конференция
РусКрипто

Ежегодная международная научно-практическая конференция

«РусКрипто'2023»

Анализ безопасности проекта национального стандарта «Нейросетевые алгоритмы в защищенном исполнении. Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации»

Маршалко Г.Б., ТК26
Романенков Р.А., ТК26
Труфанова Ю.А., ТК26

Угрозы информационной безопасности при использовании методов машинного обучения:

- угрозы нарушения конфиденциальности данных (извлечение данных о параметрах обученных моделей, извлечение данных об обучающей выборке из обученных моделей);
- угрозы нарушения доступности данных (искажение («отравление») обучающей выборки с целью ухудшения качества модели);
- угрозы нарушения целостности данных (формирование т.н. состязательных входных данных, некорректно обрабатываемых (например, классифицируемых) моделью).

Проект национального стандарта «Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации»

В 2022 году в Техническом комитете по стандартизации «Искусственный интеллект» (ТК 164) Омским государственным техническим университетом разработан проект национального стандарта «Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации»

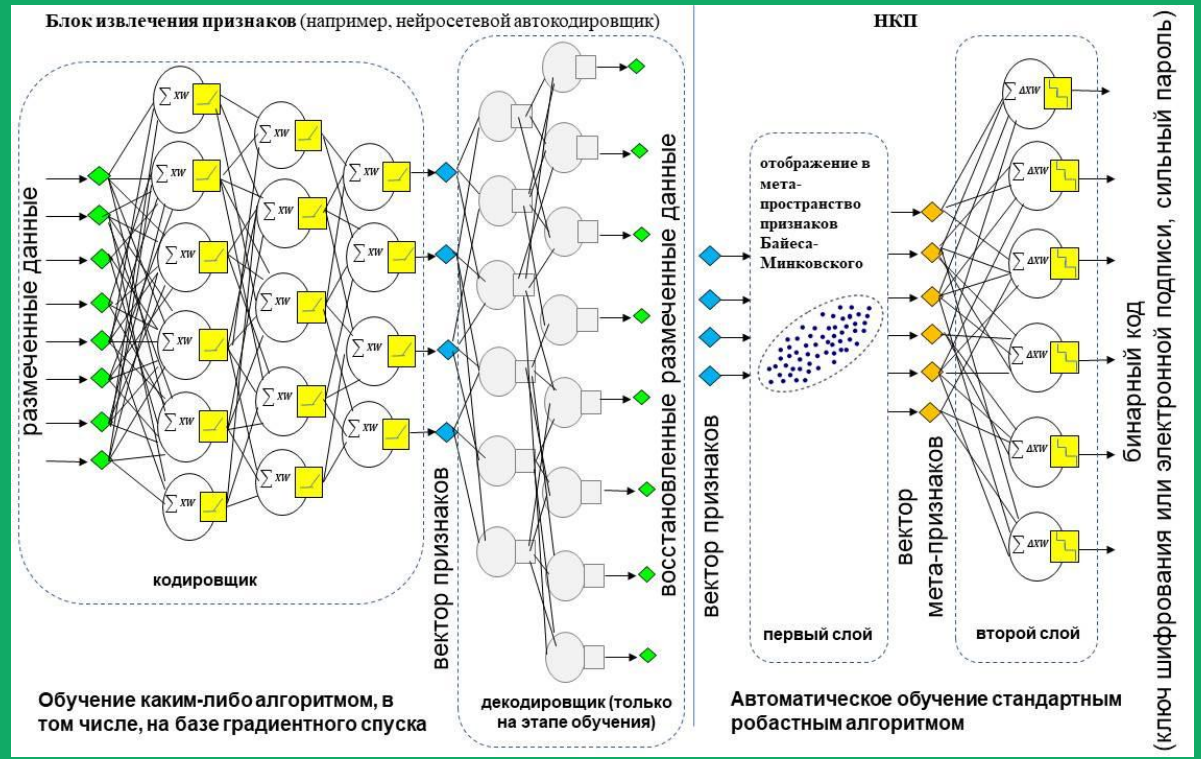


Рисунок 1. Пример схемы связывания ключа и класса образов «Свой»: блок извлечения признаков (слева) и НКП (справа)

Описание модели

Исследуемый способ построения нейросетевых моделей (в терминах стандарта - нейро-корреляционный преобразователь, т.е. НКП) рассматривает два типа входных данных:

- «Свой» - данные класса, которые должны корректно классифицироваться обученным НКП (например, лицо конкретного человека);
- «Чужие» - произвольные входные данные, не являющиеся «Своими» (случайная выборка без возвращения лиц различных людей).

Нейро-корреляционный преобразователь – нейросетевой преобразователь образов в код на основе корреляционных связей между нейронами.

Краткое описание предлагаемого проекта стандарта

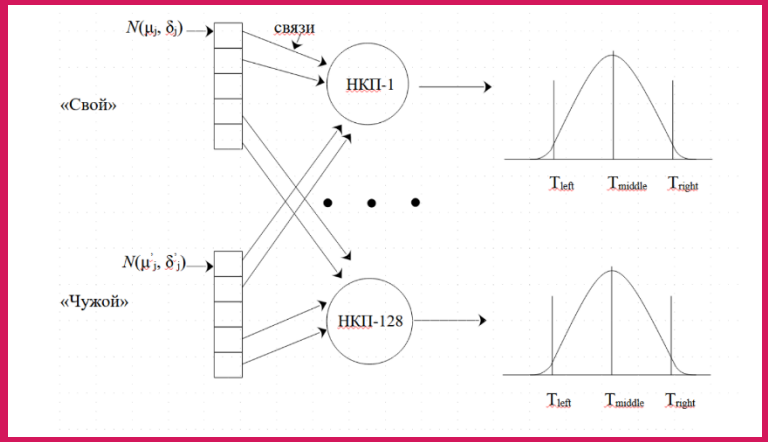
1. При обучении подаются вектора признаков классов «Свой» и «Чужой»
2. Из исходного вектора признаков «Свой» $\bar{a}$ формируется вектор мета-признаков вида $\left|\left(\frac{a_j}{\delta_j}\right)^s - \left(\frac{a_i}{\delta_i}\right)^s\right|$, являющиеся оценками корреляции исходных признаков
3. a_j - j-й мета-признак вектора признаков
4. s - параметр
5. δ_j- нормирующий коэффициент - среднеквадратичное отклонение j-го признака для классов «Чужие»
6. Входы корреляционного нейрона выбираются псевдослучайно с учетом корреляции мета-признаков
7. Весовой коэффициент корреляционного нейрона – взвешенная среднеквадратичное отклонение мета-признака от среднего значения по классу «Свой»
8. Предполагается, что распределение каждого признака близко к нормальному
9. Распределение на выходе нейрона также формируется близким к нормальному
10. Для выходного распределения каждого нейрона подбираются границы $T_{left}, T_{middle}, T_{right}$, так чтобы вероятности попадания в каждую область были примерно равны

Обозначения и краткое описание предлагаемого проекта стандарта

Обученный НКП хранит значения следующих параметров:

1. связи корреляционных нейронов с признаками
2. нормирующие коэффициенты признаков δ_i
3. веса нейронов
4. границы $T_{left}, T_{middle}, T_{right}$

$$\phi(y) = \begin{cases} 3, & y < T_{left} \\ 2, & T_{left} \le y < T_{middle} \\ 1, & T_{middle} \le y < T_{right} \\ 0, & y \ge T_{right} \end{cases}$$



Эксперимент. Идея атаки

Реализуется атака проверки принадлежности обучающему множеству:

1. Предполагаем, что у нарушителя есть:
    1. Обученный ранее НКП
    2. Набор примеров из различных классов «Свой», при этом среди них есть один, который использовался для обучения доступного нарушителю НКП
2. Задачей является определение соответствия между обученным НКП и истинным классом «Свой» из доступной нарушителю выборки
3. Нам достаточно показать возможность реализации атаки, когда у нарушителя всего два класса: «Свой» и какой-то еще - «Другой»

Эксперимент. Этапы атаки

1. Извлечь из атакуемого НКП связи корреляционных нейронов, границы T_{left}, T_{middle}, T_{right}, нормирующие коэффициенты признаков δ_i.
2. Для каждого доступного ему набора данных (из некоторого класса) попытаться обучить новый НКП, используя связи корреляционных нейронов, и получить новые границы T_{left}, T_{middle}, T_{right}.
3. Выбрать среди обученных НКП тот, у которого границы ближе всего по некоторой метрике к границам атакуемого нейрона.

Эксперимент. Обучающие множества

- G_1 - выборка примеров «Свой»;
- I_1 - выборка примеров «Чужой»;
- G_1' - еще одна выборка примеров «свой» из той же генеральной совокупности, из которой выбиралось множество G_1;
- G_2 - выборка примеров «Свой» из другой генеральной совокупности;
- I_2 - выборка примеров «Чужой», такая что $\delta_{I_1,i} = \delta_{I_2,i}$.

Номер эксперимента	Параметры нормального распределения для выборок G_1 и G_1'	Параметры нормального распределения для выборок I_1 и I_1'	Параметры нормального распределения для выборок G_2
1	$\mathcal{N}(\mu,1), \mu = \{2,3\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,3\}$
2	$\mathcal{N}(\mu,1), \mu = \{2,3,4\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,3,4\}$
3	$\mathcal{N}(\mu,1), \mu = \{2,...,5\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,...,5\}$
4	$\mathcal{N}(\mu,1), \mu = \{2,...,6\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,...,6\}$
5	$\mathcal{N}(\mu,1), \mu = \{2,...,7\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,...,7\}$
6	$\mathcal{N}(\mu,1), \mu = \{2,...,8\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,...,8\}$
7	$\mathcal{N}(\mu,1), \mu = \{2,...,9\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,...,9\}$
8	$\mathcal{N}(\mu,1), \mu = \{2,...,10\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,...,10\}$
9	$\mathcal{N}(\mu,1), \mu = \{2,...,11\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,...,11\}$
10	$\mathcal{N}(\mu,1), \mu = \{2,...,12\}$	$\mathcal{N}(\mu(i),1), \mu = \{2,..,13\}$	$\mathcal{N}(\mu,1), \mu = \{2,...,12\}$

Таблица № 1. Пример параметров генерации выборок

Описание эксперимента

- обучить НКП-1 на исходных множествах G_1 и I_1, в результате мы получим наборы T^1_{left}, T^1_{middle}, T^1_{right};
- обучить НКП-2 на множествах G'_1 и I_2, используя связи нейронов и нормирующие коэффициенты из НКП-1, в результате мы получим наборы T^2_{left}, T^2_{middle}, T^2_{right};
- обучить НКП-3 на множествах G_2 и I_2, используя связи нейронов и нормирующие коэффициенты из первого НКП-1, в результате мы получим наборы T^3_{left}, T^3_{middle}, T^3_{right};
- вычислить значение статистики *F*.

$$z_2(i) = \max(\left|T^1_{left}(i) - T^2_{left}(i)\right|, \left|T^1_{middle}(i) - T^2_{middle}(i)\right|, \left|T^1_{right}(i) - T^2_{right}(i)\right|),$$

$$z_3(i) = \max(\left|T^1_{left}(i) - T^3_{left}(i)\right|, \left|T^1_{middle}(i) - T^3_{middle}(i)\right|, \left|T^1_{right}(i) - T^3_{right}(i)\right|), \text{ где } i = 1, \dots, d.$$

$$F = \sum_{i=1}^{d} I\{z_2(i) < z_3(i)\}$$

Результаты атаки

Номер эксперимента	Количество «успешно» обученных нейронов НКП-2	Количество «успешно» обученных нейронов НКП-3	Среднее значение статистики F	Число запусков, в которых значение статистики $F > d/2$
1	128	128	68	45
2	128	128	78	48
3	128	128	81	49
4	128	128	72	46
5	128	128	91	50
6	128	128	88	50
7	128	128	83	49
8	128	128	77	48
9	128	128	65	46
10	128	128	93	50

Таблица № 2. Среднее число «успешно» обученных нейронов НКП-2 и НКП-3, среднее значение статистики F и число запусков, в которых значение статистики $F > {}^{d}/_{2}$ для каждого эксперимента

Выводы

- Предложен статистический критерий, который позволяет реализовать атаку проверки принадлежности обучающему множеству
- Предложенный критерий позволяет корректно определять входные данные, использовавшиеся для обучения атакуемого нейросетевого преобразователя
- Параметры обученного нейрона допускают утечку конфиденциальных данных об обучающей выборке
- Предлагаемый к стандартизации механизм не обеспечивает заявленных свойств безопасности
- Целесообразно оценить возможность реализации других типов атак

Вопросы

???