Состязательные атаки на нейронные сети для работы с временными рядами

Владимиров Эдуард
vladimirov.ea@phystech.edu

Зайцев Алексей
a.zaytsev@skoltech.ru

16 декабря 2023 г.

Аннотация

Существует проблема применения состязательных атак в домене временных рядов, и она заключается в том, что эти атаки очень легко обнаружены. В качестве решения этой задачи предлагается использование различных регуляризаторов, которые обеспечивают сохранение свойств исходного временного ряда. На текущий момент рассмотрен аналог L2-регуляризации. Проведён вычислительный эксперимент с моделями из семейства TS2Vec и с различными датасетами, в котором показано существенное увеличение скрытности атаки.

Ключевые слова: временной ряд · состязательная атака · IFGSM

1 Введение

В работе [1] рассмотрено множество модификаций для IFGSM: добавление случайного шума, включение инерции по аналогии с Nesterov Momentum. В работе [2] используется регуляризатор гладкости. В будущем стоит рассмотреть и другие регуляризаторы: на периодичность, на размерность вложения.

2 Теоретическая часть

2.1 IFGSM

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \varepsilon \cdot sign\big(\nabla_x \mathcal{L}(\mathbf{f}_\theta(\mathbf{x}_t), \mathbf{y})\big)$$

$$\begin{aligned}
\mathbf{h}_{t+1} &= \mathbf{x}_t + \varepsilon \cdot sign\big(\nabla_x \mathcal{L}(\mathbf{f}_\theta(\mathbf{x}_t), \mathbf{y})\big) \\
\Delta_{t+1} &= ||\mathbf{x}_0 - \mathbf{h}_{t+1}|| \\
\mathbf{x}_{t+1} &= \mathbf{x}_t + \varepsilon_{\max} clip\big(\nabla_x \mathcal{L}(\mathbf{f}_\theta(\mathbf{x}_t), \mathbf{y}), -\exp(-\Delta_{t+1}^2), \exp(-\Delta_{t+1}^2)\big)
\end{aligned}$$

3 Постановка задачи

Пусть $\mathbf{f}_\theta : \mathbb{R}^{E \times T} \longrightarrow [0, 1]$ — обученный классификатор временных рядов, $\mathbf{g}_\varkappa : \mathbb{R}^{E \times T} \longrightarrow [0, 1]$ — обученный дискриминатор, выдающий вероятность искажения данных. Тогда задача поиска

4 Вычислительный эксперимент

Таблица 1: Сравнение ванильного и улучшенного IFGSM

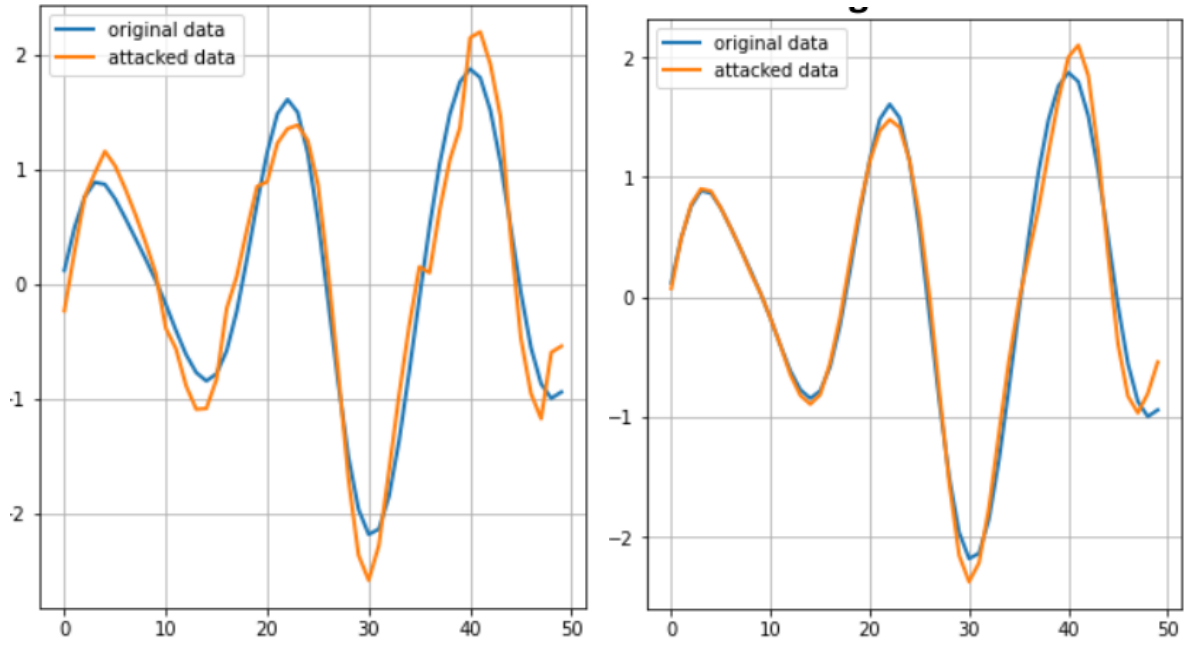
Attack	Dataset	Coffee	FordA
	Target model	TS2Vec	TS2Vec
Vanilla IFGSM	Effectiveness	**1.00**	**1.00**
	Concealability	0.08	0.28
Modified IFGSM	Effectiveness	**1.00**	0.99
	Concealability	**0.97**	**0.92**

5 Заключение

TODO

Список литературы

[1] BS Vivek and R Venkatesh Babu. Regularizers for single-step adversarial training. *arXiv preprint arXiv:2002.00614*, 2020.



(a) Vanilla IFGSM (b) Modified IFGSM

Рис. 1: Визуализация состязательных атак

[2] Gautier Pialla, Hassan Ismail Fawaz, Maxime Devanne, Jonathan Weber, Lhassane Idoumghar, Pierre-Alain Muller, Christoph Bergmeir, Daniel F Schmidt, Geoffrey I Webb, and Germain Forestier. Time series adversarial attacks: an investigation of smooth perturbations and defense approaches. *International Journal of Data Science and Analytics*, pages 1–11, 2023.