# 7. МЕТОДЫ ПРОГНОЗИРОВАНИЯ ИНТЕРНЕТ-ТРАФИКА

*Млёник Н.С., студент гр.273602, Семенович Е.И., студент гр.273602*

*Белорусский государственный университет информатики и радиоэлектроники*
*г. Минск, Республика Беларусь*

*Федосенко В.А. – канд. тех. наук*

**Аннотация.** В статье рассматривается задача прогнозирования интернет-трафика в условиях роста объёмов данных и сложности сетевых паттернов. Описаны методы машинного обучения: Random Forest для прогноза объёма трафика и классификации типов, SVM для выявления сложных паттернов и аномалий, DBSCAN для кластеризации и обнаружения всплесков и атак. Показана эффективность комбинированного подхода для повышения точности прогнозов и устойчивости к сетевым аномалиям.

**Ключевые слова.** интернет-трафик, прогнозирование, машинное обучение, Random Forest, SVM, DBSCAN, сетевые аномалии.

В условиях стремительного роста объёмов интернет-трафика задача его прогнозирования приобретает особое значение. Эффективное прогнозирование позволяет интернет-провайдерам и операторам сетей оптимизировать использование ресурсов, повышать качество обслуживания (QoS) и обеспечивать стабильную работу сетей. Оно также играет ключевую роль в выявлении сетевых аномалий, способствуя повышению уровня безопасности и быстрому реагированию на потенциальные угрозы.

Сложность интернет-трафика обусловлена разнообразием используемых протоколов (TCP, UDP), изменчивостью объёмов данных и высокой степенью неопределённости в поведении пользователей. Объём глобального интернет-трафика стремительно растёт из-за увеличения числа пользователей, внедрения облачных технологий, развития интернета вещей (IoT) [1]. Это создаёт значительную нагрузку на сетевую инфраструктуру.

Традиционные статистические методы прогнозирования, основанные на анализе средних значений и дисперсий, часто оказываются недостаточно точными в условиях высокой изменчивости сетевых параметров и сложности поведения трафика. Реальный интернет-трафик характеризуется всплесками, сезонными колебаниями и нелинейными зависимостями, что затрудняет его точное моделирование базовыми методами.

Эффективное прогнозирование интернет-трафика позволяет интернет-провайдерам и операторам сетей оптимизировать использование ресурсов, предотвращать перегрузки и обеспечивать стабильную работу сервисов. Это также способствует улучшению качества обслуживания (QoS) и своевременному выявлению сетевых аномалий, таких как DDoS-атаки и попытки несанкционированного доступа.

Для решения задачи прогнозирования интернет-трафика широко используются методы машинного обучения, которые позволяют выявлять сложные зависимости в данных, учитывать нелинейные паттерны и быстро адаптироваться к изменениям в сетевом поведении. В отличие от традиционных статистических методов, машинное обучение даёт возможность строить более точные модели, учитывающие временные зависимости, всплески трафика, сезонные колебания и сетевые аномалии.

В данной статье будут рассмотрены те методы прогнозирования, которые больше подходят именно для анализа интернет-трафика.

Одним из таких методов является метод **Random Forest** (алгоритм случайного леса). Данный алгоритм является универсальным. Суть его состоит в том, что он использует композицию решающих деревьев. Композиция берется так как само по себе решающее дерево не имеет высокое качество классификации.

Random Forest используется для решение чуть-ли не любых задач, которые могут быть решены с помощью машинного обучения. Однако можно выделить задачи классификации, регрессии, отбор признаков, выбросы/аномалии и кластеризации [2].

Что касается теоретической части у этого метода она достаточно простая. В основном необходима формула итогового классификатора a(x):

$$a(x) = \frac{1}{n}\sum_{i=0}^{n} b_i(x) \ , \qquad (1)$$

где n – количество деревьев, i – счетчик деревьев, b – решающее дерево, x – сгенерированное значение на основе выборки.

В начале для каждого дерева создается случайная подвыборка из базового набора с возможностью быть выбранным несколько раз или ни разу. Данный подход повышает обучаемость выборки, повышает устойчивость модели.

Далее строится дерево решений для каждой выборки таким образом, что на каждом узле дерева выбирается случайное подмножетсво признаков разделения.

После это происходит усреднение предсказаний деревьев (для регрессии) или определение наиболее частого класса (для классификации).

Данный метод может использовать в прогнозировании интернет-трафика для таких целей как: прогноз объема трафика в пиковые часы на основе истории данных за предыдущие сутки или построение прогноза для различных протоколов (например, HTTP, FTP) для балансировки нагрузки в сети.

Можно выделить еще один метод машинного обучения, который на наш взгляд подходит для прогнозирования интернет трафика. Это метод опорных векторов (**SVM, Support Vector Machine**), который применяется для задач линейной и нелинейной классификации, регрессии и обнаружения аномальных данных.

В контексте интернет-трафика данный метод применяется для выявления паттернов в поведении трафика на основе признаков уровня сети и транспорта (например, протоколов, портов, флагов TCP). Это может помочь в построении моделей для распознавания типов трафика (например, HTTP, FTP, P2P) и обнаружения сетевых атак.

Для описания алгоритма работы данного метода возьмем картинку с набором каких-то элементов, которые принадлежат двум разным группам. Данные элементы в нашем контексте это

наборы о сетевом трафике, разделенные на две группы: трафик от веб-сайтов (HTTP) и трафик от потокового видео (RTSP). Пример будет показан на рисунке 1.

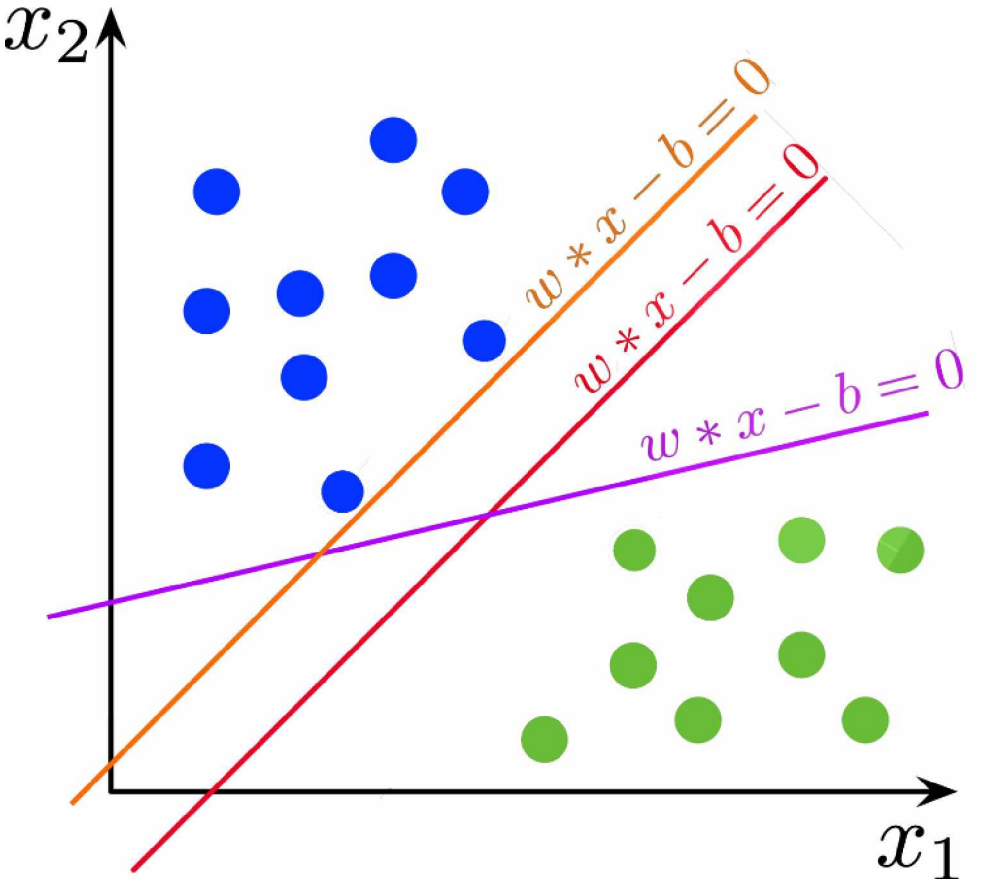


Рисунок 1 – Пример классификации точек с помощью SVM [3]

SVM помогает “нарисовать” линию (или более сложную поверхность в многомерном пространстве), которая максимально четко разделяет эти две группы трафика.

Допустим, у нас есть признаки: размер пакета (packet size), частота отправки пакетов (packet rate), протокол (например, TCP или UDP).

SVM будет анализировать эти признаки и пытаться найти такую границу, чтобы с одной стороны оказались все запросы HTTP, а с другой – все потоки видео. Если пакеты маленькие и передаются с низкой частотой по протоколу TCP → это HTTP-трафик. Если пакеты крупные и передаются с высокой частотой по протоколу UDP → это потоковое видео.

Если данные сложно разделить напрямую, SVM может использовать ядерные функции (например, радиально-базисную функцию, RBF), чтобы "развернуть" данные в пространство с большей размерностью. В этом новом пространстве данные станут линейно разделимыми, и алгоритм сможет провести более точную границу между группами трафика.

Таким образом, в результате работы SVM мы получаем модель, которая может эффективно классифицировать новые пакеты – определять, относятся ли они к HTTP или RTSP, даже если признаки пересекаются и имеют сложные зависимости [4].

Гиперплоскость – это граница, которая разделяет точки данных разных классов в пространстве признаков. В случае линейной классификации эта граница описывается уравнением вида:

$$w \cdot x + b = 0, \tag{2}$$

где w – вектор весов (нормаль к гиперплоскости), x – вектор признаков, b – смещение гиперплоскости относительно начала координат.

Опорные векторы – это объекты из обучающей выборки, которые находятся ближе всего к гиперплоскости и определяют положение границы разделения. Именно они влияют на выбор оптимального положения гиперплоскости.

Зазор – это расстояние между гиперплоскостью и ближайшими опорными векторами. Основная задача алгоритма SVM – найти такое положение гиперплоскости, при котором зазор будет максимальным. Более широкий зазор означает лучшую способность модели к обобщению.

Ядро – это функция, которая преобразует исходное пространство признаков в пространство более высокой размерности. Это позволяет сделать данные линейно разделимыми, даже если они в исходном пространстве имеют сложную структуру.

Еще одним подходом к анализу является **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) – это метод кластеризации, основанный на плотности, который используется для группировки данных и выявления аномалий. В отличие от методов, основанных на центрировании (например, K-means), DBSCAN не требует заранее задавать количество кластеров и хорошо обрабатывает данные с произвольной формой кластеров и наличием шумов [5].

Основные параметры метода: ε (эпсилон) – радиус окрестности для поиска соседей, minPts – минимальное количество точек в пределах радиуса ε, чтобы точка считалась "основной" (core point).

Алгоритм работы DBSCAN: если в окрестности радиуса ε содержится не менее minPts точек, то точка становится основной и образует новый кластер, далее все точки в окрестности основной точки добавляются в кластер. Если среди этих точек есть основные, их окрестности также добавляются в кластер (происходит «разрастание» кластера), если в окрестности точки меньше minPts точек, но она входит в окрестность основной точки, то она становится пограничной и добавляется в кластер. В конечном итоге если точка не попадает ни в один кластер, она считается шумом и не включается в кластер.

Методы машинного обучения играют ключевую роль в задачах прогнозирования интернет-трафика, так как позволяют выявлять сложные зависимости в данных, учитывать нелинейные паттерны и адаптироваться к изменениям в сетевом поведении. В отличие от традиционных статистических методов, машинное обучение учитывает временные зависимости, всплески трафика и сетевые аномалии, что повышает точность прогнозов.

В задачах прогнозирования интернет-трафика все три рассмотренных метода обладают высокой эффективностью в зависимости от условий задачи: Random Forest подходит для задач прогнозирования объёма трафика и определения типов трафика, благодаря способности обрабатывать большие объёмы данных и сложные зависимости, SVM эффективен для задач классификации и обнаружения аномалий, особенно в случае сложных паттернов поведения трафика, за счёт использования ядерных функций, DBSCAN применяется для кластеризации и обнаружения аномалий в интернет-трафике. Он хорошо выявляет группы данных с плотной структурой и позволяет эффективно обнаруживать выбросы (например, DDoS-атаки или всплески трафика).

Таким образом, для комплексных задач прогнозирования интернет-трафика рекомендуется использовать комбинацию методов: Random Forest — для прогнозирования объёма трафика и определения типов трафика; SVM — для классификации сложных сетевых паттернов и выявления нетривиальных зависимостей; DBSCAN — для кластеризации трафика и обнаружения сетевых аномалий, таких как всплески трафика и несанкционированная активность.

Таблица 1 – Сравнение методов Random Forest и SVM.

| Критерий | Random Forest | SVM |
|---|---|---|
| Тип задачи | Классификация, регрессия, отбор признаков, кластеризация | Классификация, регрессия |
| Работа с нелинейными данными | Хорошо обрабатывает сложные зависимости за счёт ансамбля деревьев | Требует использования ядра для работы с нелинейными данными |
| Обработка выбросов и шумов | Устойчив к выбросам за счёт ансамблирования | Может быть чувствителен к выбросам |
| Обучаемость | Быстро обучается на больших объёмах данных | Требует больше вычислительных ресурсов для обучения |
| Скорость предсказания | Высокая за счёт параллельной работы деревьев | Зависит от сложности ядра |
| Объём данных | Хорошо обрабатывает большие объёмы данных | Ограничения на объём данных из-за сложности вычислений |

***Список использованных источников:***

*1. Классификация IP-трафика методами машинного обучения / О.И. Шелухин, С.Д. Ерохин, А.В. Ванюшина // Научное издание, 2018. – 6-25 с.*

*2. Машинное обучение для начинающих: алгоритм случайного леса (Random Forest) // Proglib [Электронный ресурс]. – Режим доступа: https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12 – Дата доступа: 27.03.2025.*

*3. SVM Hyperplane // NEERC [Электронный ресурс]. – Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=%D0%A4%D0%B0%D0%B9%D0%BB:Svm_hyperplane.png&mobileaction – Дата доступа: 27.03.2025.*

*4. Математические основы теории машинного обучения и прогнозирования / В.В. Вьюгин // Московский центр непрерывного математического образования, 2014. –88-158 с.*

*5. Алгоритм кластеризации DBSCAN // ProProProgs [Электронный ресурс]. – Режим доступа: https://proproprogs.ru/ml/ml-algoritm-klasterizacii-dbscan. – Дата доступа: 27.03.2025.*

# INTERNET TRAFFIC FORECASTING TECHNIQUES

*N.S. Mlyonik, E.I. Semenovich*

*Belarusian State University of Informatics and Radioelectronics[1], Minsk, Republic of Belarus*

*Fedosenko V.A. - Candidate of Technical Sciences*

**Annotation.** The paper considers the task of Internet traffic forecasting in conditions of data volume growth and complexity of network patterns. The methods of machine learning are described: Random Forest for traffic volume forecasting and type classification, SVM for detection of complex patterns and anomalies, DBSCAN for clustering and detection of bursts and attacks. The effectiveness of the combined approach for improving prediction accuracy and robustness to network anomalies is shown.

**Keywords.** Internet traffic, prediction, machine learning, Random Forest, SVM, DBSCAN, network anomalies.