Advancing Adversarial Attacks in Tabular Machine Learning: A Deep Dive into CAA

Mohammad Sadat Hossain, Nafis Tahmid, Shattik Islam Rhythm

January 4, 2025

Introduction

In the rapidly evolving landscape of machine learning security, adversarial attacks have predominantly focused on computer vision and natural language processing. However, a significant portion of real-world machine learning applications actually process tabular data, especially in critical domains like finance, healthcare, and cybersecurity. The paper *"Constrained Adaptive Attack: Effective Adversarial Attack Against Deep Neural Networks for Tabular Data"* addresses this crucial gap by introducing novel approaches to generate adversarial examples for tabular data while respecting real-world constraints.

The Challenge of Tabular Adversarial Attacks

Unlike images or text, tabular data comes with inherent constraints that make traditional adversarial attack methods ineffective. For instance, in a financial dataset, features like "total debt" and "monthly payment" must maintain specific mathematical relationships. Similarly, categorical features like "education level" can't be arbitrarily modified to continuous values. These constraints make generating valid adversarial examples particularly challenging.

The authors identify several key limitations in existing approaches:

- Most existing attacks ignore feature relationships.
- Current methods either don't handle categorical features or can't process mixed data types.
- Available attacks like CPGD (Constrained Projected Gradient Descent) show low success rates.
- Search-based methods like MOEVA, while effective, are computationally expensive.

Technical Innovation: CAPGD

The first major contribution is **CAPGD (Constrained Adaptive PGD)**, a gradient-based attack that introduces several innovative mechanisms:

Adaptive Step Size

Instead of using a fixed step size decay schedule, CAPGD adaptively adjusts the step size based on optimization progress. The step size is halved when either:

- The loss hasn't increased in 75% of steps since the last checkpoint.
- The maximum loss hasn't improved since the last checkpoint.

This is mathematically represented as:

$$\eta^{k+1} = \begin{cases} \eta^k, & \text{if progress is good} \\ \frac{\eta^k}{2}, & \text{otherwise.} \end{cases}$$

Momentum Integration

CAPGD incorporates momentum to improve optimization stability:

$$z^{k+1} = P_S\left(x^k + \eta^k \nabla L'(x^k)\right)$$

$$x^{k+1} = R_\Omega\left(P_S\left(x^k + \alpha\left(z^{k+1} - x^k\right) + (1-\alpha)\left(x^k - x^{k-1}\right)\right)\right),$$

where $\alpha = 0.75$ balances between the current gradient and the previous update.

Repair Operator

A novel repair operator R_Ω projects examples back into the valid data space after each iteration, ensuring constraint satisfaction throughout the optimization process.

Formulation of Constraints

CAPGD and CAA handle adversarial attacks in tabular data by incorporating domain-specific constraints that preserve the validity of adversarial examples. These constraints are modeled using a structured grammar, ensuring the perturbed data adheres to real-world requirements.

Types of Constraints

1. **Immutability** Certain features cannot be modified. *Example:* In a financial dataset, "loan ID" or "account number" cannot be altered.

2. **Boundaries** Features must remain within specific ranges. *Example:* In a credit scoring dataset, "loan amount" must remain within $[5000, 100000]$:

$$5000 \leq \text{Loan Amount} \leq 100000$$

3. **Type** Features must retain their data type (categorical, discrete, or continuous). *Example:* "Education Level" in a dataset should stay categorical (e.g., {High School, Bachelor's, Master's}).
4. **Feature Relationships** Mathematical or logical relationships between features must be preserved. *Example (Financial Dataset):*

$$\text{Total Debt} \geq \text{Monthly Payments}$$

Example (Phishing Dataset):

$$\text{Length of Hostname} \leq \text{Length of URL}$$

Constraint Grammar

Constraints are formally defined using the following grammar:

$$\omega := \omega_1 \wedge \omega_2 \mid \omega_1 \vee \omega_2 \mid \psi_1 \geq \psi_2 \mid f \in \{\psi_1, \ldots, \psi_k\}$$

Where:

- ω: A constraint.
- ψ: A numeric expression.
- f: A feature.
- $\in$: Denotes membership in a set.

These constraints are enforced during optimization using the repair operator R_Ω, ensuring the generated adversarial examples respect all specified rules.

The Power of Ensemble: CAA

The second major contribution is **CAA (Constrained Adaptive Attack)**, which cleverly combines CAPGD with MOEVA. The key insight is that gradient-based attacks are faster but less successful, while search-based attacks are more effective but slower. CAA applies them sequentially:

1. First attempt: CAPGD for quick wins.
2. If unsuccessful: MOEVA for harder cases.

This simple yet effective combination achieves:

- Up to 96.1% decrease in model accuracy.
- 5x faster than pure MOEVA.
- Best performance in 19 out of 20 experimental settings.

Experimental Validation

The authors conducted extensive experiments across:

- **Datasets:** URL (phishing detection), LCLD (credit scoring), CTU (botnet detection), WiDS (medical).
- **Architectures:** TabTransformer, RLN, VIME, STG, TabNet.

Key findings include:

- CAPGD subsumes all other gradient-based attacks.
- CAA maintains effectiveness while significantly reducing computational cost.
- Adversarial training shows varying effectiveness across architectures.

Future Directions

The paper opens several promising research directions:

- **Defense Mechanisms:** Development of specific defenses against constrained adversarial attacks in tabular domains.
- **Constraint Modeling:** More sophisticated approaches to handling complex feature relationships.
- **Architecture Development:** Design of inherently robust architectures for tabular data.
- **Efficiency Improvements:** Further optimization of search-based components in CAA.

Conclusion

This work represents a significant advancement in adversarial machine learning for tabular data. By introducing CAPGD and CAA, the authors have not only created more effective attacks but also established a new baseline for evaluating the robustness of tabular machine learning models. The paper's contributions are particularly valuable given the prevalence of tabular data in critical applications.

Most importantly, this work highlights the unique challenges of generating adversarial examples for tabular data and provides a framework for future research in both attack and defense mechanisms. As machine learning continues to be deployed in sensitive domains, understanding and addressing these vulnerabilities becomes increasingly crucial.