Структура, функции системы интеллектуальной обработки данных

Structure, functions of intelligent data processing system

Д. Д. Игошин,
зам. начальника отдела организации оперативной службы – начальник дежурной смены, ЦУКС ГУ МЧС России по Нижегородской области/соискатель, Санкт-Петербургский университет ГПС МЧС России, г. Нижний Новгород

D. D. Igoshin,
deputy head of the department of organization of operational service – head of the shift on duty, Central control center of the Main directorate of the Ministry of emergency situations of Russia for the Nizhny Novgorod region/competitor, St. Petersburg university of the State fire service of the Ministry of emergency situations of Russia, Nizhny Novgorod

О. С. Маторина,
начальник сектора, отдел 1.3 НИЦ ОУП ПБ, ФГБУ ВНИИПО МЧС России, г. Балашиха
✉ odp1313@yandex.ru

O. S. Matorina,
head of sector, department 1.3 of the research center of the OUP PB, Federal state budgetary institution VNIIPO of the Ministry of emergency situations of Russia, Balashikha

С. В. Нестерова,
старший научный сотрудник, отдел 5.1 НИЦ ИТ, ФГБУ ВНИИПО МЧС России, г. Балашиха

S. V. Nesterova,
senior researcher, department 5.1 NRC IT, FGBU VNIIPO EMERCOM of Russia, Balashikha

О. В. Чирко,
старший научный сотрудник, отдел 2.6 НИЦ ПТ и ПА, ФГБУ ВНИИПО МЧС России, г. Балашиха

O. V. Chirko,
senior researcher, department 2.6, Research center for fire protection and PA, FGBU VNIIPO EMERCOM of Russia, Balashikha

В статье рассмотрены перспективы и особенности развития систем интеллектуальной обработки данных. Отдельное внимание в процессе исследования уделено требованиям, которые предъявляются к таким системам, их функциям и структуре. Также обозначен круг решаемых задач и сферы использования. Кроме того, обозначены факторы, предопределяющие популярность и широкое распространение методов нейросетевого моделирования на основании глубокого обучения. Особый акцент сделан на алгоритме проведения интеллектуального анализа, описана разработанная автором универсальная структурная схема интеллектуального анализа данных. На примере решения задачи выявления зашифрованного и вредоносного сетевого трафика с использованием одномерной сверточной нейронной сети описаны особенности интеллектуального подхода к обработке и анализу данных.

The article considers the prospects and features of the development of intelligent data processing systems. Special attention in the process of research is given to the requirements that apply to such systems, their functions and structure. The range of tasks to be solved and the scope of use are also indicated. In addition, the factors that predetermine the popularity and widespread use of neural network modeling methods based on deep learning are identified. Particular emphasis is placed on the algorithm for conducting data mining, and the universal block diagram of data mining developed by the author is described. On the example of solving the problem of detecting encrypted and malicious network traffic using a one-dimensional convolutional neural network, the features of an intelligent approach to data processing and analysis are described.

Ключевые слова: данные, интеллектуальный анализ, машинное обучение, цифровизация, база, нейронная сеть.

Keywords: data, data mining, machine learning, digitalization, database, neural network.

Введение

Стремительный научно-технический прогресс в области цифровизации, данных и аналитики изменяет бизнес-ландшафт, повышает производительность, способствует появлению новых бизнес-инноваций и форм конкуренции. Развитие и усложнение технологий в свою очередь стимулирует новые волны достижений в области робототехники, аналитики, искусственного интеллекта и особенно машинного обучения. В совокупности все это представляет собой шаг вперед в расширении технических возможностей, которые несомненно будут иметь глубокие последствия для бизнеса, экономики и, в более широком смысле, для общества в целом.

В таких условиях предприятия во всех отраслях промышленности сталкиваются с огромным давлением, требующим совершать операции быстрее, точнее и в больших объемах. Все больше субъектов хозяйствования фокусируются на улучшении опыта для клиентов и сотрудников, чтобы удержать таланты и увеличить доходы. Одним из способов преодолеть это давление, является внедрение новых аналитических технологий, таких как интеллектуальная обработка данных (IDA), которые основаны на инструментах распознавания образов, компьютерном зрении, нечеткой логике, системах баз данных [1]. Задача интеллектуальной обработки данных включает в себя обнаружение знаний, прогнозирование, моделирование процессов или создание новых систем, основанных на знаниях. Можно ли применять метод IDA для данных, полученных при решении плохо формализуемых или вовсе не формализуемых задач с используемой статистической обработкой исходных данных (корреляционный и ре-

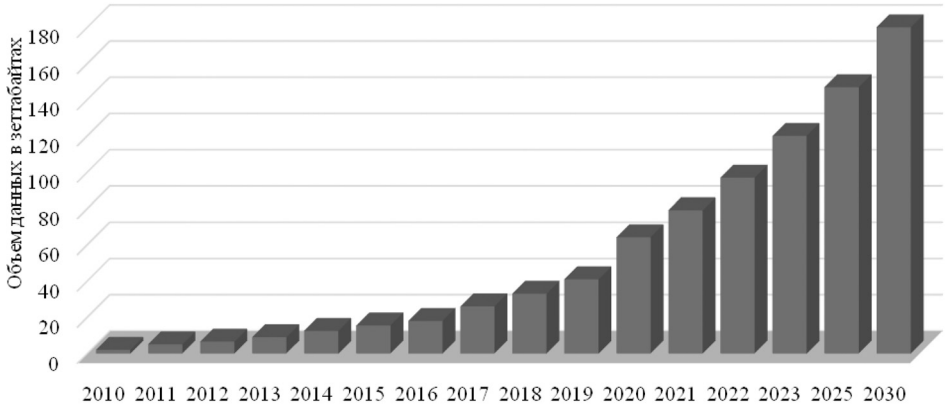

Рис. 1. Объем созданных, использованных и сохраненных данных по состоянию на 2022 г. (прогноз 2023-2030 гг.) [2]

грессионный анализ, дисперсионный анализ и т. п.)? И если да, то работают ли и как работают представленные автором алгоритмы и схемы на рис. 4 и 5?

Интеллектуальная обработка данных является эффективным средством решения сложных слабо структурированных задач, когда качественные и количественные данные известны, однако присутствует малоизвестные сторонние факторы. К этому классу традиционно относятся задачи классификации, кластеризации, аппроксимации многомерных отображений, прогнозирования временных рядов, нелинейной фильтрации, управления сложными технологическими объектами. Системы интеллектуальной обработки данных, основанных на знаниях, способны решать неструктурированные задачи, где присутствуют только качественное описание данных, но отсутствуют зависимости между характеристиками. В данном случае система направлена на извлечение знаний, изучение проблемы, обнаружение закономерностей, их структуризацию и разработку моделей представления знаний (логической, эволюционной, имитационной, структурной, статической).

Только качественное описание, основанное на суждениях лиц, принимающих решения, количественные зависимости между основными характеристиками задачи не известны

Сегодня можно отметить уже множество достижений в применении методов IDA в различных областях, таких как маркетинг, медицина, финансы, промышленность, исследования и разработки. Инструменты и приложения IDA дают положительные результаты и постоянно стимулируют поиск новых областей приложения благодаря преимуществам и возможностям, которые дает эта технология.

Быстро растущий объем информации в режиме реального времени, возникающий в результате активности в Интернете, мультимедиа, электронной коммерции, smart-производству способствует появлению более сложных методов IDA. Согласно исследованиям, общий объем данных, создаваемых, фиксируемых, копируемых и потребляемых во всем мире растет стремительными темпами и достигнет 147 зеттабайт в 2025 г. В течение следующих пяти лет до 2030 г. прогнозируется рост объема создаваемых данных в мире до более чем 180 зеттабайт [2] (см. рис. 1).

В контексте вышеизложенных тенденций необходимо отметить, что общая идея анализа больших объемов данных является привлекательной и интуитивно понятной, но с технической точки зрения — это гораздо более сложный и трудоемкий процесс. При работе с такого рода информацией уже недостаточно относительно простой и прямолинейной статистики. Для более эффективного использования данных, собранных из больших и сложных баз, должны существовать определенные стратегии, методы и приемы.

Процесс IDA, поиска и построения соответствующей модели часто является итеративным, так как нужно найти и выявить разные сведения, которые можно извлечь. Необходимо также понимать, как их связать, сгруппировать и объединить с другими данными для получения результата. После обнаружения новых элементов и аспектов подход к поиску источников и форматов данных с последующим сопоставлением этой информации с заданным результатом может измениться.

Таким образом, с учетом вышеизложенного, стратегии IDA, технические особенности и используемые приемы составляют важное направление научного исследования, что и обуславливает выбор темы данной статьи.

Анализ публикаций по теме исследования

Над разработкой алгоритмов и методов, которые используются в процессе IDA, их усовершенствованием трудятся такие авторы как М. В. Федотов, В. В. Грачев, О. И. Шелухин, Д. В. Костин, А. В. Велигура, Э. К. Мусаева, Yao, Hui, Zhao, Shibo, Gao, Zhiwei.

Созданию методологии разработки точных лингвистических моделей для IDA посвятили свои труды Д. В. Катасева, А. О. Баринова, Е. В. Тимощенко, А. Ф. Ражков, В. Н. Вероха, К. А. Борисенко, O. Cordon, F. Herrera, D. J. Hand, J. N. Kok, M. Berthold.

Необходимость, способы и приемы предварительной обработки данных в интеллектуальном анализе рассматриваются в публикациях К. Н. Жаткиной, О. А. Крейдера, О. И. Христодуло, С. В. Павлова, Е. С. Брекоткиной, A. Famili, G. E. Lasker; X. Liu, C. L. Tsien, J. C. Fackler, X. Liu, P. Cohen.

Рис. 2. Алгоритм IDA

Нерешенные части общей проблемы

Однако, несмотря на имеющиеся труды и наработки, ряд вопросов в данной предметной плоскости с учетом достижений Четвертой промышленной революции, сквозной цифровизации остается нераскрытым и достаточно дискуссионным. Так, в более детальной проработке нуждаются технические проблемы IDA и их последствия. Особого внимания заслуживает потенциал интеграции «умных» датчиков с системами интеллектуального анализа и методы обмена информацией между их ключевыми элементами.

Таким образом, цель статьи заключается в изучении особенностей, структуры и функции системы интеллектуальной обработки данных.

Результаты

Процесс изучения наборов данных для вывода полученных результатов называется анализом данных. Он предполагает сочетание различных техник и методов обработки информации [3]. В свою очередь IDA — это практика использования компьютерных систем и процессов для выполнения аналитических задач с минимальным вмешательством человека или без него.

Назначение IDA заключаются в том, что он позволяет преобразовать неструктурированную и полуструктурированную информацию в пригодные для использования данные. На сегодняшний день, несмотря на широкое внедрение и применение цифровых технологий во многих сферах деятельности, 80% всех бизнес-данных хранятся в неструктурированных форматах, таких как деловые документы, электронная почта, изображения и файлы PDF [4].

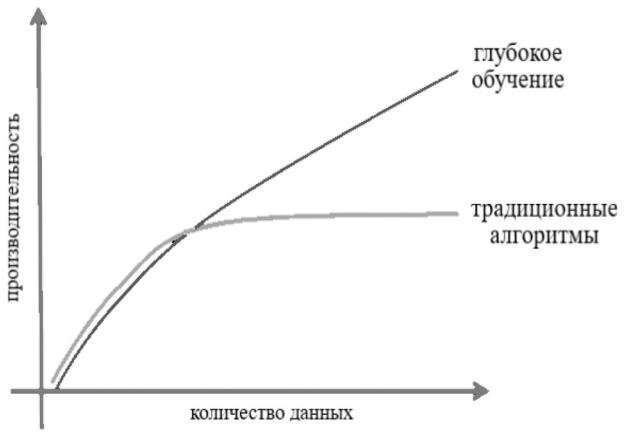

Рис. 3. Сравнение глубокого обучения и традиционных алгоритмов в IDA [8]

IDA — это следующее поколение автоматизации, способное собирать, извлекать и обрабатывать данные из различных форматов. Системы IDA используют технологии искусственного интеллекта, такие как обработка естественного языка, глубокое обучение и машинное обучение для классификации и категоризации необходимой информации, а также проверки извлеченных данных.

Типичный процесс IDA начинается с определения проблемы, зависящей от интересов аналитика данных. Затем определяются все источники информации и из накопленных данных формируется подмножество баз для применения IDA. Для обеспечения качества набор данных предварительно обрабатывается путем удаления шумов, обработки недостающей информации и преобразования в соответствующий формат. Затем к полученному набору применяется метод IDA или комбинация методов, соответствующих типу знаний, которые необходимо обнаружить. После этого обнаруженные знания обрабатываются, оцениваются и интерпретируются, как правило, с использованием некоторых инструментов постобработки, таких как методы визуализации. И на завершающем этапе информация представляется пользователю [5].

Обобщенный алгоритм IDA представлен на рис. 2.

Основные функции системы интеллектуальной обработки данных включают в себя: статистическую обработку данных (корреляционный и регрессионный анализ, дисперсионный анализ и т. п.), распознавание образов (классификация с обучением), кластеризация (классификация без обучения) идентификация (обнаружение опознавательных признаков исследуемых объектов), прогнозирование (определение тенденций развития процессов), извлечение знаний из данных (data mining) и текстов (text mining).

В соответствии с целями и интересами конечного пользователя, такими как характеристика содержимого набора данных в целом или установление связи между подмножествами деталей в наборе данных, IDA может преследовать три возможные задачи — предиктивное моделирование, кластеризация и анализ связей.

Целью предиктивного моделирования является составление прогнозов на основе существенных характеристик данных. Задача состоит в том, чтобы построить модель для сопоставления элемента данных с одним из нескольких заранее определенных

классов или с реальной переменной прогноза. Любой контролируемый алгоритм машинного обучения, который обучает модель на предыдущих или на существующих данных, может быть использован для выполнения прогностического моделирования. В модель вводятся некоторые уже известные факты с правильными ответами, на основе которых она учится делать точные прогнозы [6]. Нейронные сети, деревья решений, байесовские классификаторы, генетические алгоритмы, грубые множества и нечеткие множества — некоторые из методов, используемых для отображения дискретных целевых переменных. Методы регрессии, деревья индукции, нейронные сети и радиальные базисные функции — пример подходов, используемых для отображения целевых переменных с непрерывным значением.

Целью кластеризации является выявление элементов с похожими характеристиками и, таким образом, создание иерархии классов из существующего набора событий. Для выполнения кластеризации может использоваться любой алгоритм машинного обучения без наблюдения, для которого не известен заранее определенный набор категорий данных. В модель вводятся некоторые уже известные факты, на основе которых она выводит категории с похожими характеристиками [7]. К числу основных методов кластеризации относятся следующие: разбиение на части, иерархический алгоритм, метод определения ближайших соседей.

Анализ связей устанавливает внутренние отношения между элементами в заданном наборе данных. Эта цель достигается с помощью задач обнаружения ассоциаций, нахождения последовательных шаблонов и выявления аналогичных временных последовательностей. В ходе подобного анализа становится возможным идентифицировать образцы и тенденции, предсказывая корреляцию элементов, которая иначе

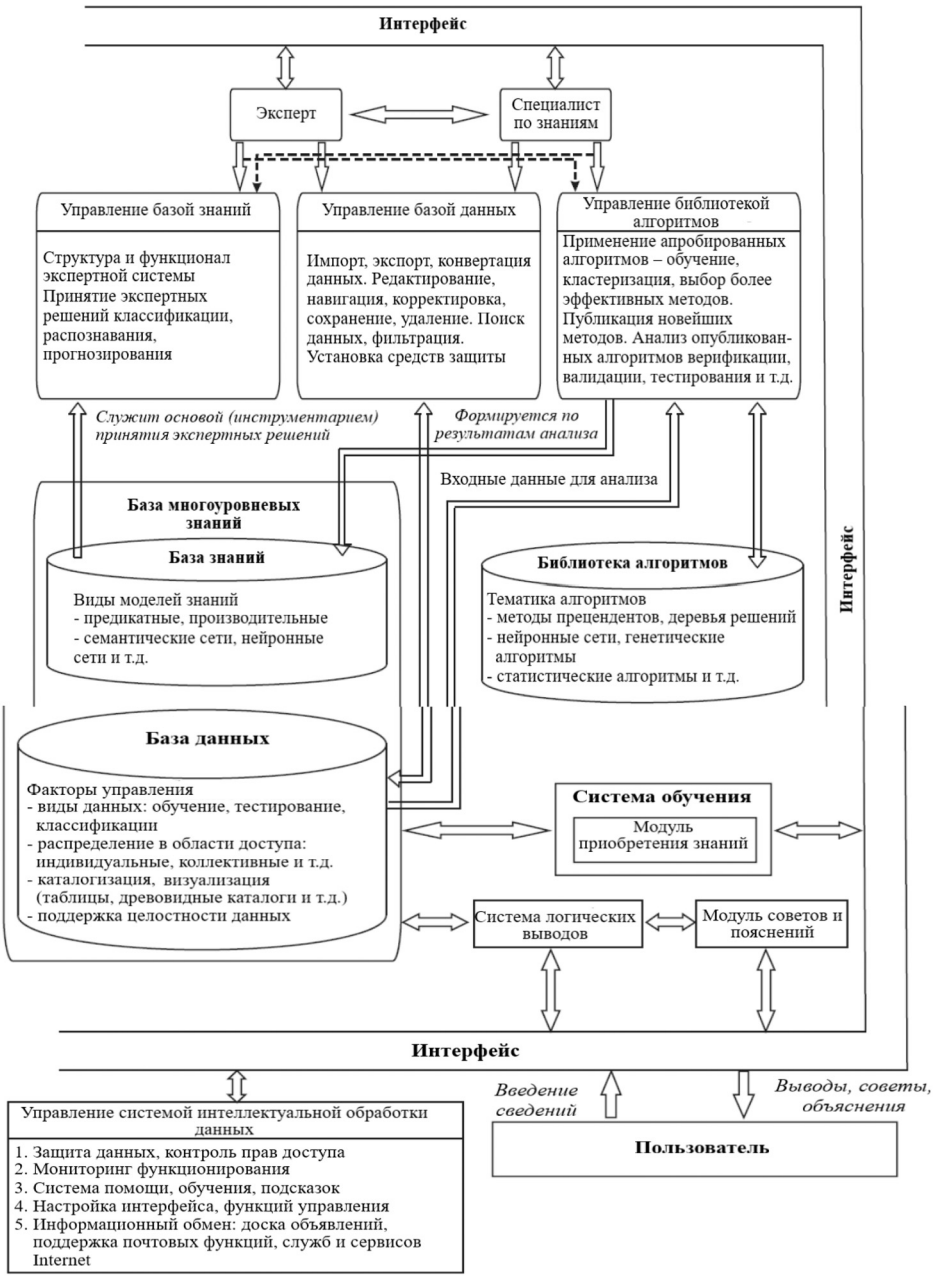

Рис. 4. Универсальная структурная схема IDA

не очевидна. Методы анализа связей основаны на подсчете встречаемости всех возможных комбинаций элементов. Одними из наиболее широко используемых алгоритмов являются Apriori и его вариации.

В последние несколько лет особую популярность в IDA получили методы нейросетевого моделирования на основании глубокого обучения. Причиной этому являются следующие факты.

1. Аккумуляция и регулярное генерирование большого количества данных. Огромное количество доступных данных, собранных за последнее десятилетие, в значительной степени способствовало популярности глубокого обучения. Это позволило нейронным сетям по-настоящему раскрыть свой потенциал, поскольку они становятся тем лучше, чем больше данных в них подается. Приведенный на рис. 3 график наглядно это иллюстрирует.
2. Вычислительная мощность. Еще одной очень важной причиной развития глубокого обучения является доступная сегодня вычислительная мощность, которая позволяет обрабатывать больше данных. По словам Рэя Курцвейла, одного из лидеров в области искусственного интеллекта, вычислительная мощность умножается на постоянный коэффициент в каждую единицу времени (например, удваивается каждый год), а не просто увеличивается постепенно. Это означает, что вычислительная мощность растет экспоненциально [9].
3. Алгоритмы. Третий фактор, который повысил популярность глубокого обучения, — это успехи, достигнутые в алгоритмах анализа. Последние прорывы в разработке алгоритмов в основном связаны с тем, что они работают намного быстрее, чем раньше, что позволяет использовать все больше и больше данных.

Производительность методов нейросетевого моделирования на основании глубокого обучения при классификации и генерации данных для изображений/аудиозаписей является очень высокой. Однако, если речь идет о необходимости анализа наборов табличных данных нейронные сети демонстрируют крайне низкие результаты. Табличные данные неоднородны и могут приводить к плотным числовым и разреженным категориальным признакам. Кроме того, корреляция между признаками слабее, чем пространственная или семантическая связь в изображениях или речевых данных. Помимо этого, проблема с табличными данными заключается в существовании категориальных атрибутов, поскольку нейронная сеть принимает в качестве входных данных только действительные числа.

Опыт и практика свидетельствуют о том, что IDA можно использовать несколькими способами. Например, можно автоматизировать полные процессы обработки данных, создавать комплексные информационные панели бизнес-аналитики, разрабатывать самоуправляемые модели машинного обучения или автоматизировать отдельные задачи. Независимо от выбранного способа, масштаба и назначения развертывание IDA предполагает наличие определенной структуры, элементов и связей между ними.

На рис. 4 представлена разработанная автором универсальная структурная схема IDA.

На следующем этапе исследования представляется целесообразным рассмотреть, как именно IDA реализуется на практике. Для примера возьмем задачу выявления зашифрованного и вредоносного сетевого трафика на основе одномерной сверточной нейронной сети.

Для решения этой задачи была разработана одномерная сверточная нейронная сеть с шестнадцатеричными данными, которая объединяет механизмы нормализованной обработки и внимания. Добавление модулей механизма внимания Global Attention Block (GAB) и Category Attention Block (CAB) позволяет классифицировать и идентифицировать сетевой трафик. Извлекая информацию об эффективной нагрузке из шестнадцатеричного сетевого трафика, модель может идентифицировать большинство категорий сетевого трафика, включая зашифрованные и вредоносные данные.

На рис. 5 изображена блок-схема одномерной сверточной нейронной сети.

Исходные данные потока сначала вводятся в модуль предварительной обработки, а затем выводятся данные, которые могут быть непосредственно использованы сверточной нейронной сетью, в три этапа: обработка информации заголовка, извлечение ключевой информации и повторная обработка данных. Предварительно обработанные данные затем передаются в модуль обучения сети, где модель обучается путем последовательного извлечения признаков, упрощения

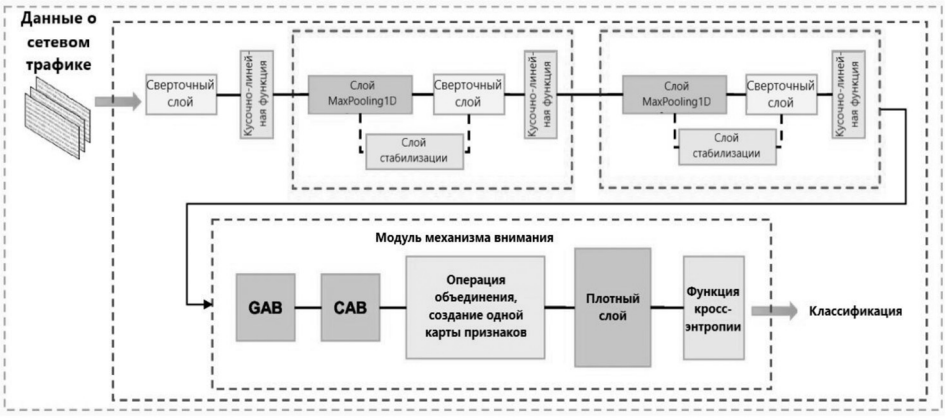

Рис. 5. Блок-схема одномерной сверточной нейронной сети

данных, оценки категории и корректировки обратной связи. В конце тестовые данные подаются в тестовый модуль, который содержит обученную модель сверточной нейронной сети, система оценивается и дорабатывается на основе результатов классификации.

Учитывая большой объем и нагрузку обрабатываемых данных сетевого трафика, традиционная модель одномерной сверточной нейронной сети не может удовлетворить требование идентификации типов и категорий зашифрованного и вредоносного трафика с высокой точностью. Поэтому в модель был добавлен нормализованный модуль обработки и модуль механизма внимания. Для экспериментов использовались наборы данных публичных сетей ISCX и USTC. Соотношение тестирования и обучающего набора было установлено на 7:3.

Как отображено в табл. 1, набор данных USTC-TFC показывает, что модель имеет более чем 98% точность идентификации вредоносного трафика, такого как Zeus, Virut и Nsis-ay. Это дает основания утверждать, что одномерная сверточная нейронная сеть, описанная в статье, обладает способностью обнаруживать вредоносный трафик.

Длина пакетов вредоносного трафика больше, чем длина пакетов обычного трафика. Значит модель может извлекать достоверные поля данных и точно идентифицировать различные типы вредоносного трафика с ограниченной длиной пакета и порядок их следования. Кроме того, определение правил анализа длин запросов клиента и ответов сервера позволяет однозначно идентифицировать вредоносный трафик. Имеются ли другие критерии (кроме длины пакетов) для идентификации вредоносного трафика?

Таблица 1

Идентификация вредоносного трафика с помощью одномерной сверточной нейронной сети

Вредоносный трафик	Точность, %	Восстановление, %	F1-оценка, %
Zeus	99,8	99,1	99,3
Virut	99,1	98,4	98,6
Nsis-ay	98,1	98,3	97,9

Заключение

Анализ данных — это итеративный и интерактивный процесс, включающий формулировку проблемы, обеспечение качества данных, построение модели, интерпретацию и последующую обработку результатов. IDA позволяет преобразовывать как неструктурированные, так и частично структурированные данные в полезную информацию для дальнейшего использования.

В статье описан интегрированный инструментарий для выполнения задач IDA, разработана универсальная структурная схема IDA и продемонстрированы на реальном примере возможности использования одномерной сверточной нейронной сети для выявления зашифрованного и вредоносного сетевого трафика.

Список использованных источников

1. Liu Xiaoming. Study on Intelligent Analysis and Processing Technology of Computer Big Data Based on Clustering Algorithm//Recent advances in electrical & electronic engineering. 2023. Vol. 16. № 2. P. 150-158.
2. Advances in intelligent data analysis XXI: 21st International Symposium on Intelligent Data Analysis, IDA 2023, Louvain-la-Neuve, Belgium, April 12-14, 2023 proceedings/Edited by Bruno Crémilleux, Sibylle Hess, Sigfried Nijssen. Cham: Springer, 2023. 499 p.
3. С. В. Пальмов, А. А. Диязитдинова, Е. С. Артюшкина. Сравнительный анализ возможностей интеллектуальных систем при выявлении скрытых закономерностей в данных//Электросвязь. 2020. № 2. С. 52-58.
4. Д. М. Лосева. Интеллектуальный анализ текстовых данных для решения задачи категоризации информации//Международная научная конференция по проблемам управления в технических системах. 2021. Т. 1. С. 291-293.
5. Л. С. Звягин. Интеллектуальный анализ данных: Big Data и Data Science//Мягкие измерения и вычисления. 2022. Т. 54. № 5. С. 81-90.
6. Ji Pu. A fuzzy intelligent group recommender method in sparse-data environments based on multi-agent negotiation//Expert systems with applications. 2023. Vol. 213. Number PC. P. 99-112.
7. Advances in scalable and intelligent geospatial analytics: challenges and applications/Edited by Surya Durbha [and six others]. Boca Raton: CRC Press, 2023. 428 p.
8. Intelligent image and video analytics: clustering and classification applications/Edited by El-Sayed M. El-Alfy, George Bebis, Mengchu Zhou. Boca Raton: CRC Press, 2023. 368 p.
9. Cloud-based intelligent informative engineering for Society 5.0/Edited by Kaushal Kishor, Neetesh Saxena, Dilkeshwar Pandey. Boca Raton: Chapman & Hall/CRC, 2023. 232 p.

References

1. Liu Xiaoming. Study on Intelligent Analysis and Processing Technology of Computer Big Data Based on Clustering Algorithm//Recent advances in electrical & electronic engineering. 2023. Vol. 16. № 2. P. 150-158.
2. Advances in intelligent data analysis XXI: 21st International Symposium on Intelligent Data Analysis, IDA 2023, Louvain-la-Neuve, Belgium, April 12-14, 2023 proceedings/Edited by Bruno Crémilleux, Sibylle Hess, Sigfried Nijssen. Cham: Springer, 2023. 499 p.
3. S. V. Palmov, A. A. Diyazitdinova, E. S. Artyushkina. Comparative analysis of the capabilities of intelligent systems in identifying hidden patterns in data//Elektrosvyaz. 2020. № 2. P. 52-58.
4. D. M. Loseva. Text data mining for solving the problem of information categorization//International scientific conference on control problems in technical systems. 2021. Vol. 1. P. 291-293.
5. L. S. Zvyagin. Data Mining: Big Data and Data Science//Soft Measurements and Computing. 2022. V. 54. № 5. P. 81-90.
6. Ji Pu. A fuzzy intelligent group recommender method in sparse-data environments based on multi-agent negotiation//Expert systems with applications. 2023. Vol. 213. Number PC. P. 99-112.
7. Advances in scalable and intelligent geospatial analytics: challenges and applications/Edited by Surya Durbha [and six others]. Boca Raton: CRC Press, 2023. 428 p.
8. Intelligent image and video analytics: clustering and classification applications/Edited by El-Sayed M. El-Alfy, George Bebis, Mengchu Zhou. Boca Raton: CRC Press, 2023. 368 p.
9. Cloud-based intelligent informative engineering for Society 5.0/Edited by Kaushal Kishor, Neetesh Saxena, Dilkeshwar Pandey. Boca Raton: Chapman & Hall/CRC, 2023. 232 p.