ПОСТРОЕНИЕ СОСТЯЗАТЕЛЬНОЙ АТАКИ НА НЕЙРО-КОРРЕЛЯЦИОННЫЙ ПРЕОБРАЗОВАТЕЛЬ ОБРАЗОВ В КОД ИЗ ПРОЕКТА СТАНДАРТА

Сухарев И.В., ВМК МГУ, ООО «КРИПТО-ПРО»

Маршалко Г.Б., МИЭМ НИУ ВШЭ, Академия криптографии Российской Федерации

небезопасный

ГОСТ Р 52633.5–2011

«Автоматическое обучение нейросетевых преобразователей биометрия-код доступа»

2012 г.

???

ПРОЕКТ ГОСТ Р XXXX

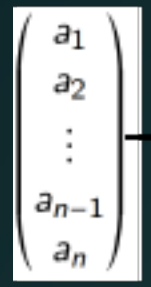
«Нейросетевые алгоритмы в защищенном исполнении. Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации»

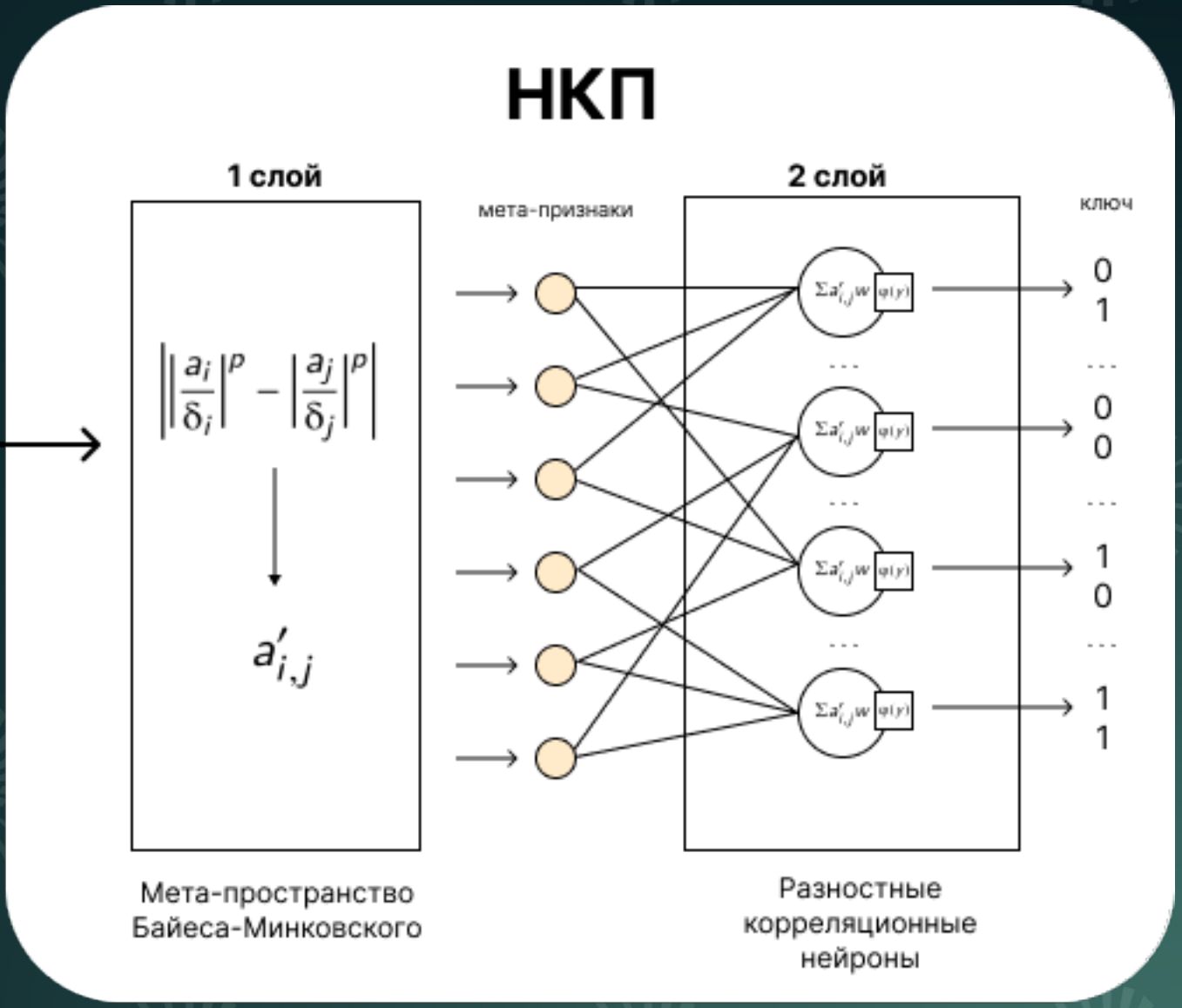
2022 г.

ПРОЕКТ СТАНДАРТА, ПРЕДЛОЖЕННЫЙ В 2022 ГОДУ

«Чужой»

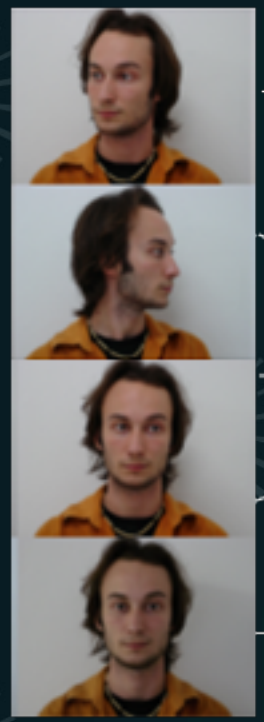
«Свой»





НЕЙРО-КОРРЕЛЯЦИОННЫЙ ПРЕОБРАЗОВАТЕЛЬ (НКП)

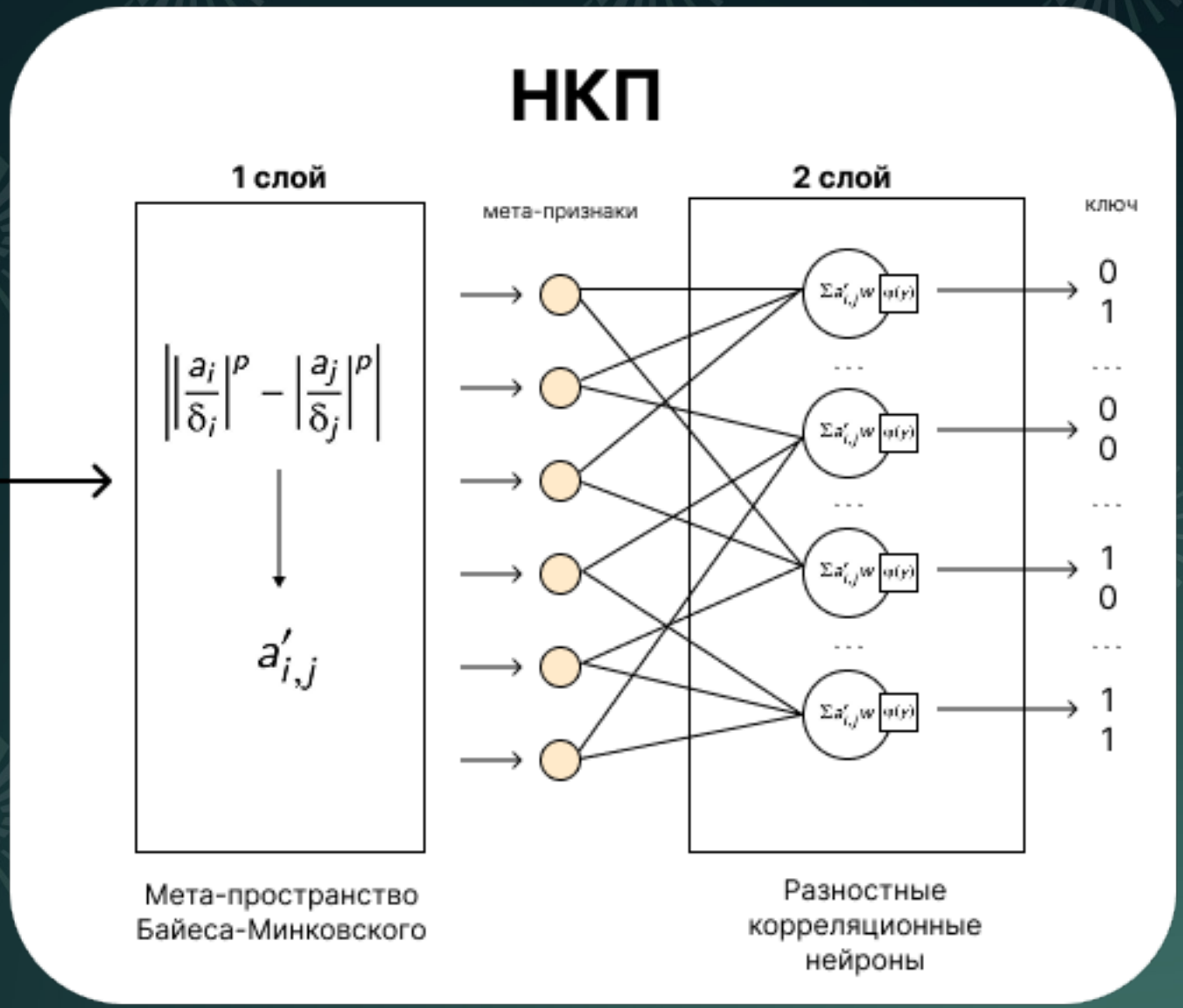
В работе [1] была продемонстрирована утечка информации из НКП, позволяющая построить атаку принадлежности обучающей выборке



Энкодер

Латентное пространство

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{pmatrix}$$



НЕЙРО-КОРРЕЛЯЦИОННЫЙ ПРЕОБРАЗОВАТЕЛЬ (НКП)

[1] РОМАНЕНКОВ Р.А., МАРШАЛКО Г.Б., ТРУФАНОВА Ю.А. Анализ безопасности проекта национального стандарта «Нейросетевые алгоритмы в защищенном исполнении. Автоматическое обучение нейросетевых моделей на малых выборках в задачах классификации». *Труды Института системного программирования РАН*. 2023;35(6):179-188. https://doi.org/10.15514/ISPRAS-2023-35(6)-11

МОДЕЛЬ НАРУШИТЕЛЯ

WhiteBox модель («белый ящик»)

Дано:

1) Обученный $НКП_A$ для класса $\mathcal{A}$

2) Изображение A^0 из класса $\mathcal{A}$

3) Ключ $\boldsymbol{k_{target}}$, т. к. $НКП_A(\boldsymbol{A^0}) = \boldsymbol{k_{target}}$

Задача:

$$\min_{\boldsymbol{x^*}} \mathbf{HD}\Big(НКП(\boldsymbol{x^*}),\ \boldsymbol{k_{target}}\Big), \quad \|\boldsymbol{x^*} - \boldsymbol{x_0}\|_2 \leq \boldsymbol{\varepsilon}$$

$\mathrm{HD}(k^*, k) = \sum_{i=1}^{N} |k_i^* - k_i| -$ расстояние Хемминга между ключом k^* и k

x_0 – исходный вектор "Чужого"

АЛГОРИТМ АТАКИ

1. Нейроны разделяются на *«правильные»* и *«неправильные»*;
2. Для каждого *«неправильного»* вычисляется градиент выхода по входным признакам;
3. Перед изменением соответствующего признака a вычисляется масштабирующий коэффициент:

   $\mathbf{scale}(a) = \mathbf{exp}\left(-\mathbf{relations}(a) \cdot n_{\mathbf{correct}}(a)\right)$, где $\mathbf{relations}(a) \in [0,1]$ - доля количества связей признака a от макс. кол-ва связей, $n_{\mathbf{correct}}(a)$ - число уже *«правильных»* нейронов, затрагиваемых этим признаком;
4. Для *«правильных»* нейронов оценивается близость отклика к границе активации. Нейроны, расположенные вблизи границы, защищаются отталкивающим градиентом (минимальным изменением признаков) к центру нужного интервала.

РЕЗУЛЬТАТЫ АТАКИ

ИЗНАЧАЛЬНОЕ ФОТО

ПОСЛЕ ИЗВЛЕЧЕНИЯ ПРИЗНАКОВ + РЕКОНСТРУКЦИИ

Целевой класс:

Атакующий класс:

Результат:
(для ключа 256 бит)

НАЧАЛЬНОЕ РАССТОЯНИЕ ХЕММИНГА **94**

КОНЕЧНОЕ РАССТОЯНИЕ ХЕММИНГА **3**

ВЫВОДЫ

- Заявленные свойства безопасности НКП не выполняются в полной мере, так как в модели «белого ящика» нарушитель может построить состязательный пример, практически неотличимый от атакующего
- В качестве повышения устойчивости НКП к такому виду состязательных атак предлагается «усложнить» алгоритм построения структуры связей мета-признаков и нейронов на этапе обучения

СПАСИБО ЗА ВНИМАНИЕ!

Сухарев И.В. sukharev@cryptopro.ru

Маршалко Г.Б. marshalko_gb@tc26.ru