МНОГОМЕРНЫЙ ДВОИЧНЫЙ КЛАССИФИКАТОР ДЕРЕВА РЕШЕНИЙ НА ОСНОВЕ НЕГЛУБОКОЙ НЕЙРОННОЙ СЕТИ ДЛЯ ОБНАРУЖЕНИЯ ФИШИНГА

Авазжон Марахимов, Улугбек Охундадаев, Кабул Худайбергенов
Кафедра Обработки информации и систем управления, Ташкентский государственный технический университет, Ташкент, Узбекистан
kabul.kudaybergenov@gmail.com

Аннотация. В работе представлен подход к обнаружению фишинговых веб-сайтов на основе многомерного двоичного дерева решений, узлы которого формируются с помощью неглубокой нейронной сети. Метод сочетает интерпретируемость классических деревьев решений с обобщающей способностью нейронных сетей, что особенно важно в задачах кибербезопасности, где требуется как высокая точность классификации, так и объяснимость принимаемых решений.

Ключевые слова: *обнаружение фишинга, многомерное дерево решений, неглубокая нейронная сеть, функция активации, классификация.*

1. Введение

Фишинг остаётся одной из наиболее распространённых и опасных угроз информационной безопасности. По данным Anti-Phishing Working Group, ежегодно фиксируются сотни тысяч уникальных фишинговых веб-сайтов, направленных на кражу учётных данных, финансовой информации и персональных данных пользователей.

В данной работе предлагается гибридный подход, то есть многомерный двоичный классификатор дерева решений на основе неглубокой нейронной сети (NNDT), адаптированный для задачи обнаружения фишинговых веб-сайтов. Метод объединяет преимущества обоих подходов: иерархическую структуру и интерпретируемость дерева решений с нелинейной обобщающей способностью нейронной сети.

Задача обнаружения фишинга формулируется как задача бинарной классификации. Пусть дан набор данных $D = (x_i, y_i)_{i=1}^n$, где $x_i \in \square^d$ - вектор признаков, извлечённых из веб-сайта (URL, HTML-контент, метаданные), а $y_i \in 0,1$ - метка класса (0 - легитимный сайт, 1 - фишинговый). Требуется построить классификатор $f : \square^d \to 0,1$, минимизирующий ошибку классификации и обеспечивающий высокую скорость принятия решения в реальном времени.

2. Предлагаемый метод

Предлагаемый классификатор NNDT строится рекурсивно в нисходящем порядке. В каждом внутреннем узле дерева обучается неглубокая нейронная сеть с одним скрытым слоем, формирующая разделяющую гиперплоскость между двумя подмножествами объектов. Процесс построения дерева описывается следующим алгоритмом:

Вход: Обучающая выборка D фишинговых и легитимных URL с известными метками классов.

Выход: Двоичное дерево решений, во внутренних узлах которого хранятся параметры w неглубокой нейронной сети, а в листьях - метки классов.

1. Инициализировать $\Psi := D$, $X := \varnothing$.
2. Извлечь из Ψ объекты одного класса и обозначить их как X.
3. Пока Ψ не пусто:
   - Обучить неглубокую нейронную сеть с выбранной функцией активации, разделяющую объекты X и Ψ, X.
   - Если выполнено условие кросс-валидации, создать листовой узел с меткой класса большинства и исключить эти объекты из Ψ.
   - Иначе создать дочерний узел и продолжить разделение.

Объединённая функция активации, $f_{jrelu}(x) = \beta f_{prelu}(x) + (1-\beta) f_{pelu}(x)$, где $\beta \in [0,1]$ - обучаемый коэффициент объединения, определяющий относительный вклад каждой компоненты в процессе обучения.

3. Численные эксперименты

Результаты численных экспериментов показывают, что предложенный метод NNDT с объединённой функцией активации превосходит все базовые методы по всем метрикам качества. Точность классификации достигает 99.18%, что на 3-5 процентных пунктов выше, чем у классических одномерных деревьев решений (CART, C4.5), и на 2-3 процентных пункта выше, чем у многомерных моделей MDT-1 и MDT-2.

Таблица 1. Сравнение качества классификации методов.

Метод	**Accuracy, %**	**Precision, %**	**Recall, %**	**F1, %**	**Время обуч., с**	**Время класс., мс**	**Число узлов**
CART	93.24	93.10	92.85	92.97	0.52	0.08	187
C4.5	94.11	93.95	93.78	93.86	0.48	0.09	165
Random Forest	96.82	96.75	96.60	96.67	3.24	0.45	-
MLP(deepNN)	97.15	97.02	96.95	96.98	12.85	0.38	-
MDT	95.68	95.52	95.41	95.46	2.45	0.21	412
NNDT с (1)	**99.18**	**99.12**	**99.05**	**99.08**	1.15	0.12	**124**

4. Заключение

В работе предложен гибридный метод обнаружения фишинговых веб-сайтов на основе многомерного двоичного дерева решений с неглубокой нейронной сетью. Метод сочетает интерпретируемость дерева решений с обобщающей способностью нейронной сети и использует новую объединённую функцию активации с обучаемым коэффициентом.

Список литературы

1. Marakhimov A.R., Kudaybergenov J.K., Khudaybergenov K.K., Ohundadaev U.R. A multivariate binary decision tree classifier based on shallow neural network // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2022. Т. 22, № 4. С. 725–733.

2. Wang F., Wang Q., Nie F., Yu W., Wang R. Efficient tree classifiers for large scale datasets // Neurocomputing. 2018. V. 284. P. 70–79.