Extração de Características em Dados Tabulares

Orientador: Prof. Rodrigo Cesar Pedrosa Silva
Bolsista: Alecia

24 de maio de 2023

Resumo

Este projeto de pesquisa concentra-se no desenvolvimento de técnicas avançadas de extração de atributos adaptadas para dados tabulares. O objetivo é superar os desafios únicos apresentados por conjuntos de dados tabulares com alta dimensionalidade, atributos categóricos e valores ausentes. Ao utilizar conhecimento específico do domínio e incorporar métodos inovadores, como redução de dimensionalidade, codificação de atributos categóricos e imputação de valores ausentes, o projeto tem como objetivo aprimorar a interpretabilidade, robustez e desempenho preditivo de modelos de aprendizado de máquina aplicados a dados tabulares. Serão realizados experimentos extensivos em conjuntos de dados do mundo real para avaliar os métodos propostos e compará-los com técnicas existentes. Os resultados deste projeto têm o potencial de melhorar o desempenho dos modelos e fornecer insights sobre os padrões subjacentes presentes em dados tabulares, contribuindo assim para o campo da ciência de dados e aprendizado de máquina.

Palavras-chave

Extração de Características, aprendizado de máquina, inteligência artificial

1 Introdução

Nos últimos anos, o crescimento exponencial na coleta de dados tem levado ao surgimento de conjuntos de dados tabulares em larga escala em diversos domínios, como finanças, saúde e comércio eletrônico [2]. Dados tabulares geralmente consistem em informações estruturadas organizadas em linhas e colunas, sendo que cada coluna representa um atributo específico. Técnicas eficazes de extração de atributos são cruciais para extrair informações significativas desses conjuntos de dados e melhorar o desempenho de algoritmos de aprendizado de máquina [5].

A extração de atributos tem como objetivo transformar dados tabulares brutos em um conjunto reduzido e representativo de atributos que capturam as características essenciais dos dados [1]. Métodos tradicionais de extração de atributos, como Análise de Componentes Principais (PCA) e Análise de Componentes Independentes (ICA), têm sido amplamente aplicados em domínios de processamento de imagem e sinal. No entanto, os desafios únicos apresentados por dados tabulares, como alta dimensionalidade, atributos categóricos e valores ausentes, requerem técnicas de extração de atributos adaptadas.

Este projeto de pesquisa tem como foco explorar e desenvolver métodos avançados de extração de atributos especificamente projetados para dados tabulares. Ao abordar as limitações das técnicas existentes e incorporar conhecimento específico do domínio, visamos melhorar a interpretabilidade, robustez e desempenho preditivo de modelos de aprendizado de máquina aplicados a conjuntos de dados tabulares.

2 Objetivos

Os principais objetivos deste projeto de pesquisa são os seguintes:

a) Investigar e avaliar as técnicas existentes de extração de atributos para dados tabulares: Este objetivo envolve uma revisão abrangente dos métodos de extração de atributos de última geração adaptados para conjuntos de dados tabulares. Analisaremos os pontos fortes, limitações e pressupostos subjacentes de cada técnica, considerando fatores como seleção de atributos, transformação de atributos e construção de atributos.

b) Desenvolver novos métodos de extração de atributos para dados tabulares: Com base nos insights obtidos na revisão da literatura, proporemos algoritmos inovadores de extração de atributos que abordem os desafios únicos dos dados tabulares. Esses métodos podem incluir técnicas de redução de dimensionalidade, codificação de atributos categóricos, tratamento de valores ausentes e incorporação de conhecimento específico do domínio.

c) Avaliar o desempenho e a interpretabilidade dos métodos propostos: Para avaliar a eficácia das técnicas desenvolvidas de extração de atributos, realizaremos experimentos extensivos em diversos conjuntos de dados tabulares do mundo real. Métricas de desempenho, como acurácia de classificação, erro de regressão e tempo de treinamento do modelo, serão usadas para comparar os métodos propostos com as técnicas existentes. Além disso, analisaremos a interpretabilidade dos atributos extraídos para obter insights sobre os padrões e relacionamentos subjacentes presentes nos dados.

3 Justificativa/Relevância

Este projeto de pesquisa possui uma relevância significativa no campo da ciência de dados e aprendizado de máquina. A extração de atributos informativos de dados tabulares é crucial para obter previsões ou insights precisos e confiáveis

[3, 4, 6]. Os resultados deste projeto terão as seguintes implicações:

a) Melhoria no desempenho do modelo: Ao utilizar técnicas avançadas de extração de atributos, temos como objetivo aprimorar o desempenho preditivo de modelos de aprendizado de máquina aplicados a dados tabulares. Isso permitirá previsões mais precisas, tarefas de classificação e regressão em diversos domínios.

b) Aumento da interpretabilidade: A interpretabilidade dos atributos desempenha um papel vital na compreensão das relações entre as variáveis de entrada e as previsões do modelo. Ao desenvolver métodos que extraem atributos interpretáveis de dados tabulares, este projeto de pesquisa contribuirá para a transparência e confiabilidade dos modelos de aprendizado de máquina.

4 Atividades/Metodologias

O projeto de pesquisa seguirá a seguinte metodologia:

a) Revisão da literatura: Realizar uma revisão abrangente de artigos e pesquisas relacionados à extração de atributos para dados tabulares. Analisar e resumir os pontos fortes, limitações e princípios subjacentes das técnicas existentes.

b) Identificação de desafios: Identificar os desafios únicos enfrentados ao aplicar técnicas de extração de atributos a dados tabulares. Esses desafios podem incluir alta dimensionalidade, atributos categóricos, valores ausentes e dados ruidosos.

c) Desenvolvimento de algoritmos: Com base nos desafios identificados e nos insights obtidos na revisão da literatura, desenvolver algoritmos inovadores de extração de atributos especificamente projetados para dados tabulares. Esses algoritmos podem envolver técnicas como redução de dimensionalidade, codificação de atributos, métodos de imputação para valores ausentes e incorporação de conhecimento específico do domínio.

d) Configuração experimental: Adquirir diversos conjuntos de dados tabulares do mundo real de diferentes domínios para avaliar o desempenho dos métodos propostos de extração de atributos. Pré-processar os conjuntos de dados, abordando questões como limpeza de dados, normalização e tratamento de variáveis categóricas.

e) Avaliação de desempenho: Aplicar os métodos desenvolvidos de extração de atributos aos conjuntos de dados pré-processados e comparar seu desempenho com as técnicas existentes. Avaliar o desempenho usando métricas apropriadas, como acurácia de classificação, erro médio quadrático e eficiência computacional. Realizar testes de significância estatística para validar as melhorias alcançadas pelos métodos propostos.

f) Análise de interpretabilidade: Avaliar a interpretabilidade dos atributos extraídos por meio de análises de importância de atributos, visualizações e técnicas de interpretabilidade do modelo agnósticas, como LIME (Local Interpretable Model-Agnostic Explanations). Analisar os insights obtidos a partir dos atributos interpretáveis e sua concordância com o conhecimento do domínio.

g) Validação experimental: Validar a robustez e generalizabilidade dos métodos propostos de extração de atributos aplicando-os a conjuntos de dados tabulares adicionais de diferentes domínios. Comparar os resultados com métodos de referência e avaliar a consistência do desempenho em diversos conjuntos de dados.

h) Análise comparativa: Realizar uma análise comparativa dos métodos propostos de extração de atributos com as técnicas existentes, destacando seus pontos fortes, limitações e aplicabilidade em diferentes cenários. Identificar os cenários específicos nos quais os métodos propostos superam ou complementam abordagens existentes.

Referências

[1] Isabelle Guyon and André Elisseeff. *An Introduction to Feature Extraction*, pages 1–25. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[2] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[3] Anil Jain and Douglas Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE transactions on pattern analysis and machine intelligence*, 19(2):153–158, 1997.

[4] Alan Jović, Karla Brkić, and Nikola Bogunović. A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1200–1205. Ieee, 2015.

[5] Wamidh K Mutlag, Shaker K Ali, Zahoor M Aydam, and Bahaa H Taher. Feature extraction methods: a review. In *Journal of Physics: Conference Series*, volume 1591, page 012028. IOP Publishing, 2020.

[6] Rizgar Zebari, Adnan Abdulazeez, Diyar Zeebaree, Dilovan Zebari, and Jwan Saeed. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, 1(2):56–70, 2020.