Armand Mounsi - Mathieu Latournerie - Noé Bonne

Fair Classification via Domain Adaptation: A Dual Adversarial Learning Approach

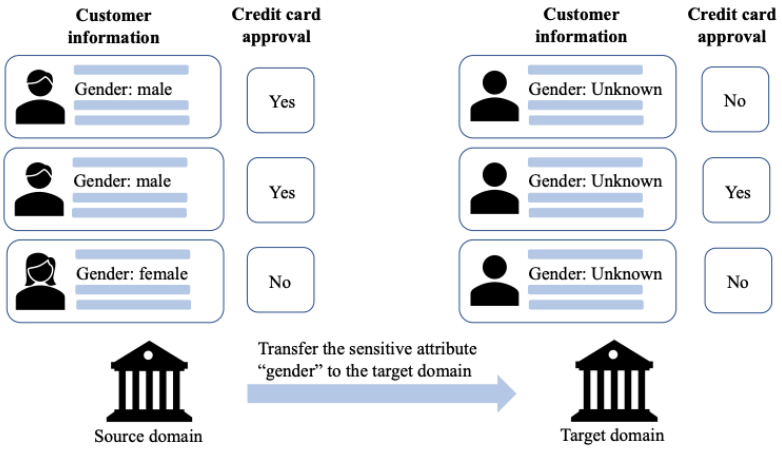
DNN project - paper reimplementation

Abstract:

The FairDA framework proposes an innovative approach to achieve fairness in machine learning predictions, particularly focusing on scenarios where sensitive attributes are not available in the target domain. By leveraging dual adversarial networks, FairDA estimates sensitive attributes in the target domain and ensures fair classification. This methodology not only addresses privacy and legal concerns associated with sensitive attributes but also advances fairness in machine learning applications across diverse domains.

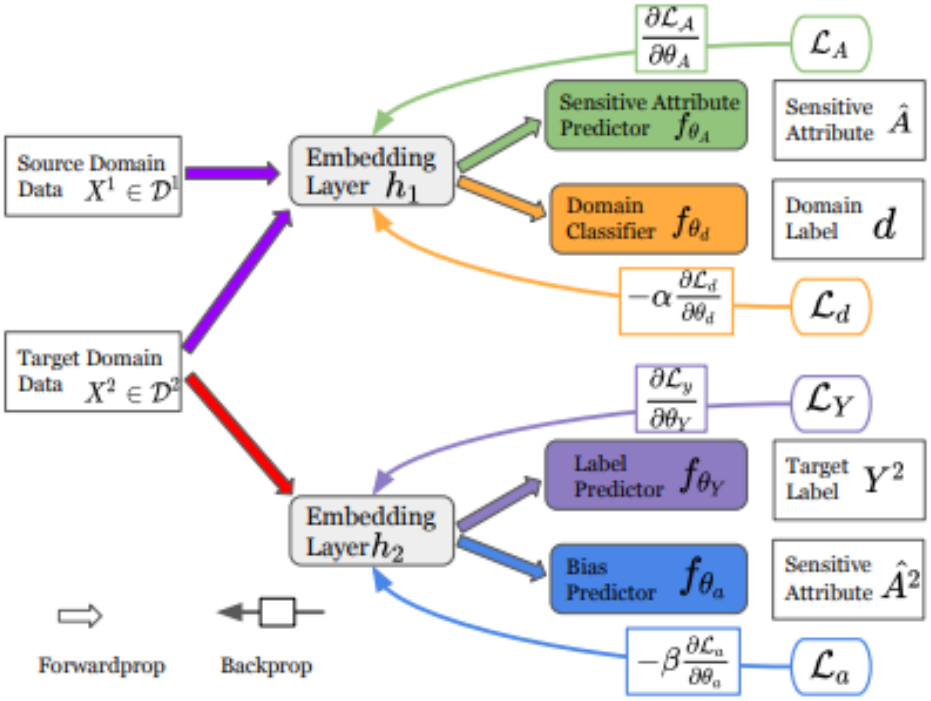
Introduction and Background:

FairDA stands at the intersection of fairness in machine learning and domain adaptation. Traditional approaches often struggle in the absence of sensitive attributes in the target domain, limiting their applicability and fairness. FairDA, through its novel use of adversarial learning, offers a solution that is both practical and effective, as demonstrated across multiple datasets and settings.



Methodology:

The core of FairDA comprises two adversarial components: one for domain adaptation to estimate sensitive attributes in the target domain and another for adversarial debiasing to ensure fair classification. This dual approach allows for the transfer of knowledge from a source domain, where sensitive attributes are known, to a target domain, enhancing prediction fairness without compromising privacy.



Our Reimplementation Efforts:

In our project, we reimplemented the FairDA framework to validate its effectiveness and explore its adaptability. Our implementation focused on closely following the proposed architecture, incorporating the dual adversarial networks, and tailoring the training process to our datasets. We learned a lot about custom losses and backpropagation in pytorch.

Results and Discussion:

Our results mirrored the promising outcomes presented in the paper, with our reimplementation achieving comparable fairness metrics across several datasets. Our work underscores the viability of FairDA in real-world applications and its potential as a cornerstone for future fairness-oriented machine learning research.

Conclusion:

The FairDA framework represents a significant step forward in the pursuit of fairness in machine learning. Our reimplementation efforts have not only validated the original findings but also contributed to the ongoing dialogue on ethical AI practices. By addressing the challenges associated with sensitive attributes and domain adaptation, FairDA paves the way for more equitable machine learning applications.