Введение

Классификация сетевого трафика — процесс сопоставления этого трафика и приложений, которые его создают. Это также можно назвать идентификацией протокола приложения. Классификация являет собой основу для широкого набора возможностей работы с сетью: от управления сетью до сетевой безопасности, от дифференциации сервисов до управления трафиком, от анализа современных тенденций до проведения сетевых исследований.

В данном контексте объектами классификации являются потоки сетевого трафика, которые представляют собой последовательности из сетевых пакетов, которыми обмениваются пары конечных узлов с целью коммуникации посредством компьютерных сетей. Классификация может быть основана на различной информации о потоках трафика, такой как номера портов, полезная нагрузка приложений или же статистические особенности потоков.

Классификация сетевого трафика позволяет обеспечить ясное понимание типа трафика, проходящего через сеть. Она является наиболее существенной частью современных сетевых систем. Для удобства управления администраторы сетевых систем всегда стараются получить точное и ясное соответствие сетевых приложений и создаваемого им трафика, тем самым обеспечив полноценный контроль над теми приложениями, которые используют их сеть. Основываясь на этой информации они могут принимать определенные меры управления и безопасности путем внедрения набора тщательно выверенных правил относительно доступа к конкретным типам приложений, сервисов или контента.

Классификация трафика является основой систем сетевой безопасности, включая файерволлы или сетевые системы обнаружения вторжения. Например, классификация позволяет осуществлять динамический контроль доступа в файерволлах и обеспечить контроль безопасности и аудит на уровне приложений в NIDS, что помогает обнаруживать и предотвращать аномальные и опасные действия со стороны злоумышленников, вредоносных программ и при проведении распределенных атак отказа доступа (DDoS).

В последнее время правительства стран требуют у интернет-провайдеров обеспечить возможность законного перехвата сетевого

трафика по аналогии с телефонными компаниями. Решения для перехвата трафика основываются на технологиях классификации сетевого трафика, с помощью которых они позволяют получить информацию об использовании сети конкретным человеком в любой момент времени с минимальными трудностями.

Непрерывное получение четкой информации о том, как различные типы приложений используют Интернет, очень полезны для большого круга задач, таких как анализ тенденций, планирование сетевой инфраструктуры, разработка сетевых устройств и т. д.

Для решения подобных задач в настоящее время широкое распространение получили методы, основанные на технологиях математической статистики и машинного обучения, с помощью которых даже неизвестные вредоносные приложения могут быть детектированы с определенной вероятностью.

Такие методы позволят разрабатываемой системе легко адаптироваться к постоянно изменяющейся природе Интернет ресурсов и учитывать специфику анализа сетевого трафика.

Исследования сетевого трафика показали, что он представляет собой сложный динамический процесс и является суперпозицией многих потоков с множественными взаимосвязанными характеристиками, которые генерируются различными протоколами. В последние годы сильно расширилось исследование полезности характеристик трафика для его классификации. Процесс состоит из двух отдельных этапов: первый — обучить классификатор, используя набор обучающих данных, второй — предсказать класс новых данных, используя классификатор.

В зависимости от того, помечены ли тренировочные данные или нет, мы можем применить либо алгоритм обучения с учителем, либо без учителя (кластеризации) на первом этапе.

Целью классификации сетевого трафика является отображение потока сетевых данных в определенные типы приложений или классы трафиков. Задача классификации заключается в разбиении объектов на классы. Объекты в пределах одного класса считаются эквивалентными с точки зрения критерия разбиения.

Для решения известных проблем классификации были предложены технологии машинного обучения (МО) (ML — Machine Learning) и интеллектуального анализа данных (Data Mining), оказавшиеся наиболее эффективными.

В первой главе анализируется современное состояние и задачи классификации IP-трафика. Внимание акцентируется на использо-

вании методов машинного обучения, включая методы классификации с учителем и без учителя (кластеризации).

Во второй главе рассмотрены классические парадигмы машинного обучения и интеллектуального анализа данных, используемые в задачах классификации IP-трафика. Анализируются современные методы и алгоритмы машинного обучения, используемые в дальнейших разделах: Искусственные нейронные сети (ИНС), метод опорных векторов, решающие деревья алгоритмы ID3C4.5, CART (Classification and Regression Tree), CHAID (Chisquare Automatic Interaction Detection), QUEST Quick, Unbiased, Efficient, Statistical Tree, «случайный лес» (Random forest, Bootstrap, Bagging и AdaBoost и др.). Рассмотрены основные методы кластеризации: иерархические и неиерархические. В заключении главы анализируются основные метрики оценки качества алгоритмов классификации, а также программные инструменты, применяемые в интеллектуальном анализе данных IP-трафика.

Третья глава посвящена проблемам контроля и анализа сетевого трафика. Рассмотрены современные средства и технологии, используемые в сетевых анализаторах трафика.

Четвертая глава занимает центральное место и посвящена изложению результатов исследований в области контролируемой и неконтролируемой классификации трафика методами машинного обучения. Здесь рассматривается широкий круг задач, начиная выбора атрибутов и кончая влияния фонового трафика на качество классификаци.

Работа по написанию книги распределилась среди авторов следующим образом: О.И. Шелухин: Главы 2, 3, разделы 4.11, 4,12, 4.13, 4.14, 4.15; С.Д. Ерохин: Глава 1, разделы 4.1, 4.5, 4.8, 4.9, 4.10, 4.12; А.В. Ванюшина: разделы 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.10.