MECAI 2024 III Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Abordagem baseada na seleção *Wrapper* para identificação da relevância de variáveis e redução de dimensionalidade

Cássio Antonio Andrade[1]
ICMC-USP
Ednaldo José Ferreira [2]
Embrapa Instrumentação / C4AI

1 Introdução

Nos últimos anos, o termo *Big Data* ganhou destaque devido à produção massiva de dados, sendo que a análise desses dados, chamada *Big Data Analytics*, tornou-se uma tendência importante. *Machine Learning* (ML) e *Deep Learning* (DL), subáreas da inteligência artificial (IA), desempenham papéis cruciais, permitindo que sistemas aprendam e melhorem automaticamente, com aplicações em dados estruturados e não estruturados, especialmente em grandes volumes de dados.

Nos últimos anos, entretanto, a ênfase no volume tem cedido ao chamado "*Small Data*", com tônica maior na qualidade ao invés de quantidade. Em vez de produzir e armazenar dados indiscriminadamente, o paradigma e a atenção têm se atido aos conjuntos de dados bem filtrados e de alta qualidade (veracidade), que podem proporcionar *insights* mais precisos e acionáveis [1]. Embora o conceito de *Big Data* ainda se mantenha relevante, muitas empresas e áreas de pesquisa têm lidado com conjuntos de dados menores, restritos, com atualizações menos frequentes e, em geral, estruturados.

Muitos modelos de ML e DL aplicados ao *Small Data* operam como "caixas-pretas", algo que dificulta a explicação das decisões. A chamada *IA Explicável* é demanda recorrente em diversos setores regulados como finanças e saúde. Nesse contexto, seleção de variáveis ou *feature selection* (FS), uma tarefa clássica da estatística, pode ajudar a tornar os modelos mais explicáveis e a resolver problemas no âmbito do *Small Data*.

A FS é uma fase do pré-processamento de dados que visa identificar variáveis relevantes, eliminando as redundantes, irrelevantes e correlacionadas, visando a qualidade das predições e a redução da dimensionalidade [2]. Em geral, FS melhora a capacidade de generalização do modelo

[1] andrade.cassio@hotmail.com
[2] ednaldo.ferreira@embrapa.br

e torna a tarefa de predição mais eficiente, além de diminuir o tempo de treinamento e processamento e a demanda de recursos computacionais. Além disso, auxilia na explicabilidade das decisões dos modelos e possibilita a redução do custos operacionais (coleta de dados) e financeiros.

Há três abordagens distintas de FS: (1) filtragem, que seleciona variáveis com base em métricas estatísticas ou critérios de relevância; (2) *Wrapper* que testa combinações de variáveis, utilizando o próprio modelo de ML para avaliar a performance de subconjunto combinado de variáveis; e (3) embutida (do inglês: *embedded*), quando o subconjunto de variáveis selecionado é parte inerente ao processo de treinamento e às características do algoritmo de ML [3].

Em geral, a abordagem *Wrapper* apresenta desempenho superior às demais uma vez que utiliza o próprio algoritmo de ML e o desempenho do modelo treinado para avaliar a subconjunto de variáveis selecionadas. No entanto, sua principal desvantagem é o custo computacional, sendo impraticável para a grande maioria dos dados reais. Estratégias baseadas na combinação de abordagens de FS têm sido propostas [4]. Em linhas gerais, esses métodos visam a melhoria do desempenho de meta-heurísticas de busca ou combinações com filtros. Contudo, embora melhorem o desempenho, ainda são altamente restritivas para uso geral. Nesse contexto, o objetivo deste trabalho é o desenvolvimento de um novo método para FS de baixo custo computacional capaz de, a partir de estatísticas descritivas do conjunto de dados e das variáveis preditoras (descritores), prever a relevância das variáveis independentes que resulta da execução da FS *Wrapper* em tarefa de classificação.

Não foram encontrados trabalhos que visem a previsão da relevância pautados na abordagem *Wrapper*. Assim, o estudo e o desenvolvimento estão caracterizados por seu ineditismo para solução do problema de seleção.

2 Metodologia

A metodologia proposta foi desenvolvida em dissertação de mestrado do Programa MECAI do ICMC-USP [5]. A tarefa alvo escolhida para a FS foi a classificação utilizando a regressão logística (RL) como modelo de interesse. O método proposto, ilustrado na Figura 1, tem como primeira etapa a criação de subespaços computáveis para bases de dados de grande dimensionalidade, permitindo a execução da FS *Wrapper*. Em seguida, os dados foram divididos em conjuntos de treinamento e teste (processo *Split(CV)*) utilizando a estratégia *leave-one-dataset-out*. Uma base de referência foi estruturada com descritores de variáveis (estatísticas discriminativas de filtros) e descritores dos conjuntos de dados (descritores de contexto). Cada instância da base de referência é constituída, portanto, de métricas do valor discriminativo da variável e estatísticas relativas ao conjunto de dados do qual a variável pertence. Em adição, uma variável-alvo (*target*) informa o rótulo de "relevância" ou "irrelevância", oriundo da execução de FS *Wrapper*, daquela instância na base de referência.

A partir da base de referência, um modelo de *ensemble* criado pelo algoritmo *Histogram-based Gradient Boosting Trees* (HGBT) foi treinado para prever a relevância das variáveis. Os hiperparâmetros do modelo foram otimizados em validação cruzada com o método *leave-one-dataset-out*.

A avaliação de desempenho do método proposto foi realizada em três eixos analíticos. Pri-

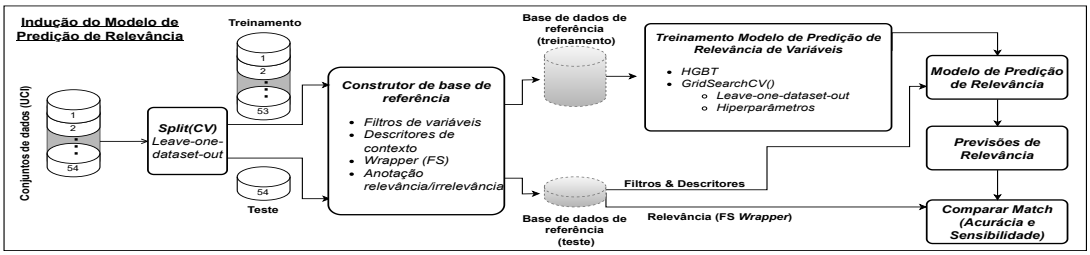

Figura 1: Diagrama do método proposto.

meiro, são analisadas métricas comparativas (acurácia e sensibilidade) da relevância da seleção e redução de dimensionalidade. Segundo, são realizadas comparações pareadas das acurácias obtidas em tarefas de classificação utilizando RL em diferentes cenários: (a) $\mathbf{RL_{Todas}}$ **x** $\mathbf{RL_{FS\ \textit{Wrapper}}}$, (b) $\mathbf{RL_{Todas}}$ **x** $\mathbf{RL_{FS\ Proposta}}$ e (c) $\mathbf{RL_{FS\ \textit{Wrapper}}}$ **x** $\mathbf{RL_{FS\ Proposta}}$

No terceiro eixo de análise, o modelo é aplicado em um conjunto de dados com mais de 250 mil instâncias relacionadas a uma pesquisa sobre diabetes e outro sobre dados financeiros com mais de 60 variáveis. Neste último eixo, o modelo HGBT adotado foi aquele com maior acurácia geral entre todos os modelos obtidos em cada iteração do processo *Split(CV)*.

3 Resultados

A otimização de hiperparâmetros via *GridSearchCV* (Biblioteca Python, Scikit Learn), resultou em uma acurácia média de validação, considerando os modelos otimizados, de 67,38%, sendo a maior acurácia média 69,33%.

O modelo preditivo de relevâncias, ao ser aplicado nos conjuntos reservados para teste, alcançou uma acurácia média de 64% e sensibilidade média de 78%. Portanto, é esperado que 6 em cada 10 variáveis sejam corretamente identificadas como relevantes ou irrelevantes em um novo conjunto de dados e que 7 em cada 10 variáveis relevantes para a abordagem FS *Wrapper* serão apontadas pelo modelo de predição de relevâncias para um novo conjunto de dados.

Ambas abordagens de FS resultaram em uma redução de dimensionalidade superior a 30%, com a abordagem FS *Wrapper* apresentando desempenho ligeiramente superior. No entanto, a diferença média foi de apenas 9%, estatística que evidencia a eficiência do método proposto para redução de dimensionalidade.

As comparações pareadas das acurácias mostraram, no cenário (a), que as acurácias de FS com *Wrapper* tendem a ser ligeiramente maiores e foram estatisticamente significativas em 4 dos 54 conjuntos de dados. Em (b), a RL com todas as variáveis superou significativamente a proposta em 6 conjuntos de dados, sendo superada em um deles. O método proposto manteve desempenho indistinguível do modelo com todas as variáveis em 90,74% das comparativas, considerando a diferença entre ganhos e perdas significativas.

Na configuração comparativa (c), os resultados foram estatisticamente relevantes para 11 dos 54 conjuntos de dados em favor de $\mathbf{RL_{FS\ \textit{Wrapper}}}$. Note-se que não foram constatadas evidências estatísticas para diferenças entre as acurácias em cerca de 80% dos conjuntos de dados.

Finalmente, o último eixo analítico visou as comparações pareadas entre $\mathbf{RL_{FS\ Proposta}}$ e

$\mathbf{RL_{Todas}}$ em conjuntos de dimensionalidades horizontais e/ou verticais consideravelmente maiores, altamente restritivos à abordagem de FS *Wrapper* com busca exaustiva. Não foram observadas diferenças estatisticamente significativas entre as acurácias médias em $\mathbf{RL_{Todas}}$ e $\mathbf{RL_{FS\ Proposta}}$. Por outro lado, o impacto na redução de dimensionalidade foi considerável. Para o conjunto de dados Diabetes, o método proposto selecionou 14 das 21 variáveis, ou seja, operou uma redução de pouco mais de 33% das variáveis do conjunto. Já para o conjunto de dados com informações financeiras, a redução de dimensionalidade foi ainda maior: seleção de 28 variáveis das 64 que compõem o conjunto, ou seja, uma redução de 56,25%.

4 Conclusões

Este trabalho propôs um novo método de FS que emprega um modelo de ML para prever a relevância das variáveis com base em descritores de variáveis e do conjunto de dados. O modelo de predição de relevância é capaz de selecionar, com baixo custo computacional, um subconjunto reduzido de variáveis, melhorando a explicabilidade e reduzindo custos de aquisição do dados inerentes às variáveis removidas.

Em termos de acurácia e sensibilidade, o modelo de predição de relevâncias alcançou 64% e 78%, respectivamente, com redução média de dimensionalidade superior a 30%. Para cerca de 80% dos conjuntos de dados utilizados, não foram constatadas evidências estatísticas significativas entre as acurácias produzidas por FS *Wrapper* e FS Proposta com a RL. Quando aplicado em conjuntos de dados altamente restritivos, o método proposto apresentou desempenho equivalente ao uso de todas as variáveis, com reduções significativas de dimensionalidade.

Futuras pesquisas devem focar na ampliação de descritores de variáveis e de contexto, inclusão de mais conjuntos de dados na base referencial, e validação com outros modelos mais simples. Além disso, a substituição da métrica de acurácia pela medida F, que combina sensibilidade e precisão, pode ser uma alternativa promissora.

Referências

[1] E. Strickland. *Andrew Ng: Unbiggen AI - IEEE Spectrum*. 2022.

[2] S. Visalakshi and V. Radha. A literature review of feature selection techniques and applications: Review of feature selection in data mining. *2014 IEEE International Conference on Computational Intelligence and Computing Research*, p. 1–6. 2014.

[3] B. Venkatesh and J. Anuradha. A review of feature selection and its methods. *Cybernetics and Information Technologies*,v. 19, n. 1, p 3–26 2019.

[4] Maryam and N. A. Setiawan. A wrapper feature selection based on ensemble learning algorithm for high dimensional data. *International Journal of Advanced Trends in Computer Science and Engineering*, v.8, p. 2782-2787. 2019.

[5] C. A. Andrade, Método otimizado para predição da relevância de variáveis e redução de dimensionalidade pautado pela abordagem de seleção Wrapper, Dissertação de Mestrado - MECAI, ICMC, USP, 2024.