor as Random Forests, combinando múltiplas árvores com elementos aleatórios para conseguir um desempenho notável e grande robustez.

Métodos de Margem Máxima

As Support Vector Machines (SVM), desenvolvidas por Cortes & Vapnik (1995), abordam o problema de uma forma elegante: procuram o hiperplano que maximiza a margem entre classes, ou seja, a distância mínima entre o hiperplano de separação e os pontos mais próximos de cada classe. O recurso a funções kernel (funções que medem a semelhança entre pontos) constitui uma inovação fundamental, permitindo transformar problemas não-lineares em lineares através de mapeamentos implícitos para espaços de maior dimensão, como demonstrado por Schölkopf et al. (1999). A decisão de classificação baseia-se apenas nos vectores de suporte - os pontos de treino que se encontram exactamente na margem - tornando o método robusto a outliers distantes.

Métodos de Ensemble

Para além das Random Forests, outros métodos de ensemble ganharam destaque. O AdaBoost (Freund & Schapire, 1997) combina classificadores fracos de forma sequencial, ajustando os pesos das observações. O Gradient Boosting, formalizado por Friedman (2001), constrói modelos de forma aditiva, otimizando uma função de perda através de gradiente descendente. Chen & Guestrin (2016) deram-nos o XGBoost, uma implementação otimizada do Gradient Boosting que tem captado cada vez mais interesse na área.

Métodos de Aprendizagem Profunda

As redes neuronais têm raízes profundas, inspiradas nos trabalhos pioneiros de McCulloch & Pitts (1943) e Rosenblatt (1958) sobre o *perceptron* - o primeiro modelo matemático de um neurónio artificial. Após décadas de desenvolvimento, o advento da aprendizagem profunda revolucionou o campo. LeCun et al. (1998) demonstraram como as redes convolucionais conseguem "ver" padrões em imagens tal como o cérebro humano, tornando-se a base do reconhecimento facial e da visão computacional moderna. Paralelamente, Hochreiter & Schmidhuber (1997) resolveram o problema da memória em redes neuronais ao desenvolver as LSTM (Long Short-Term Memory), capazes de "recordar" informação relevante ao longo de sequências temporais - essenciais para processamento de linguagem e previsão de séries temporais. Goodfellow et al. (2016) sintetizam estas e outras inovações no seu manual de referência sobre aprendizagem profunda.

2.3 Integração de Imputação e Classificação

A relação entre imputação e classificação é mais profunda do que pode parecer à primeira vista. A forma como tratamos os valores em falta pode ter um impacto significativo no desempenho dos nossos modelos preditivos, o que tem motivado o desenvolvimento de abordagens que consideram ambos os processos de forma integrada.

Impacto da Imputação no Desempenho de Classificadores

Batista & Monard (2003) foram dos primeiros a estudar sistematicamente como diferentes métodos de imputação afetam o desempenho de classificadores. A sua conclusão foi reveladora:

a escolha do método de imputação pode ser tão importante quanto a escolha do próprio algoritmo de classificação. Donders et al. (2006) mostraram que até métodos simples como a imputação pela média podem introduzir distorções significativas em modelos de regressão logística, afetando as probabilidades que calculamos.

García-Laencina et al. (2010) focaram-se especificamente nas redes neuronais, mostrando como a incerteza se propaga através das camadas e pode amplificar os erros de imputação. Jadhav et al. (2019) confirmaram que métodos de imputação que preservam a estrutura de covariância dos dados originais levam consistentemente a melhores resultados preditivos.

Conceptualmente, qualquer algoritmo de classificação pode ser interpretado como um imputador, na medida em que pode ser utilizado para prever valores ausentes com base em padrões aprendidos nos dados observados. Esta observação reforça a ligação intrínseca entre as duas tarefas e abre caminho a abordagens que as tratam de forma integrada.

Abordagens Sequenciais vs. Integradas

Tradicionalmente, fazemos imputação e classificação em duas etapas separadas. Mas Saar-Tsechansky & Provost (2007) questionaram esta prática, argumentando que a imputação ideal para explorar dados pode não ser a melhor para fazer previsões. Esta observação abriu caminho para métodos que consideram o objetivo final de classificação durante a própria imputação.

Valdiviezo & Van Aelst (2015) propuseram uma abordagem interessante: guiar a imputação pela importância das variáveis para a classificação, usando medidas derivadas de Random Forests. É uma mudança de paradigma em relação aos métodos tradicionais que tratam todas as variáveis de forma igual.

Frameworks de Otimização Conjunta

A ideia de otimizar imputação e classificação simultaneamente tem ganho força. Bertsimas et al. (2021) desenvolveram o framework OptImpute, formulando o problema como otimização conjunta e usando programação inteira mista para encontrar imputações que maximizam o desempenho do classificador. Os resultados reportados são promissores, especialmente quando os padrões de valores em falta são complexos.

Liao et al. (2014) propuseram um método baseado em maximização da expectativa (EM) que alterna entre estimar parâmetros do classificador e imputar valores, convergindo para uma solução que considera ambos os objetivos. Deng et al. (2016) seguiram uma linha similar com otimização de matrizes de baixo posto.

Aplicação de Teoria da Informação

A informação mútua tem emergido como uma ferramenta poderosa para unificar imputação e classificação. Peng et al. (2005) estabeleceram os fundamentos teóricos para seleção de *features* baseada em informação mútua, mostrando propriedades interessantes como a invariância a transformações monotónicas. García-Laencina et al. (2012) estenderam estas ideias para imputação, propondo ponderar a influência de variáveis observadas com base na sua informação mútua com a variável a imputar.

Mohan & Pearl (2021) trouxeram uma perspetiva causal ao problema, argumentando que precisamos considerar o mecanismo que gera os valores em falta. A sua visão sugere que métodos baseados em informação mútua podem capturar dependências relevantes sem precisar de assumir relações causais específicas.

Desafios e Direções Futuras

Integrar imputação e classificação não é tarefa fácil. Do ponto de vista computacional, a otimização conjunta pode ser muito mais complexa que abordagens sequenciais, especialmente com grandes volumes de dados (Josse & Reiter, 2018). Para além disso, teoricamente, precisamos garantir que não ocorre propagação de incerteza da imputação através do classificador para evitar sobreajustamento (Rubin, 1996).

Os desenvolvimentos recentes em aprendizagem profunda têm explorado arquiteturas que lidam nativamente com valores em falta, como as redes MIDA propostas por Gondara & Wang (2018). Estas abordagens prometem uma integração ainda mais profunda entre imputação e classificação.

2.4 Proposta do Método

Este projeto propõe uma abordagem que combine o melhor de dois mundos: a informação mútua e a seleção dinâmica de amigos para imputação e classificação. Inspiramo-nos nos avanços recentes e procuramos ir além das limitações dos métodos tradicionais. Um dos pontos fulcrais deste trabalho - o "fator de confiança" que ajusta dinamicamente o número de amigos – deixa-nos expectantes se, de facto, na prática, consegue melhorar significativamente a adaptabilidade do método a diferentes estruturas de dados.

Fundamentos de Teoria da Informação

A informação mútua (MI) é um conceito fascinante introduzido por Shannon (1948) que quantifica quanta informação uma variável nos dá sobre outra. Matematicamente, para duas variáveis X e Y, definimos:

$$MI(X;Y) = \sum\sum p(x,y)\log\left[\frac{p(x,y)}{p(x)p(y)}\right],$$

onde $p(x,y)$ é a probabilidade conjunta e $p(x)$, $p(y)$ são as probabilidades marginais. O que torna a MI especial é que, ao contrário da correlação de Pearson que só capta relações lineares, ela deteta qualquer tipo de dependência estatística entre variáveis - seja linear, não-linear ou até não-monotónica (Cover & Thomas, 2006). No contexto da imputação, isto permite-nos identificar quais variáveis observadas contêm mais informação sobre uma variável com valores em falta, criando um sistema de ponderação mais informado que os métodos baseados apenas em distâncias.

Aplicação à Seleção Dinâmica de Amigos

A nossa proposta de usar MI para ponderar *features* na seleção de amigos é uma extensão natural dos princípios estabelecidos por Battiti (1994) para seleção de *features*. Calculamos a MI entre cada par de variáveis, construindo uma matriz simétrica que captura toda a estrutura de dependências do *dataset*. Esta matriz torna-se a base para determinar pesos adaptativos que refletem a relevância real de cada variável observada para prever valores ausentes.

Seguidamente esta informação é combinada com um mecanismo de seleção dinâmica do número de amigos. O nosso "fator de confiança", calculado como a razão entre a mediana e o máximo das distâncias ponderadas, permite ao algoritmo adaptar-se automaticamente à densidade local dos dados. Em regiões homogéneas (onde a mediana se aproxima do máximo), o método seleciona mais vizinhos com confiança. Em regiões heterogéneas (onde a mediana é menor que o máximo), o método torna-se mais seletivo e cauteloso. É uma abordagem diferente do k-NN tradicional com k fixo ou dos métodos de raio fixo que não contemplam a estrutura informacional das variáveis.

3. METODOLOGIA A APLICAR

A metodologia deste projeto segue uma estrutura inspirada no ciclo de vida CRISP-DM (*Cross-Industry Standard Process for Data Mining*), mas adaptada às especificidades do método proposto.

1. Compreensão do Problema

O primeiro passo é definir claramente os nossos objetivos. Precisamos de compreender os desafios específicos da imputação e classificação com dados incompletos, e também de que forma podemos contribuir para o actual estado da arte.

2. Compreensão dos Dados

Neste ponto vamos analisar em detalhe as características dos *datasets* que utilizaremos. É crucial entender os diferentes padrões de valores em falta (MCAR, MAR, MNAR) e como a estrutura das variáveis pode influenciar o nosso método.

3. Preparação dos Dados

Esta fase envolve desenvolver procedimentos robustos para gerar dados sintéticos com padrões controlados de valores em falta. Também implementaremos protocolos para introduzir artificialmente valores ausentes em conjuntos de dados completos, seguindo diferentes mecanismos. Para *datasets* reais, aplicaremos os tratamentos padrão que precedem qualquer abordagem de classificação.

4. Modelação

4.1 Desenvolvimento do Algoritmo de Imputação

O nosso algoritmo de imputação assenta em três pilares fundamentais: primeiro, quantificamos as dependências entre variáveis através de informação mútua; segundo, selecionamos adaptativamente as instâncias mais similares; e terceiro, estimamos os valores ausentes de forma ponderada.

Cálculo da Matriz de Informação Mútua

Começamos por construir uma matriz simétrica M onde cada elemento M[i,j] quantifica a informação mútua entre as variáveis i e j. Esta matriz captura todas as dependências estatísticas do *dataset*, incluindo aquelas relações não-lineares que os métodos baseados em correlação não conseguem detetar. Para garantir robustez com diferentes tipos de distribuições, utilizaremos estimadores não-paramétricos conforme proposto por Kraskov et al. (2004).

Sistema de Ponderação Adaptativo

Para cada variável com valores em falta, derivamos um vetor de pesos w a partir da matriz M:

$$w[k] = \frac{M[k,target]}{\Sigma\, M[i,target]}.$$

Estes pesos dizem-nos exatamente quão relevante é cada variável observada para prever a variável-alvo, permitindo ao algoritmo focar-se na informação que realmente importa.

Seleção Dinâmica de Amigos

O número de amigos é determinado dinamicamente através do fator de confiança:

$$fc = \frac{mediana(distâncias)}{máximo(distâncias)}.$$

Este fator reflete a homogeneidade local dos dados. Quando estamos numa região densa (fc próximo de 0), podemos considerar mais amigos com segurança. Em regiões dispersas (fc próximo de 1), precisamos ser mais seletivos. O número final de amigos é calculado como:

$$k = k_{\min + \left(k_{\max} - k_{min}\right)} \times (1 - fc)$$

Processo de Imputação Iterativo

Em casos de variáveis com um número muito elevado de valores em falta, parece-nos teoricamente possível que o algoritmo acima descrito possa não ser capaz de preencher todos

os *missings*, Para prevenir que tal aconteça, o algoritmo termina com um processo de verificação de valores em falta no *dataset* após imputação e, no caso de ainda haver algum, inicia um processo iterativo de preenchimento dos mesmos.

O algoritmo começa por recolher todas as instâncias sem qualquer valor em falta e cria um *"sub"-dataset* com estas; de seguida, gera novo *dataset* apenas com as entradas que ainda têm valores em falta e, instância por instância, vai aplicando todo o método a cada entrada, garantindo assim que não fica qualquer valor em falta no final.

4.2 Extensão para Classificação

Inicialmente pensado apenas como um método de imputação, foi de forma natural que compreendemos que os mesmos princípios fundamentais podem ser também aplicados à classificação supervisionada; mantendo a lógica de ponderação e seleção dinâmica, mas adaptando-a ao contexto preditivo.

Adaptação da Informação Mútua

Para classificação, expandimos a matriz M para incluir a informação mútua entre cada feature e a variável-alvo categórica. Usamos a formulação apropriada para variáveis discretas, quantificando assim a relevância de cada feature para discriminar entre classes.

Votação Ponderada

A estimativa da classe é feita através de votação ponderada entre os amigos selecionados. Cada amigo contribui com um voto ponderado por:

$$v[i] = \frac{1}{\log d[i]}$$

Onde:

- d[i] é a distância ponderada até ao amigo i (quanto mais próximo, maior o peso)

Tratamento Multi-classe

Para problemas com múltiplas classes, implementamos uma estratégia de classificação direta que evita a fragmentação das abordagens *one-vs-all* ou *one-vs-one*. A informação mútua multi-classe é calculada conforme Vinh et al. (2010), preservando as dependências entre todas as classes simultaneamente.

4.3 Integração dos Componentes

A integração entre imputação e classificação é fundamental para a coerência e eficiência do método.

Framework Unificado

Ambos os processos partilham a mesma infraestrutura: a matriz de informação mútua é calculada uma única vez e reutilizada; o mecanismo de seleção de amigos é idêntico, mudando apenas o tratamento final; e as estruturas de dados são otimizadas para evitar redundâncias.

Pipeline Integrado

Desenvolvemos um pipeline que permite treinar o modelo com dados incompletos (fazendo imputação automática durante o treino) e aplicar o modelo a novos dados potencialmente incompletos sem pré-processamento adicional.

Gestão de Complexidade Computacional

Para garantir que o método é escalável, implementamos algumas otimizações: estratégias de cache para evitar recálculos da informação mútua e paralelização onde possível.

Validação Cruzada Adaptada

O processo de validação cruzada é modificado para respeitar a natureza integrada do método. Cada fold preserva a proporção de valores em falta, garantindo uma avaliação realista. Implementamos também uma estratégia de "missing-stratified" cross-validation para testar a robustez sob diferentes padrões de valores em falta.

5. Avaliação e Validação

A avaliação do método proposto é realizada com base numa análise abrangente que contempla tanto a imputação de valores em falta como a qualidade da classificação supervisionada. Para além disso, recorre-se a estratégias de validação específicas que permitem testar a robustez e generalização da abordagem em diferentes contextos e sob condições variadas.

Avaliação - Imputação

No que respeita à imputação, optou-se por um protocolo sistemático que envolve a omissão controlada de valores em variáveis originalmente completas. Essa omissão artificial permite comparar os valores imputados com os reais e calcular métricas de erro como o erro quadrático médio (MSE), o erro absoluto médio (MAE), o erro percentual absoluto médio (MAPE) e a raiz do erro quadrático médio (RMSE). Estas métricas oferecem uma visão quantitativa do desempenho do algoritmo em diferentes escalas.

No entanto, tal como destacado por Lin e Tsai (2020), a avaliação da imputação não se pode cingir à precisão pontual dos valores estimados. É igualmente fundamental verificar se a estrutura estatística dos dados é preservada. Por essa razão, realizam-se análises complementares que incluem a comparação das correlações entre variáveis antes e depois da imputação, testes de Kolmogorov-Smirnov para avaliar a semelhança entre distribuições, e o exame de estatísticas descritivas como média, variância, assimetria e curtose. A capacidade do método para conservar estas propriedades é crucial para garantir que os modelos subsequentes não são construídos sobre bases enviesadas.

A título comparativo, o desempenho do algoritmo é contextualizado através da sua aplicação lado a lado com outros métodos de imputação bem estabelecidos, tais como a imputação pela média ou mediana, abordagens iterativas como o MICE, métodos baseados em distância como o k-NN tradicional, e estratégias baseadas em modelos, como a imputação com Random Forests. Esta comparação permite identificar não apenas a eficácia da nova abordagem, mas também as situações em que se destaca ou revela limitações.

Avaliação - Classificação

Quanto à componente de classificação, a avaliação segue uma linha igualmente rigorosa. Para além da métrica tradicional de accuracy, são consideradas medidas como precisão, recall, F1-score e AUC-ROC, bem como estatísticas específicas como a Kappa de Cohen, particularmente úteis em cenários com classes desbalanceadas. Estes indicadores fornecem uma visão abrangente do comportamento preditivo do modelo, permitindo compreender como este responde perante diferentes distribuições de classes ou desequilíbrios nos dados.

Complementarmente, procede-se à comparação do classificador com algoritmos de referência amplamente utilizados na literatura. O k-NN tradicional serve como linha de base direta, uma vez que o método proposto representa uma extensão conceptualmente próxima. Por outro lado, incluem-se também classificadores como o SVM, útil para problemas com fronteiras de decisão complexas, e modelos ensemble como o Random Forest e o XGBoost, reconhecidos pelo seu bom desempenho em contextos heterogéneos.

Validação

Por fim, a validação do método assume um papel central na demonstração da sua robustez e aplicabilidade real. Para além da tradicional validação cruzada estratificada — que assegura a manutenção da distribuição de classes em cada fold —, é introduzida uma variante adaptada, designada como missing-stratified cross-validation. Nesta abordagem, diferentes padrões de valores em falta são gerados de forma controlada em cada fold, permitindo avaliar a consistência do desempenho em contextos simulados de crescente complexidade.

Adicionalmente, o método é aplicado tanto a conjuntos de dados sintéticos como reais, testando a sua resiliência perante diferentes mecanismos de geração de valores em falta (MCAR, MAR, MNAR). Esta diversidade de cenários visa demonstrar que a eficácia do modelo não depende de pressupostos específicos sobre o padrão de missingness.

Por fim, avalia-se a estabilidade dos resultados entre execuções, através da repetição de experiências com diferentes seeds de randomização, o que permite verificar a consistência dos outputs e reduzir o risco de conclusões baseadas em configurações particulares.

De forma integrada, esta secção procura não apenas quantificar o desempenho do método, mas também garantir que este se mantém sólido em contextos diversos, mantendo a coerência estatística dos dados e apresentando resultados fiáveis e generalizáveis.

6. Implementação

A implementação será feita em Python, aproveitando o ecossistema rico de bibliotecas: NumPy para cálculos numéricos eficientes, pandas para manipulação de dados, e scikit-learn para funcionalidades de aprendizagem automática. O código será modular e bem documentado, garantindo que outros possam reproduzir e estender o nosso trabalho.

A eficiência computacional é uma prioridade, especialmente pensando em aplicações com grandes volumes de dados. Exploraremos oportunidades de paralelização, particularmente no cálculo da informação mútua e no processamento de múltiplas imputações. O objetivo final é encontrar o equilíbrio perfeito entre rigor teórico e viabilidade prática.

4. DADOS

O método será validado usando dois tipos complementares de dados:

1. Dados Simulados

Vamos gerar *datasets* sintéticos onde controlamos completamente os padrões de valores em falta (MCAR, MAR, MNAR). Isto permite-nos avaliar o método em cenários bem definidos onde conhecemos a "verdade". Ferramentas como o make_classification do scikit-learn serão essenciais neste processo.

2. Datasets de Benchmark

Utilizaremos conjuntos de dados públicos bem estabelecidos na comunidade: Iris, Wine e Breast Cancer Wisconsin do UCI Machine Learning Repository. Estes *datasets* são amplamente utilizados em estudos comparativos, permitindo-nos posicionar o nosso método em relação ao estado da arte.

Do ponto de vista ético e legal, os dados simulados não levantam questões por serem sintéticos. Os *datasets* de *benchmark* são públicos e não contêm informações pessoais identificáveis, cumprindo todas as regulamentações de proteção de dados. Como plano de contingência, se houver dificuldades na geração de dados simulados, podemos adaptar *datasets* existentes introduzindo padrões controlados de valores em falta.

5. CRONOGRAMA

Planificação das macro tarefas a realizar no ano letivo 2025/2026 na UC Não Letiva.

Observações: As macro tarefas devem ser compatibilizadas com as etapas da(s) metodologia(s) proposta(s), podendo haver sobreposição entre essas macro tarefas. Todas as UCs Não Letivas (Dissertação/Trabalho de Projeto/Estágio) são de 1600 horas, correspondendo a, sensivelmente, 9 meses de trabalho a tempo inteiro, i.e., 40h/semana. No caso do Estágio, destas 1600 horas, 960 horas são de contacto com a entidade acolhedora e as restantes 640 horas são de estudo e escrita da tese.

			MÊS	09	09	10	10	11	11	12	12	01	01	02	02	03	03	04	04	05	05	06	06	07	07	08	08	09	09	10	10
			QUINZENA	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª	1.ª	2.ª
TAREFA	**STATUS**	**INÍCIO**	**FIM**																												
Documentação para inscrição no 2.º ano	Não iniciado	01/09	30/09																												
Revisão da literatura	Iniciado																														
Desenvolvimento do método de imputação e testes iniciais em dados simulados	Não iniciado																														
Extensão para classificação e ajustes no algoritmo	Não iniciado																														
Avaliação em datasets de benchmark e comparações	Não iniciado																														
Redação da dissertação e preparação de resultados	Não iniciado																														
Contingência / Preparação de publicação científica ou de participação em evento científico	Não iniciado																														

REFERÊNCIAS BIBLIOGRÁFICAS

- Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6), 519-533.
- Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 5(4), 537-550.
- Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2021). From predictive methods to missing data imputation: An optimization approach. Journal of Machine Learning Research, 18(196), 1-39.
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC Press.
- Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society: Series B (Methodological), 22(2), 302-306.
- Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. Foundations of Computational Mathematics, 9(6), 717-772.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.
- Cover, T. M., & Thomas, J. A. (2006). Elements of information theory (2nd ed.). John Wiley & Sons.
- Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215-242.

- Deng, Y., Chang, C., Ido, M. S., & Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. Scientific Reports, 6, 21689.
- Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59(10), 1087-1091.
- Enders, C. K. (2010). Applied missing data analysis. Guilford Press.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179-188.
- Fix, E., & Hodges, J. L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties (Report No. 4). USAF School of Aviation Medicine.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.
- García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. Neural Computing and Applications, 19(2), 263-282.
- García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2012). Classifying patterns with missing values using multi-task learning perceptrons. Expert Systems with Applications, 40(4), 1333-1341.
- Gondara, L., & Wang, K. (2018). MIDA: Multiple imputation using denoising autoencoders. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 260-272). Springer.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
- Hosmer, D. W., & Lemeshow, S. (1989). Applied logistic regression. John Wiley & Sons.
- Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1-11.
- Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence, 33(10), 913-933.
- Josse, J., & Reiter, J. P. (2018). Introduction to the special section on missing data. Statistical Science, 33(2), 139-141.
- Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. Physical Review E, 69(6), 066138.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
- Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. C., & Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: Imputable or not, and how? BMC Bioinformatics, 15, 346.
- Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: A review and analysis of the literature (2006–2017). Artificial Intelligence Review, 53(2), 1487-1534.
- Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. John Wiley & Sons.
- Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). John Wiley & Sons.
- Maron, M. E. (1961). Automatic indexing: An experimental inquiry. Journal of the ACM, 8(3), 404-417.
- Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research, 11, 2287-2322.

- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115-133.
- Mohan, K., & Pearl, J. (2021). Graphical models for processing missing data. Journal of the American Statistical Association, 116(534), 1023-1037.
- Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226-1238.
- Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.
- Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386-408.
- Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91(434), 473-489.
- Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. Journal of Machine Learning Research, 8, 1623-1657.
- Schölkopf, B., Burges, C. J., & Smola, A. J. (Eds.). (1999). Advances in kernel methods: Support vector learning. MIT Press.
- Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379-423.
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525.
- Valdiviezo, H. C., & Van Aelst, S. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. Information Sciences, 311, 163-181.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1-67.

- Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning (pp. 1096-1103). ACM.
- Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research, 11, 2837-2854.
- Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 10, 207-244.

APÊNDICE

A pesquisa bibliográfica que fundamenta este projeto foi conduzida de forma sistemática e abrangente. Utilizámos as principais bases de dados académicas: Google Scholar, Scopus, IEEE Xplore, PubMed, ScienceDirect e ACM Digital Library.

As palavras-chave foram escolhidas de forma a cobrir tanto aspetos gerais como específicos do nosso trabalho: *"classification algorithms", "data imputation methods", "missing data imputation", "mutual information", "k-NN classification"* e *"dynamic nearest neighbors"*. Esta seleção permitiu-nos captar a evolução histórica dos métodos e as tendências mais recentes.

Embora o foco principal tenha sido em artigos publicados nos últimos 5 anos (para garantir que estamos a par dos desenvolvimentos mais recentes), também incluímos trabalhos fundamentais da área, o que nos levou até à década de 40. Esta abordagem deu-nos uma perspetiva histórica robusta, permitindo compreender como os métodos evoluíram e quais os problemas que persistem.

As nossas consultas foram estruturadas para identificar artigos que cobrissem todo o espectro: desde técnicas amplamente utilizadas em ciência de dados até métodos avançados de imputação e classificação, com especial atenção a abordagens adaptativas ou baseadas em teoria da informação.

Os critérios de inclusão passaram por:

- apenas artigos com *peer review*, escritos em inglês ou português;
- não foram excluídos da revisão logo à partida, mas evitamos *case studies* dada a natureza especifica de cada um;
- qualquer artigo cujo titulo ou *abstract* não revelasse ser importante para o tema, foi excluído

Esta pesquisa resultou na identificação de mais de 50 artigos relevantes que fundamentam a revisão de literatura e suportam as decisões metodológicas do projeto. A organização seguiu uma lógica que progride do geral (métodos clássicos de classificação e imputação) para o

particular (técnicas híbridas e adaptativas), facilitando a compreensão da evolução do campo e do posicionamento da nossa proposta.

As referências e respectivos artigos foram organizados utilizando o *software* Zotero.