**1.3 Методика тестирования состязательных атак на модели машинного обучения для решения задач time series forecasting и anomy detection**

Методика предназначена для исследования зависимости эффективности проведения состязательных атак на модели машинного обучения (МО) от параметров атак и моделей МО.

Структурная схема методики представлена на рисунке 1.

|  |
| --- |
| D:\Наука\Лаборатория\На поездку\1.3\1 Методика.jpg |
| Рисунок 1. – Структурная схема методики тестирования состязательных атак на модели машинного обучения для решения задач time series forecasting и anomy detection. |

На **первом этапе** реализации методики осуществляется определение исходных данных. Исходными данными являются:

исследуемый временной ряд – *T*;

множество моделей МО для решения задач TSF и TSAD – *M*;

множество видов состязательных атак на модели МО – *A*;

показатель оценивания эффективности состязательной атаки на модели МО – *I*;

критерий оценивания эффективности состязательной атаки – *c*.

Для выбранного временного ряда (одномерный или многомерный) осуществляется выбор и обучение **моделей машинного обучения**.

А) Для решения задач TSF:

модель с долговременной краткосрочной памятью – LSTM,

модель сверточных нейронных сетей – CNN,

комбинированная модель – CNN+LSTM.

Помимо этого, целесообразно рассмотреть статистические модели:

интегрированная модель авторегрессионного скользящего среднего – ARIMA(p, d, q),

сезонная интегрированная модель авторегрессионного скользящего среднего – SARIMA(X)(p,d,q)(P,D,Q)m,

модель векторной авторегрессии – VAR.

Б) Для решения задач TSAD:

модель автоэнкодера – AE,

модель сверточного автоэнкодера – ConvAE,

модель с долговременной краткосрочной памятью – LSTM,

модель автоэнкодера с долговременной краткосрочной памятью – LSTM-AE,

модель вариационного автоэнкодера с долговременной краткосрочной памятью – LSTM-VAE,

В качестве **исследуемых атак** целесообразно рассматривать следующие реализации состязательных атак для временных рядов:

FGSM (Fast Gradient Sign Method);

I-FGSM (Iterative Fast Gradient Sign Method);

MI-FGSM (Momentum Iterative Fast Gradient Sign Method);

JSMA (Jacobian-based Saliency Map Attack);

Метод Карлини и Вагнера (C&W);

DeepFool;

L-BFGS (Limited-memory Broydewn-Fletcher-Goldfarb-Shanno);

PGD (Projected Gradient Descent);

L-BFGS (Limited-memory Broydewn-Fletcher-Goldfarb-Shanno);

UAP (Universal Adversarial Perturbations);

LogBarrier;

ZOO (Zeroth Order Optimization);

Для каждой состязательной атаки определяется множество параметров ее проведения – *Pa*.

В качестве **показателей оценивания эффективности** состязательной атаки на модели МО возможно рассмотреть:

А) При решении задач TSF:

*Величина изменения среднеквадратической ошибки* () рассчитывается в соответствии выражению

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | (1) |
| где: |  | – | среднеквадратическая ошибка прогнозирования значений временного ряда при проведении состязательной атаки на исходный временной ряд; | |
|  |  | – | среднеквадратическая ошибка прогнозирования значений временного ряда при отсутствии воздействия на исходный временной ряд. | |

Среднеквадратическая ошибка (mean squared error – MSE) рассчитывается в соответствии выражению

|  |  |
| --- | --- |
|  | (2) |

Б) При решении задач TSAD:

*Величина изменения средней гармонической классификации* () рассчитывается в соответствии выражению

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | (3) |
| где: |  | – | среднее гармоническое классификации при отсутствии воздействия на исходный временной ряд. | |
|  |  | – | среднее гармоническое классификации при проведении состязательной атаки на исходный временной ряд; | |

Среднее гармоническое классификации (F1-score) рассчитывается в соответствии выражению

|  |  |
| --- | --- |
|  | (4) |

Доля объектов, названных классификатором положительными и при этом действительно являющимися положительными (Precision).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | (5) |
| где: | TP | – | правильная положительная классификация; | |
|  | *TN* | – | правильная отрицательная классификация; | |
|  | *FP* | – | неправильная положительная классификация; | |
|  | *FN* | – | неправильная отрицательная классификация; | |

Доля объектов, найденных классификатором из всех объектов положительного класса (Recall).

|  |  |
| --- | --- |
|  | (6) |

Выбор критерия выбора предпочтительного вида и параметров состязательной атаки осуществляется в соответствии выражению

|  |  |
| --- | --- |
| *.* | (7) |

На **втором этапе** реализации методики производится расчет параметров воздействия на временной ряд.

А) Расчет порядка дифференцирования временного ряда.

Расчет порядка дифференцирования временного ряда осуществляется в соответствии алгоритму, представленному на рисунке 2.

|  |
| --- |
| D:\TSL\Отчеты\23.11\1.3\2 Алгоритм.jpg |
| Рисунок 2. – Алгоритм расчет порядка дифференцирования временного ряда. |

Исходными данными для реализации алгоритма является исследуемый временной ряд. Переменной d присваивается начальное значение равное нулю.

Производится расчет среднего значения – *M*{*yt*} временного ряда и его дисперсии – *D*{*yt*}. Если данные величины не являются постоянными на всем протяжении временного ряда, то это является основанием полагать, что ряд не является стационарным. Для приведения его к стационарному виду производится его дифференцирование, значение переменной d увеличивается на единицу.

Для полученного в ходе дифференцирования временного ряда производится расчет среднего значения и дисперсии, производится проверка условий: *M*{*yt*}=const, *D*{*yt*}=const. При выполнении данного требования алгоритм завершает работу, выходными данными алгоритма является значение d.

Б) Расчет значимых коэффициентов автокорреляции.

Для расчета значимых коэффициентов автокорреляции производится дифференцирование исходного временного ряда в степени d. Далее производится построение графика автокорреляционной функции (АКФ) с лагом равным длине временной последовательности (Рисунок 3). По графику производится вычисление временного шага – *q*, после которого коэффициенты АКФ становятся незначимыми (не выходят за пределы заштрихованной области). Значение *q* определяет количество предыдущих элементов временного ряда оказывающих влияние на текущее в режиме среднего скользящего значения.

|  |
| --- |
|  |
| Рисунок 3. – Пример графика автокорреляционной функции временного ряда. |

Б) Расчет выборочного коэффициента автокорреляции.

Расчет выборочного коэффициента автокорреляции производится аналогично расчету значимых коэффициентов автокорреляции. На начальном этапе производится дифференцирование исходного временного ряда в степени d. Далее производится построение графика выборочной автокорреляционной функции (ВАКФ) с лагом равным длине временной последовательности (Рисунок 4). По графику производится вычисление временного шага – *p*, после которого коэффициенты ВАКФ становятся незначимыми (не выходят за пределы заштрихованной области). Значение q определяет количество предыдущих элементов временного ряда оказывающих влияние на текущее в режиме авторегрессионного окна.

|  |
| --- |
|  |
| Рисунок 4. – Пример графика выборочной автокорреляционной функции временного ряда. |

На **третьем этапе** методики производится непосредственное оценивание эффективности реализации состязательных атак на модели МО. Оценивание производится последовательно для каждой совокупности элементов множеств моделей МО, видов состязательных атак и параметров состязательной атаки по множеству показателей *I*.

|  |  |
| --- | --- |
| . | (1) |