ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2025 Управление, вычислительная техника и информатика № 73

Tomsk State University Journal of Control and Computer Science

Научная статья
УДК 004.738
doi: 10.17223/19988605/73/12

Подход к выделению значимых признаков сетевой активности устройств Интернета вещей

Ольга Сергеевна Исаева[1], Сергей Владиславович Исаев[2], Никита Владимирович Кулясов[3]

[1, 2, 3] *Институт вычислительного моделирования Сибирского отделения Российской академии наук, Красноярск, Россия*
[1] *isaeva@icm.krasn.ru*
[2] *si@icm.krasn.ru*
[3] *razor@icm.krasn.ru*

Аннотация. Исследуются признаки сетевой активности устройств Интернета вещей и предлагается метод сокращения размерности признакового пространства для повышения эффективности анализа данных. Предложенный подход устраняет мультиколлинеарность, нелинейную зависимость и избыточность признаков, сохраняя их семантическую интерпретируемость. В его основе лежит комбинированное использование статистических характеристик взаимной информации, корреляции, критериев стабильности и значимости для фильтрации признаков. Применение подхода позволило существенно сократить признаковое пространство и улучшить его свойства: численную устойчивость данных, обобщающую способность моделей, качество кластеризации.

Ключевые слова: Интернет вещей; датасет сетевых угроз; статистический анализ; устойчивость признаков; веса признаков в методе главных компонент.

***Благодарности*:** Работа поддержана Красноярским математическим центром, финансируемым Минобрнауки РФ в рамках мероприятий по созданию и развитию региональных НОМЦ (соглашение 075-02-2025-1606).

***Для цитирования*:** Исаева О.С., Исаев С.В., Кулясов Н.В. Подход к выделению значимых признаков сетевой активности устройств Интернета вещей // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2025. № 73. С. 100–109. doi: 10.17223/19988605/73/12

Original article
doi: 10.17223/19988605/73/12

Approach to identifying significant features of IoT device network activity

Olga S. Isaeva[1], Sergey V. Isaev[2], Nikita V. Kulyasov[3]

[1, 2, 3] *Institute of Computational Modelling SB RAS, Krasnoyarsk, Russian Federation*
[1] *isaeva@icm.krasn.ru*
[2] *si@icm.krasn.ru*
[3] *razor@icm.krasn.ru*

Abstract. The study investigates the features of network activity in Internet of Things (IoT) devices and proposes a method for reducing the dimensionality of the feature space to enhance data analysis efficiency. The proposed approach eliminates multicollinearity, nonlinear dependencies, and feature redundancy while preserving their semantic interpretability. It is based on the combined use of statistical characteristics such as mutual information, correlation, stability criteria, and significance measures for feature filtering. Applying this approach significantly reduced the feature space and improved its properties: numerical stability of the data, generalization ability of models, clustering quality.

Keywords: Internet of Things; threat datasets; statistical analysis; feature stability; feature weights in principal component analysis.

***Acknowledgments*:** This work is supported by the Krasnoyarsk Mathematical Center and financed by the Ministry of Science and Higher Education of the Russian Federation in the framework of the establishment and development of regional Centers for Mathematics Research and Education (Agreement No. 075-02-2025-1606).

***For citation*:** Isaeva, O.S., Isaev, S.V., Kulyasov, N.V. (2025) Approach to identifying significant features of IoT device network activity. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 73. pp. 100–109. doi: 10.17223/19988605/73/12

Введение

Среди ключевых трендов развития современного общества широкую популярность набирает концепция Интернета вещей (Internet of Things, IoT), направленная на построение киберфизических систем, объединяющих физические и цифровые объекты на основе информационно-коммуникационных технологий [1]. Его применение становится актуальным в различных сферах жизни: от контроля качества продукции до прогнозирования отказов технических систем и мониторинга состояния окружающей среды. Однако значительные преимущества, предоставляемые технологиями Интернета вещей, требуют усиленного внимания к вопросам безопасности и надежности связанных с обеспечением их функционирования процессов, которые являются привлекательной целью для кибератак из-за разнообразия и уязвимости устройств и протоколов их подключения [2]. Особенностью сетевого трафика Интернета вещей является изменчивость, вызванная множественностью происходящих событий, сценариев атак и их последствий, зависящих от развития средств уклонения. Сложность диагностирования аномалий увеличивается по мере роста числа взаимосвязанных систем и разнообразия типов входных данных [3]. Постоянное изменение характеристик нормального поведения затрудняет автоматическое обнаружение аномалий, которые могут возникать из-за различных факторов, таких как отказ используемых устройств или внешняя атака, что требует не только их выявления, но и классификации видов и причин деструктивных воздействий. Методы машинного обучения позволяют получать дополнительную информацию о происходящих в сетевой инфраструктуре событиях безопасности, а также предотвращать нежелательные инциденты. Но для их эффективной работы необходимы специализированные датасеты, охватывающие длительные периоды наблюдения и содержащие сведения о характерных видах атак, типичном поведение пользователей и свойствах реальных сетей, в рамках которых строится система обнаружения вторжений.

Актуальность применения методов машинного обучения для обеспечения сетевой безопасности широко освещается в современной научной литературе. В работах [4–5] представлен детальный обзор популярных наборов данных, используемых для выявления сетевых вторжений, а также предложены критерии оценки их применимости для решения практических задач. В [6] рассматривается проблема несбалансированности данных публичных датасетов. Предложен метод синтезирования и сжатия выборок для уменьшения дисбаланса классов. Важным аспектом анализа сетевых данных являются извлечение признаков из исходных файлов и их последующее сопоставление с размеченными данными [7]. При этом исследования отмечают наличие ряда ограничений в публичных датасетах, таких как использование искусственных имитационных сред, неоднородность данных, ошибки в расчетах значений признаков, их дублирование и некорректное разбиение на сессии. Эти факторы существенно влияют на качество моделей машинного обучения. В [8] отмечается важность учета контекста, включающего не только архитектурные особенности сети, но и свойства протоколов, используемых на отдельных уровнях межмашинного взаимодействия.

Для преодоления ограничений, присущих публичным датасетам, авторами настоящего исследования была разработана и внедрена инфраструктура сбора данных и имитации угроз безопасности для сети Интернета вещей Красноярского научного центра [9]. После организации сбора данных встала задача исследования признакового пространства для сокращения его размерности, исключения избыточности, повышения обобщающей способности моделей, улучшения их точности, устойчивости и интерпретируемости результатов. Существующие подходы к решению этой задачи можно разделить на четыре группы: встраиваемые (Embedded), обертывающие (Wrapper), фильтрующие (Filtered) и действующие на основе экспертных оценок. Встраиваемые методы интегрируют выбор признаков в процесс обучения, что позволяет использовать скрытую структуру данных. Однако такой подход потребует

обработки не только множественности признаков, но и данных, генерируемых IoT-устройствами. Обертывающие методы используют алгоритмы роевого интеллекта для эвристического обхода комбинаций признаков и выбора их подмножества, соответствующего целевой функции [10]. Такой способ поиска является NP-сложной задачей и также не подходит для крупномасштабных данных Интернета вещей. Процесс фильтрации признаков заключается в выявлении наиболее информативных и значимых из них. Они конструируют критерии ранжирования признаков с помощью статистических характеристик, корреляции, взаимной информации или условной энтропии [11]. Методы экспертных оценок выполняют отбор признаков на основе знаний о предметной области, но при большом размере признакового пространства их необходимо интегрировать с методами фильтрации.

Цель данной работы – анализ признаков сетевой активности IoT-устройств для уменьшения размерности признакового пространства за счет применения фильтрации, основанной на критериях стабильности и информационной значимости. Подход, представленный в исследовании, учитывает особенности функционирования реальной сети Интернета вещей, но может быть обобщен на другие типы сетевого трафика. Работа является важным шагом к построению системы информационной безопасности, специализированной для исследуемой корпоративной сети, учитывающей ее архитектурные и технические особенности.

1. Особенности сбора данных сетевой активности

В рамках исследований, проводимых в Красноярском научном центре, технология Интернета вещей была внедрена для мониторинга микроклимата в помещениях, где размещено сетевое оборудование [12]. Для анализа безопасности IoT-сети в реальных условиях эксплуатации созданы инструменты сбора данных и имитации угроз. Межмашинное взаимодействие строится на основе протокола прикладного уровня MQTT (Message Queuing Telemetry Transport). На рис. 1 приведена схема размещения основных узлов инфраструктуры сбора данных сетевой активности устройств Интернета вещей.

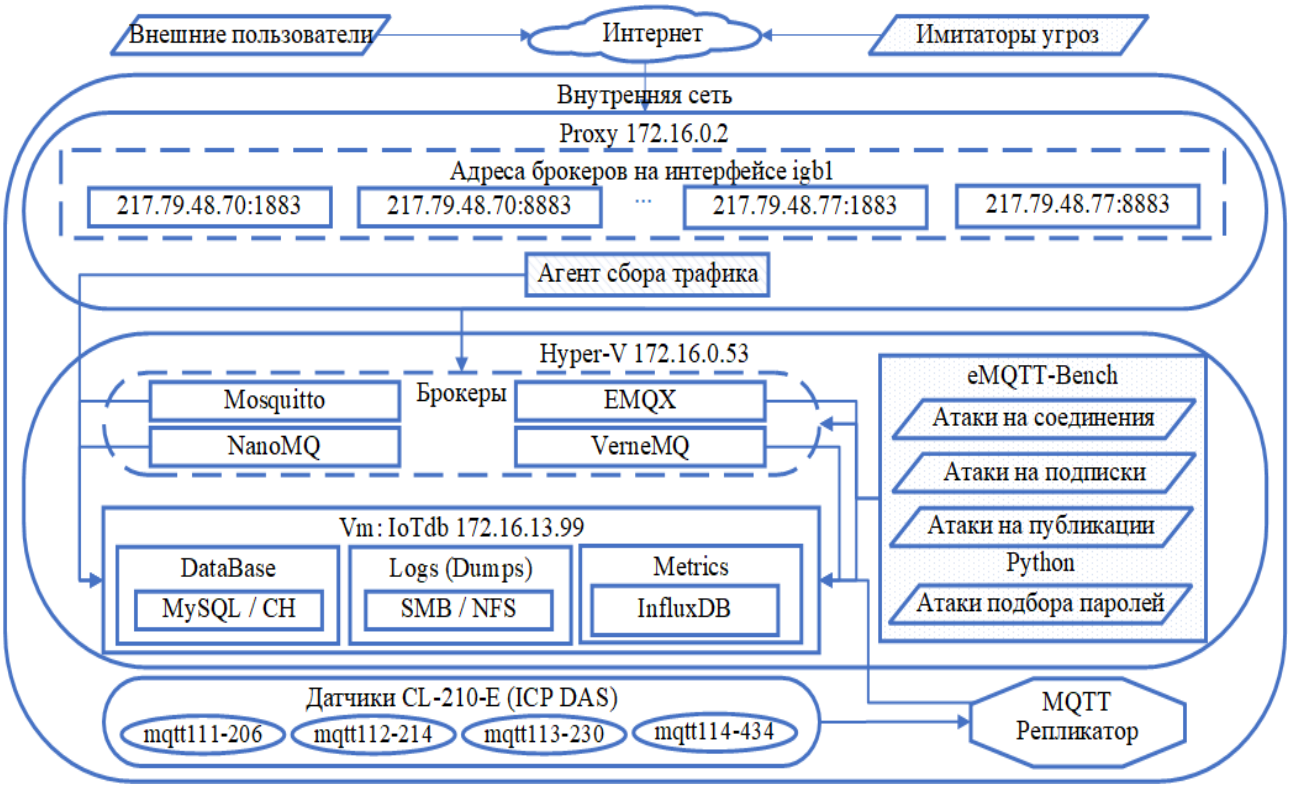

Рис. 1. Инфраструктура сбора данных устройств Интернета вещей
Fig. 1. Infrastructure for collecting data from IoT devices

Основные элементы сети: издатель (Publisher), производящий данные, подписчик (Subscriber), их использующий, и брокер (Broker), выполняющий распределение данных от издателей подписчикам. Использование технологии IoT в сочетании с протоколом MQTT позволяет создать эффективную систему мониторинга, способную оперативно обрабатывать большие потоки данных в темпе их поступления. В рамках исследования были развернуты брокеры на платформах Eclipse Mosquitto, EMQX, NanoMQ и VerneMQ. Сбор данных осуществлялся как на самих брокерах, так и на proxy-сервере, через который передается внутренний и внешний сетевой трафик по портам протокола MQTT, включая легитимные сессии, а также несанкционированные попытки соединения и сканирования различных сервисов.

Собранные сетевые данные являются неструктурированными, и для их эффективного использования требуется этап предобработки, включающий разделение трафика на сессии и вычисление параметров полученных сессий. Для этого мы выбрали программное обеспечение с открытым кодом NTLFlowLyzer, которое использовано [13] при построении публичного датасета BCCC-CIC-IDS201. Данный инструмент является усовершенствованной версией сетевого анализатора CICFlowMeter, лежащего в основе популярного датасета CICIDS2017. Мы выполнили проверку параметров выделяемых сессий для сетевых журналов, собранных в рамках нашего исследования. Для решения специфических задач, связанных с Интернетом вещей, в наших данных содержатся метрики брокеров и флаги используемых протоколов [14]. В сформированный датасет вошло более 300 признаков, охватывающих шесть ключевых категорий: временные характеристики, флаги протоколов TCP и MQTT, параметры скорости соединений, статистические данные по заголовкам пакетов, свойства полезной нагрузки и объемные характеристики при массовой передаче данных.

Дальнейшее исследование выявило ряд проблем, существенно затрудняющих применение собранных данных для анализа аспектов безопасности Интернета вещей. К таким проблемам относятся разноплановость представленной информации для типичных сессий и высокоразмерность признакового пространства, которая приводит к экспоненциальному росту объема данных, необходимого для получения надежных результатов моделей машинного обучения, а также наличие избыточных и слабоинформативных признаков, которое снижает эффективность алгоритмов кластеризации и увеличивает их временную сложность. Еще одной значимой проблемой является высокая корреляция признаков, которая искажает структуру данных и приводит к несогласованным результатам кластеризации.

Использование методов снижения размерности, заключающееся в формировании новых признаков, ограничивает возможность семантической интерпретации взаимосвязей между объектами в многомерных данных. Традиционные подходы к сокращению признакового пространства, такие как рекурсивное исключение признаков (RFE) [15], адаптируют набор признаков под конкретный класс моделей, однако не учитывают сложные линейные и нелинейные зависимости между признаками, что сказывается на качестве получаемых подпространств. Стандартные метрики оценки качества моделей (например, MSE, MAE, точность, F1-мера и др.) не могут быть применены в нашем случае, поскольку они предполагают сравнение предсказаний модели с известными целевыми значениями. В нашей задаче требуется первоначально построить целевой показатель путем кластеризации данных, что затруднительно ввиду большого признакового пространства. Таким образом, необходим механизм направленной фильтрации признаков, обеспечивающий сохранение обобщающих свойств и удовлетворяющий выбранным критериям качества данных.

2. Сокращение признакового пространства

Для решения вышеописанных проблем предложен метод сокращения размерности признакового пространства, выделяющий сложные зависимости между признаками с сохранением его информативности. Введем обозначения, необходимые для описания подхода:

$$X = \left[x_{ij} \right] \in \mathbb{R}^{n \times m}, \tag{1}$$

где X – матрица наблюдений размерности $n \times m$, n – количество объектов наблюдения, m – количество признаков, x_{ij} – значение j-го признака для i-го объекта наблюдения, $I = \{1, 2, \ldots, n\}$ – множество иден-

тификаторов, соответствующих каждому объекту наблюдения, $J = \{1, 2, \ldots, m\}$ – множество идентификаторов, соответствующих каждому признаку. Каждый объект наблюдения $i \in I$ описывается вектором $X_i = \left(x_{i1}, x_{i2}, \ldots, x_{im}\right) \in \mathbb{R}^m$ значений всех признаков для *i*-го объекта. Признак $j \in J$ представлен вектором $P_j = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right) \in \mathbb{R}^n$ значений *j*-го признака для всех объектов.

$$X = \begin{bmatrix} X_1 & X_2 & \ldots & X_n \end{bmatrix} = \begin{bmatrix} P_1 & P_2 & \ldots & P_m \end{bmatrix}^{\mathrm{T}}. \tag{2}$$

Для центрированной матрицы X_c можно построить квадратную матрицу $\Sigma_X = {X_c}^{\mathrm{T}} X_c$ размерности $m \times m$ (здесь и далее мы не умножаем на нормировочный коэффициент, поскольку данные предварительно стандартизованы) и вычислить $\lambda_1, \lambda_2, ..., \lambda_m$ – собственные числа, $v_1, v_2, ..., v_m$ – собственные векторы, $\delta_1, \delta_2, ..., \delta_m$, $\delta_i = \sqrt{\lambda_i}$ – сингулярные числа.

Введем целевую функцию, описывающую требования к признаковому пространству:

$$L(X) = \alpha_1 \cdot \kappa\left(X\right) + \alpha_2 \cdot M\left(X\right) + \sum_{l=1}^{L} \alpha_{3,l} \cdot R\left(F_l, X\right) \tag{3}$$

Весовые коэффициенты α позволяют изменять влияние каждого компонента на результат целевой функции, $\kappa\left(X\right) = \delta_{\max} / \delta_{\min}$ – число обусловленности, $\delta_{\max}$ – наибольшее, а $\delta_{\min} \neq 0$ – наименьшее сингулярные числа, показывают наличие линейной зависимости и мультиколлинеарности данных, $M\left(X\right)$ выполняет оценку меры нелинейной зависимости признаков:

$$M\left(X\right) = \frac{2}{m(m-1)} \sum_{j=1}^{m} \sum_{k=j+1}^{m} M_{jk}, \quad M_{jk} = \sum_{x_{ij} \in P_j} \sum_{x_{ik} \in P_k} \frac{p(x_{ij}, x_{ik})}{n} \cdot \ln \frac{n \cdot p(x_{ij}, x_{ik})}{p(x_{ij}) \cdot p(x_{ik})}, \tag{4}$$

где M_{jk} определяет величину взаимной информации между P_j и P_k [16], $p(x_{ij}, x_{ik})$ – частота совместного появления значений признаков P_j и P_k, $p(x_{ij})$ и $p(x_{ik})$ – частота появления каждого значения признака в отдельности.

Значение $R\left(F_l, X\right)$ – мера сложности Радемахера [17] для семейства функций F_l относительно выборки *X*:

$$R^{(l)}\left(F_l, X\right) = \sup_{f \in F_l} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f\left(X_i\right), \tag{5}$$

где ε_i – случайная величина, F_l – семейство функций вида *f*. Это теоретическая мера, которая показывает способность аппроксимировать данные и избегать переобучения.

Пусть *A* – множество действий по изменению признакового пространства *P*. На каждом шаге $h \in S^h = \{0, 1, \ldots, H\}$ действие $A^h \in A$ применяется к множеству признаков так, что $A^h\left(P^{h-1}\right) = P^h$. Тогда итоговое множество признаков P^H получается следующим образом:

$$P^H = \arg\min_{h \in S^h} \left(L\left(P^h\right) \middle| A^h \in A, P^h = A^h\left(P^{h-1}\right), L\left(P^h\right) \leq L\left(P^{h-1}\right) \right), \tag{6}$$

при условии: $P_j \in \mathbb{R}^m, m \leq m_{\max} \ \wedge \ \sum_{i=1}^{n} \sum_{j}^{m} \mathrm{I}\left(x_{ij} = \varnothing\right) = 0 \ \wedge \ \forall j \ \sigma_j \geq \tau_\sigma$, где $m_{\max}$ – ограничение на размерность пространства, $\mathrm{I}\left(\cdot\right)$ – индикационная функция, σ_j – стандартное отклонение, τ_σ – пороговое значение. Действия из множества *A* заключаются в фильтрации признакового пространства P^h размерности m^h и удалении из него признаков, удовлетворяющих следующим условиям:

I. Подмножество коррелирующих признаков:

$$C^{\mathrm{I}} = \left\{ C^{j\{k\}} \right\}_{j,k} = \bigcup_{1 \leq j \leq k \leq m^h} \left\{ \left(P_j, P_k\right) \in P^h : \left| \rho_{jk} \right| \geq \tau_\rho \right\}, \tag{7}$$

где $P_j, P_k \in P^h$, m^h – число признаков в P^h, ρ_{jk} – коэффициент корреляции между признаками, τ_ρ – пороговое значение корреляции.

II. Подмножество мультиколлинеарных признаков:

$$C^{\mathrm{II}}=\left\{C^{j\{k\}}\right\}_{j,k}=\bigcup_{k\neq j}\left(P_j\cup\left\{P_k\in P^h:\ |\beta_k|>\tau_\beta,\ V_k\geq\tau_V\right\}\right), \tag{8}$$

$P_j\in P^h=\arg\max_k\left(V_k\geq\tau_V\right)$, τ_V – граница мультиколлинеарности, V_k – коэффициент инфляции дисперсии (Variance Inflation Factor; VIF), который показывает, насколько сильно взаимная корреляция между признаками увеличивает дисперсию:

$$V_k=\left(1-D^2\right)^{-1},\ D^2=\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2\Big/\sum_{i=1}^{n}\left(x_{ij}-x_{ij}\right)^2, \tag{9}$$

где D^2 – коэффициент детерминации, $\bar{x}_j$ – среднее значение, $\tilde{x}_{ij}=\beta_0+\sum_{k\neq j}x_{ik}\beta_k$ – предсказанное значение, вычисленное через разложение $P_j=\beta_0+\beta_1P_1+\beta_2P_2+\cdots+\beta_mP_m$, $\beta_k=\rho_{jk}\left(\sigma_j/\sigma_k\right)$, ρ_{jk} – коэффициент корреляции, σ_j – стандартное отклонение.

III. Подмножество взаимозависимых признаков:

$$C^{\mathrm{III}}=\left\{C^{j\{k\}}\right\}_{j,k}=\bigcup_{1\leq j\leq k\leq m^h}\left\{\left(P_j,P_k\right)\in P^h:\ \left|M_{jk}\right|\geq\tau_M\right\}, \tag{10}$$

где M_{jk} вычисляется по (5), τ_M – порог сильной зависимости.

Для множества $C=C^{\mathrm{I}}\bigcup C^{\mathrm{II}}\bigcup C^{\mathrm{III}}$ вводятся критерии исключения признаков. Признак $P\in C^{j\{k\}}\subseteq C$ исключается из P^h и $C=C\setminus C^{j\{k\}}$, если выполняется условие

$$\forall Q\in C^{j\{k\}}\setminus P:\ \left(S(P)<S(Q)\right)\vee\left(W(P)<W(Q)\right), \tag{11}$$

где $S(P)<S_{stab}$ – коэффициент стабильности, S_{stab} – порог, $W(P)$ – агрегированный вес признака.

Коэффициент стабильности вычисляется как

$$S(P)=\frac{1}{T-1}\sum_{t=1}^{T-1}M\left(P^{(t)},P^{(t+1)}\right), \tag{12}$$

где T – количество случайных выборок, полученных из X, $M\left(P^{(t)},P^{(t+1)}\right)$ – взаимная информация (4) между признаком P на t-й и $(t+1)$-й выборках.

Агрегированный вес признака определятся как

$$W(P)=\sum_{i=1}^{k}\left|v_i(P)\right|\cdot\frac{\lambda_i}{\sum_{m=1}^{k}\lambda_m}, \tag{13}$$

где $k=\min\left\{\,j\,\middle|\left(\sum_{i=1}^{j}\lambda_i\Big/\sum_{i=1}^{m}\lambda_j\right)\geq\varsigma\right\}$ – глубина редуцированного пространства, содержащего ς дисперсии исходного набора данных, λ_i – собственное значение, $v_i(P)$ – компонента собственного вектора v_i, соответствующая признаку P. Действия по выбору и фильтрации признаков выполняются до тех пор, пока не останется групп для рассмотрения или признаков по критерию (11).

3. Применение подхода к данным сетевой активности

Предложенный подход был апробирован на данных сетевой активности устройств Интернета вещей, собранных в корпоративной сети научного центра. В рамках анализа данных были построены диаграммы рассеивания, демонстрирующие зависимости между парами признаков. На рис. 2 представлен пример визуализации, выполненный на логарифмической шкале, для признаков «Длительность потока (сек)» и «Объем переданных данных (байт)». Для каждой пары признаков добавлена тепловая карта плотности, показывающая распределение точек на графике, и построена аппроксимация данных с использованием полиномиальных или линейных моделей. Визуализация позволила оценить характер взаимосвязей между признаками, а также выделить аномалии в данных.

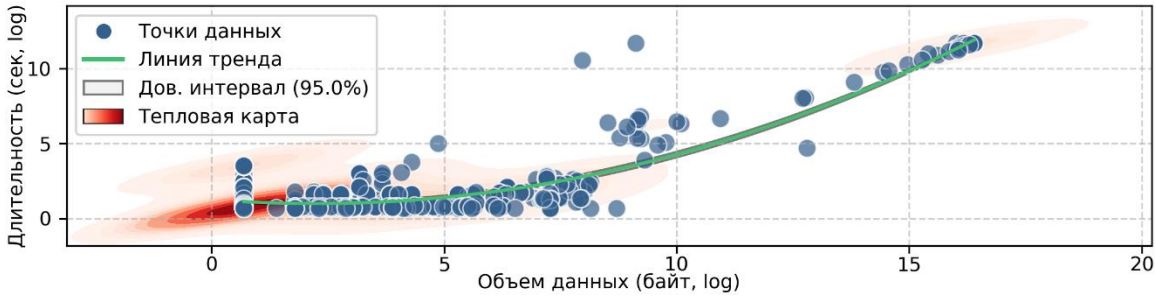

Рис. 2. Пример зависимых признаков
Fig. 2. Example of dependent features

Сформировано множество $C = C^{\mathrm{I}} \bigcup C^{\mathrm{II}} \bigcup C^{\mathrm{III}}$ зависимых признаков по формулам (7)–(10). Для каждого признака из полученного множества рассчитан коэффициент стабильности по формуле (12) и исключены признаки, удовлетворяющие критерию (11), не сохраняющие стабильность своих характеристик на различных подвыборках данных и демонстрирующие существенную зависимость от более устойчивых признаков.

Для определения весовых коэффициентов признаков проведен анализ собственных значений и накопительной объясненной дисперсии (рис. 3). Установлено минимальное количество главных компонент, достаточное для описания данных (95% объясненной дисперсии). Вычислены весовые коэффициенты признаков в этих компонентах. Признаки, которые не вносят значимый вклад в формирование главных компонент или имеют меньший вес по сравнению с другими зависимыми от них признаками, согласно критерию (13), были удалены из рассмотрения. На каждом шаге вычисляется значение целевой функции (3), и результаты изменения признакового пространства, не удовлетворяющие условию (6), не принимаются.

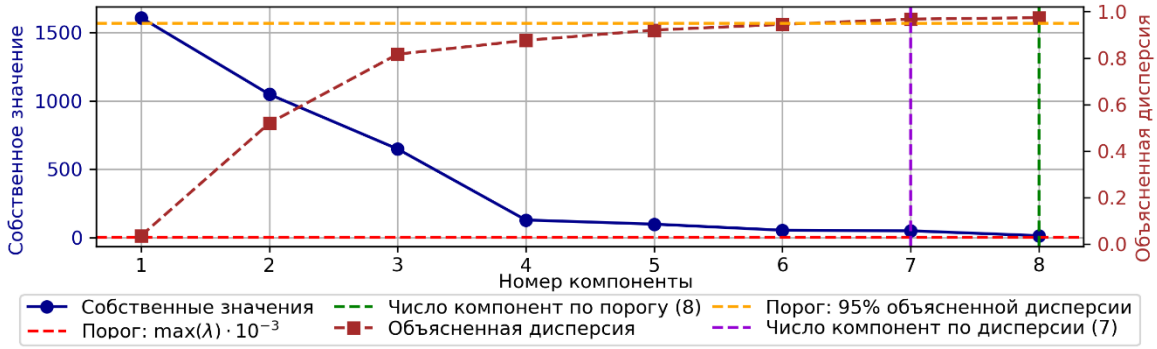

Рис. 3. Определение числа компонент для расчёта весов признаков
Fig. 3. Determining the number of components for calculating feature weights

В таблице приведены пошаговые результаты применения предложенного подхода и их сравнение с методом рекурсивного исключения признаков (RFE) для модели случайного леса.

Пошаговые результаты сокращения размерности

Выполненные действия	Признаков	Число обусл.	Сложность по Радермахеру		
			Лог. регресс.	Случ. лес	Лин. мод., случ. вес
0. Предобработка	275	768,26	0,264	0,199	0,417
1. Корр., нестабильные	223	450,6	0,253	0,196	0,415
2. Корр., незначимые	79	90,4	0,181	0,191	0,233
3. Мультикол., нестабильные	25	17,08	0,108	0,163	0,145
4. Взаимные, нестабильные	22	12,5	0,099	0,143	0,117
5. Взаимные, незначимые	17	2,23	0,053	0,079	0,087
Время выполнения: 6,72 с (~ 1 000 строк), 58,46 с (~ 10 200 строк). Корр.: 0. Взаимозависимые: 0					
RFE (случ. лес)	17	75,17	0,097	0,182	0,083
Время выполнения: 41,42 с (~ 1 000 строк), 147,42 с (~10 200 строк). Корр.: 8. Взаимозависимые: 56					

Применение подхода к сокращению признакового пространства демонстрирует не только значительное сокращение размерности, но и улучшение ключевых характеристик данных. Снижение числа обусловленности свидетельствует о повышении численной устойчивости данных, что может быть дополнительно усилено за счет регуляризации сокращенной матрицы. Заметное уменьшение сложности моделей, описывающих данные, способствует снижению риска переобучения и улучшению их обобщающей способности.

Выполнено сравнение результатов кластеризации для полного и обработанного признакового пространства. На рис. 4 приведены графики силуэтного коэффициента, которые показывают повышение компактности кластеров и их лучшую разделенность.

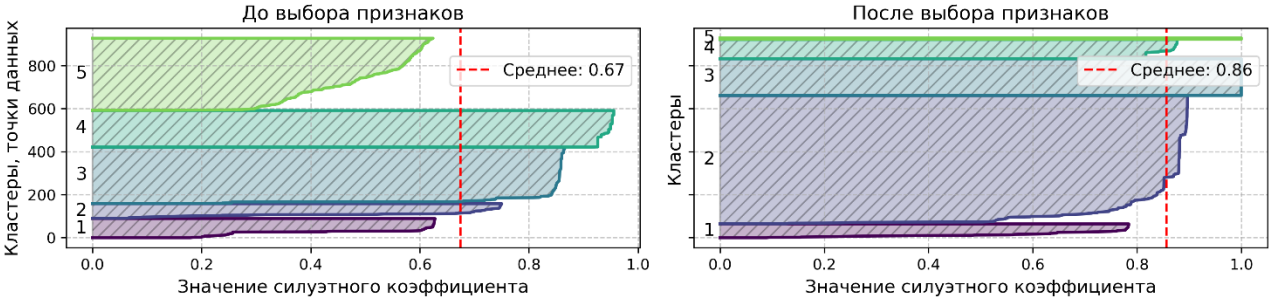

Рис. 4. Сравнение силуэтного коэффициента
Fig. 4 Comparison of silhouette coefficient

В результате анализа объектов, отнесенных к выделенным кластерам, было установлено, что сокращение размерности признакового пространства обеспечило возможность описать характеристики сформированных групп и предложить объяснения для наблюдаемого разбиения. Первый кластер объединяет попытки установления соединений без получения ответа от принимающих устройств. Второй содержит легитимные сессии, сопровождаемые передачей данных. Третий также включает легитимные сессии, но значительной длительности, характеризующие взаимодействие с устройствами интернета вещей. Четвёртый кластер состоит из запросов на соединение и коротких сессий, соответствующих сканированию сети на транспортном уровне без детализации по прикладному протоколу. Пятый включает короткие соединения, представляющие собой более глубокое сканирование сети на уровне прикладного протокола, в отличие от поверхностного анализа, характерного для предыдущего кластера. Таким образом, проведенное исследование позволило повысить степень интерпретируемости данных.

Заключение

Предложенный в исследовании подход направлен на решение ключевых проблем, возникающих при работе с высокоразмерными данными, таких как избыточность признаков, наличие корреляций и сложных нелинейных зависимостей. Он позволяет эффективно сокращать размерность признакового пространства, одновременно улучшая численные характеристики данных и снижая сложность моделей машинного обучения. Основные преимущества подхода включают:

1. Эффективное устранение избыточных зависимостей, достигаемое за счет комбинированного использования взаимной информации и анализа корреляций.
2. Обеспечение семантической интерпретируемости данных за счет сохранения исходных признаков и связей между объектами.
3. Улучшение характеристик кластеризации и обобщающей способности моделей благодаря уменьшению нестабильности признаков и снижению сложности.

Существенное сокращение размерности признакового пространства без потери информативности делает данный подход применимым для широкого спектра задач – от предварительной обработки данных до повышения эффективности алгоритмов классификации, регрессии и кластеризации.

Список источников

1. Курбатов В.И., Папа О.М. Интернет вещей: основные концепции и тренды // Гуманитарные, социально-экономические и общественные науки. 2023. № 1. С. 48–54.
2. Минаев В.А., Швырев Б.А., Ромашкин Т.Р. Безопасность интернета вещей: основные решения // Информация и безопасность. 2023. Т. 26 (2). С. 163–168.
3. Chatterjee A., Ahmed B.S. IoT anomaly detection methods and applications: A survey // Internet of Things. 2022. V. 19. Art. 100568. doi: 10.1016/j.iot.2022.100568
4. Ring M., Wunderlich S., Scheuring D., Landes D., Hotho A. A survey of network-based intrusion detection data sets // Computers & Security. 2019. V. 86. P. 147–167. doi: 10.1016/j.cose.2019.06.005
5. Xinpeng C. CICIDS2017 and UNBSW-NB15 // IEEE Dataport. 2023. doi: 10.21227/ykpn-jx78. URL: https://ieee-dataport.org/documents/cicids2017-and-unbsw-nb15
6. Liu L., Wang P., Lin J., Liu L. Intrusion detection of imbalanced network traffic based on machine learning and deep learning // IEEE Access. 2021. V. 9. P. 7550–7563. doi: 10.1109/ACCESS.2020.3048198
7. Moustafa N. A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets // Sustainable Cities and Society. 2021. V. 72. Art. 102994.
8. Горюнов М.Н., Мацкевич А.Г., Рыболовлев Д.А. Синтез модели машинного обучения для обнаружения компьютерных атак на основе набора данных CICIDS2017 // Труды Института системного программирования РАН. 2020. Т. 32 (5). С. 81–94. doi: 10.15514/ISPRAS–2020–32(5)–6.
9. Исаева О.С., Кулясов Н.В., Исаев С.В. Инфраструктура сбора данных и имитации угроз безопасности сети интернета вещей // Сибирский аэрокосмический журнал. 2025. Т. 26, № 1. С. 8–20. doi: 10.31772/2712-8970-2025-26-1-8-20
10. Altarabichi M.G., Nowaczyk S., Pashami S., Mashhadi P.S. Surrogate-assisted genetic algorithm for wrapper feature selection // IEEE Congress on Evolutionary Computation. 2021. P. 776–785. doi: 10.1109/CEC45853.2021.9504718
11. Liu S., Motani M. Improving mutual information based feature selection by boosting unique relevance // Journal of Artificial Intelligence Research. 2025. V. 82. P. 1267–1292. doi: 10.1613/jair.1.17219
12. Исаева О.С. Построение цифрового профиля устройств Интернета вещей // Информационные и математические технологии в науке и управлении. 2023. № 2 (30). С. 36–44. doi: 10.25729/ESI.2023.30.2.004
13. Shafi M., Lashkari A., Roudsari A. NTLFlowLyzer: Towards generating an intrusion detection dataset and intruders behavior profiling through network and transport layers traffic analysis and pattern extraction // Computers & Security. 2025. V. 148. Art. 104160. doi: 10.1016/j.cose.2024.104160
14. Исаева О.С. Построение онтологии для систематизации характеристик сети Интернета вещей // Онтология проектирования. 2024. Т. 14, № 2 (52). С. 243–255. doi: 10.18287/2223- 9537-2024-14-2-243-255
15. Priyatno A.M., Widiyaningtyas T. A systematic literature review: recursive feature elimination algorithms // JITK. 2024. V. 9 (2). P. 196–207. doi: 10.33480/jitk.v9i2.5015
16. Цурко В.В., Михальский А.И. Оценка статистической связи случайных величин через взаимную информацию // Автоматика и телемеханика. 2022. Вып. 5. С. 76–86. doi: 10.31857/S0005231022050063
17. Koltchinskii V. Rademacher complexities and bounding the excess risk // Journal of Machine Learning Research. 2010. V. 11. P. 2457–2485.

References

1. Kurbatov, V.I. & Papa, O.M. (2023) Internet of things: main concepts and trends. *Gumanitarnye, sotsial'no-ekonomicheskie i obshchestvennye nauki.* 1. pp. 48–54.
2. Minaev, V.A., Shvyrev, B.A. & Romashkin, T.R. (2023) Internet of things security: main solutions. *Informatsiya i bezopasnost' – Information and Security*. 26(2). pp. 163–168.
3. Chatterjee, A. & Ahmed, B. S. (2022) IoT anomaly detection methods and applications: a survey. *Internet of Things*. 19. Art. 100568. doi: 10.1016/j.iot.2022.100568.
4. Ring, M., Wunderlich, S., Scheuring, D., Landes, D. & Hotho A. (2019) A survey of network-based intrusion detection data sets. *Computers & Security*. 86. pp. 147–167. doi: 10.1016/j.cose.2019.06.005
5. Xinpeng, C. (2023) CICIDS2017 and UNBSW-NB15. *IEEE Dataport*. doi: 10.21227/ykpn-jx78.
6. Liu, L., Wang, P. & Lin, J. (2021) Intrusion detection of imbalanced network traffic based on machine learning and deep learning. *IEEE Access*. 9. pp. 7550–7563. doi: 10.1109/ACCESS.2020.3048198
7. Moustafa, N. (2021) A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets, *Sustainable Cities and Society*. 72. Art. 102994.
8. Goryunov, M.N., Matskevich, A.G. & Rybolovlev, D.A. (2020) Synthesis of a machine learning model for detecting computer attacks based on the CICIDS2017 dataset. *Trudy ISP RAN*. 32(5). pp. 81–94. doi: 10.15514/ISPRAS–2020–32(5)–6
9. Isaeva, O.S., Kulyasov, N.V. & Isaev, S.V. (2025) Infrastructure for collecting data and simulating security threats to the internet of things. *Sibirskiy aerokosmicheskiy zhurnal – Siberian Aerospace Journal*. 26(1). pp. 8–20. doi: 10.31772/2712-8970-2025-26-1-8-20

10. Altarabichi, M.G., Nowaczyk, S., Pashami, S. & Mashhadi, P.S. (2021) Surrogate-assisted genetic algorithm for wrapper feature selection. *IEEE Congress on Evolutionary Computation*. pp. 776–785. DOI: 10.1109/CEC45853.2021.9504718.
11. Liu, S. & Motani, M. (2025) Improving mutual information-based feature selection by boosting unique relevance. *Journal of Artificial Intelligence Research*. 82. pp. 1267–1292. doi: 10.1613/jair.1.17219.
12. Isaeva, O.S. (2023) Construction of a digital profile of internet of things devices. *Informatsionnye i matematicheskie tekhnologii v nauke i upravlenii – Information and Mathematical Technologies in Science and Management*. 2(30). pp. 36–44. doi: 10.25729/ESI.2023.30.2.004
13. Shafi, M., Lashkari, A. & Roudsari, A. (2025) NTLFlowLyzer: Towards generating an intrusion detection dataset and intruders' behavior profiling through network and transport layers traffic analysis and pattern extraction. *Computers & Security*. 148. Art. 104160. doi: 10.1016/j.cose.2024.104160
14. Isaeva, O.S. (2024) Construction of an ontology for systematization of characteristics of the Internet of Things network. *Ontologiya proektirovaniya – Ontology of Design*. 14(2(52)). pp. 243–255. doi: 10.18287/2223-9537-2024-14-2-243-255
15. Priyatno, A.M. & Widiyaningtyas, T. (2024) A systematic literature review: recursive feature elimination algorithms. *J I T K*. 9(2). pp. 196–207. doi: 10.33480/jitk.v9i2.5015
16. Tsurko, V.V. & Mikhalskii, A.I. (2022) Estimation of statistical connection of random variables through mutual information. *Automation and Telemechanics*. 5. pp. 76–86. doi: 10.31857/S0005231022050063
17. Koltchinskii, V. (2010) Rademacher complexities and bounding the excess risk. *Journal of Machine Learning Research*. 11. pp. 2457–2485.

Информация об авторах:
Исаева Ольга Сергеевна – доктор технических наук, старший научный сотрудник Регионального научно-образовательного математического центра «Красноярский математический центр» Института вычислительного моделирования СО РАН – обособленного подразделения ФИЦ КНЦ СО РАН (Красноярск, Россия). E-mail: isaeva@icm.krasn.ru
Исаев Сергей Владиславович – кандидат технических наук, заместитель директора Института вычислительного моделирования СО РАН – обособленного подразделения ФИЦ КНЦ СО РАН (Красноярск, Россия). E-mail: si@icm.krasn.ru
Кулясов Никита Владимирович – программист Регионального научно-образовательного математического центра «Красноярский математический центр» Института вычислительного моделирования СО РАН – обособленного подразделения ФИЦ КНЦ СО РАН (Красноярск, Россия). E-mail: razor@icm.krasn.ru

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:
Isaeva Olga S. (Doctor of Technical Sciences, Senior Researcher, Regional Mathematical Center "Krasnoyarsk Mathematical Center", Institute of Computational Modelling SB RAS – subdivision of the FRC KSC SB RAS, Krasnoyarsk, Russia). E-mail: isaeva@icm.krasn.ru
Isaev Sergey V. (Candidate of Technical Sciences, Deputy Director, Institute of Computational Modelling SB RAS – subdivision of the FRC KSC SB RAS, Krasnoyarsk, Russia). E-mail: si@icm.krasn.ru
Kulyasov Nikita V. (Programmer, Regional Scientific and Educational Mathematical Center "Krasnoyarsk Mathematical Center", Institute of Computational Modelling SB RAS – subdivision of the FRC KSC SB RAS, Krasnoyarsk, Russia). E-mail: razor@icm.krasn.ru

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Поступила в редакцию 27.04.2025; принята к публикации 02.12.2025

Received 27.04.2025; accepted for publication 02.12.2025