Рецензия на рукопись
GENADA: generative time series adversarial attack framework
Автор: Михаил Баронов

Актуальность работы оправдана широким применением методов глубокого обучения для анализа временных рядов в широком классе задач. При этом нейронные сети, несмотря на высокое качество классификации, остаются уязвимыми к adversarial-атакам.

Существующие градиентные методы атак, такие как FGSM, iFGSM и PGD вычислительно затратны, особенно при большом числе атакуемых объектов. Кроме того, в реальных сценариях доступ к градиентам целевой модели может быть ограничен или полностью отсутствовать. В связи с этим разработка эффективных генеративных методов adversarial-атак для временных рядов, способных быстро порождать возмущения без вычисления градиентов на этапе инференса, является актуальной и практически значимой задачей.

Научная новизна работы заключается в предложении фреймворка **GENADA** - генеративного подхода к построению adversarial-атак на модели классификации временных рядов. В отличие от классических градиентных атак, где возмущение подбирается заново для каждого объекта с помощью итеративной оптимизации, в данной работе обучается отдельная нейронная сеть-генератор, которая напрямую порождает допустимое adversarial-возмущение. После обучения такая модель генерирует атаку за один прямой проход, что существенно снижает вычислительные затраты на этапе применения.

Дополнительной новизной является рассмотрение как одношаговой, так и итеративной версии генеративной атаки, что позволяет сопоставить предложенный подход с классическими методами FGSM и iFGSM. В работе также используется схема обучения генератора относительно замороженной целевой модели, благодаря чему на этапе инференса не требуется доступ к градиентам атакуемого классификатора. Это делает метод потенциально применимым в условиях ограниченного доступа к модели.

Практическая значимость и новизна подтверждаются экспериментами на нескольких наборах временных рядов - PowerCons, GunPoint и Strawberry - а также на разных архитектурах моделей, включая LSTM, ResCNN и PatchTST. Полученные результаты показывают, что GENADA может достигать конкурентного качества атаки по сравнению с сильными white-box базовыми методами, при этом обеспечивая существенное ускорение генерации adversarial-примеров. Таким образом, работа вносит вклад в развитие более эффективных и масштабируемых методов оценки устойчивости моделей классификации временных рядов.

Major problems:

1. Почему-то почти нет нормальных численных результатов. В тексте написано, что GENADA сравним с iFGSM и быстрее в 10–50 раз, но таблиц с результатами почти нет.
2. Непонятно сколько раз запускались эксперименты. Нет ошибок, стандартных отклонений, доверительных интервалов.
3. С воспроизводимостью есть проблемы. Архитектуры толком не описаны: не

хватает hidden size, числа слоев, dropout, scheduler, train/validation split и т.д. GitHub тоже указан как anonymized repository, так что проверить результаты сейчас нельзя.

Minor problems:

1. В Main Results почти нет самих результатов, только текстовые утверждения. Лучше добавить таблицы и графики.
2. Обозначения плавают.

Итог:
Идея работы интересная, особенно часть про быстрый генератор атак после обучения. Но текущая версия статьи выглядит сырой. Нужно добавить нормальные таблицы с результатами, ошибки по экспериментам, подробнее описать настройку экспериментов и поправить теорию.

Рецензент:
Кабиров Р.А.