2. ГОСТ Р ИСО/МЭК ТО 9294-93. Информационная технология. Руководство по управлению документированием программного обеспечения. Введ: 1994-07-01. – М.: Изд-во стандартов, 2003 // СПС КонсультантПлюс.

3. ГОСТ Р ИСО/МЭК 15408-1-2012. Национальный стандарт Российской Федерации. Информационная технология. Методы и средства обеспечения безопасности. Критерии оценки безопасности информационных технологий. Часть 1. Введение и общая модель – Введ: 2012-11-15. – М.: Стандартиформ, 2011 // СПС КонсультантПлюс.

4. ГОСТ Р ИСО/МЭК 12207-2010. Национальный стандарт Российской Федерации. Информационная технология. Системная и программная инженерия. Процессы жизненного цикла программных средств – Введ: 2012-03-01 – М.: Стандартиформ, 2011 // СПС КонсультантПлюс.

5. ГОСТ Р ИСО 9127-94. Системы обработки информации. Документация пользователя и информация на упаковке для потребительских программных пакетов – Введ: 1995-07-01. – М.: Изд-во стандартов, 2003 // СПС КонсультантПлюс.

УДК 681.32

КЛАССИФИКАЦИЯ IP-ТРАФИКА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Лях Андрей Александрович, *магистрант Института управления и экономики АПК ФГБОУ ВО РГАУ - МСХА имени К.А. Тимирязева, andrei-lyah@rambler.ru*

Аннотация: *В статье рассматривается задача классификации сетевого трафика с использованием методов машинного обучения. Рассматривается классификация приложений в условиях априорной неопределенности.*

Ключевые слова: *анализ сетевого трафика; классификация сетевого трафика; машинное обучение.*

Классификация трафика требуется в наше время, так как полученные результаты могут применяться в различных приложениях, важных как для администрирования сети, так и для конечного пользователя.

С точки зрения поставщика, определение протоколов / приложений / типов приложений по потокам данных в сети может быть использовано для:

- управление сетью и движением в ней (например, для блокировки отдельных протоколов, таких как битторрент),
- обеспечение высокого качества обслуживания клиентов путем эффективного распределения наиболее приоритетных потоков и регулирования скорости передачи отдельных пакетов,
- регулирование цен на услуги,
- планирование размещения и использования ресурсов,
- оптимизация предоставляемых услуг и алгоритмов маршрутизации (например, для изменения приоритетов передачи различных типов данных в случае высокой нагрузки на сеть).

Оценка текущего использования сети пользователями может дать представление об оптимальном устройстве новых сетей с учетом понимания предпочтений и принципов пользователей интернета и интернет-услуг, поскольку можно получить подробную статистику по всем услугам [1, 2].

Классификация сетевого трафика – процесс сопоставления этого трафика и приложений, которые его создают. Это также можно назвать идентификацией протокола приложения. Классификация представляет 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.

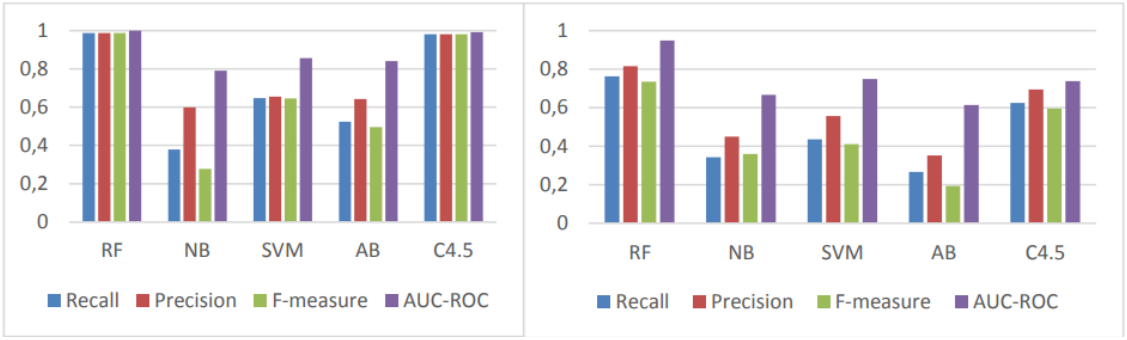
Так как потребности пользователей относительно использования сети постоянно меняются, необходимо их знать и модифицировать сеть в соответствии с актуальными запросами. Для этого нужно как уметь моделировать устройство сети на текущий момент времени, так и понимать направление движения её развития и изменения. Например, на сегодняшний день видна тенденция отказа от преобладающего ранее принципа асимметрии устройства сети в том смысле, что клиенты загружают намного больше информации, чем отправляют её в сеть.

Появление P2P-приложений, VoIP, видео звонков, потоковой передачи мультимедиа и прочих новшеств должно вызвать у интернет провайдеров соответствующие ответные действия по переустройству сети под новые запросы клиентов. Кроме того, в настоящее время увеличивается количество так называемых «умных устройств», которые должны в будущем составить Интернет вещей: он также поставит перед интернет провайдерами ряд задач для обеспечения максимальной эффективности своей работы.

Были исследованы особенности классификации приложений использующих стек TCP/IP в условиях априорной неопределенности. Последняя характеризовалась наличием фонового трафика, принадлежащего к классам, которые не участвовали в обучении алгоритма, что значительно ухудшает точность классификации. Влияние неизвестного фонового трафика на качество классификации с использованием МО рассматривалось на примере приложений: HTTP; BitTorrent; Skype; Steam; DNS. Фоновый трафик представлял собой данные, которые не были представлены в обучающей выборке, однако присутствовали в тестовом наборе: SSL; LLMNR; Quic.

Классификация при наличии фонового трафика показала, что алгоритмы МО с учителем, качество работы которых основывается на полноте и достоверности обучающих выборок данных, не способны определять новые (неизвестные) данные, что приводит к критичным ошибкам. Наличие фонового трафика, принадлежащего к классам, которые не участвовали в обучении алгоритма, значительно ухудшает точность классификации.

В качестве примера на рисунке 1 представлены гистограммы усредненных оценок качества классификации для всех рассматриваемых приложений для различных алгоритмов классификации при отсутствии (рисунок 1, а) и наличии (рисунок 1, б) фонового трафика.



а) **б)**

Рис. 1. Усредненная гистограмма по алгоритмам классификации при: а) отсутствии в наборе данных фона, б) наличии в наборе данных фона

Так для приложения Skype снижение величины AUC-ROC для алгоритма *RF* составляет 13,2%, а для C4.5 достигает 35%. Для BitTorrent эти величины составляют: для RF- 5,7%, а для C4.5- 37%. В среднем по всем приложениям снижение составляет для RF 5,2 % а для C4.5 достигает 25%, что обусловлено ложной классификацией фоновых приложений.

Для алгоритма RF величина вероятности ложной классификации FPR при наличии фона возрастает в среднем для всех приложений с величины 0.003 (при отсутствии фона) до величины 0,059, т.е. около 20 раз. Для алгоритма C4.5 увеличение вероятности ложной классификации составляет 18,5 раз. Естественным развитием в данном направлении является применение методов кластеризации, которые предназначены для первичного определения и разграничения неизвестных типов приложений, а уже затем анализа и классификации.

Однако проведенные исследования алгоритмов кластеризации k-Means и DBSCAN при обучении без учителя, для того же состава трафика протоколов, что и при классификации с учителем показали в целом неудовлетворительные результаты. Рассмотренные алгоритмы неконтролируемого обучения k-Means и DBSCAN значительно уступают в качестве алгоритму обучения с учителем Random Forest. Алгоритм k-Means справляется с кластеризацией потоков сетевого трафика, однако это происходит лишь при условии, что количество кластеров априорно известно. В противном случае качество классификации значительно ухудшается и для приложений HTTP, Skype, DNS, Steam достигает 30%. Алгоритм DBSCAN имеет значительные ошибки в количестве и содержании кластеров ряда анализируемых приложений (классы HTTP, SSL, Steam и Skype оказались разбросаны по многим кластерам).

Повысить эффективность классификации в условиях возможного появления неизвестного фона можно введением дополнительного класса под названием «Неизвестное приложение». Так при введении дополнительного класса величина TPR снизилась для алгоритма RF с 0,987 до 0,964, т.е. не более 2,5%, в то время как величина FPR снизилась на 33%.

На рисунках 2 и 3 представлены зависимости средних значений TPR и FPR

характеризующие эффективность классификации анализируемых приложений включая введение нового класса.

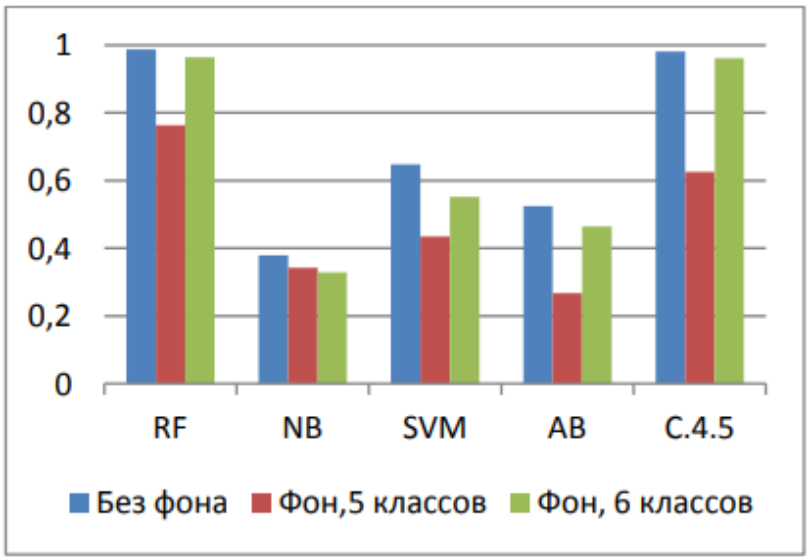


Рис. 2. Сравнительная гистограмма параметра True Positive Rate для алгоритмов классификации

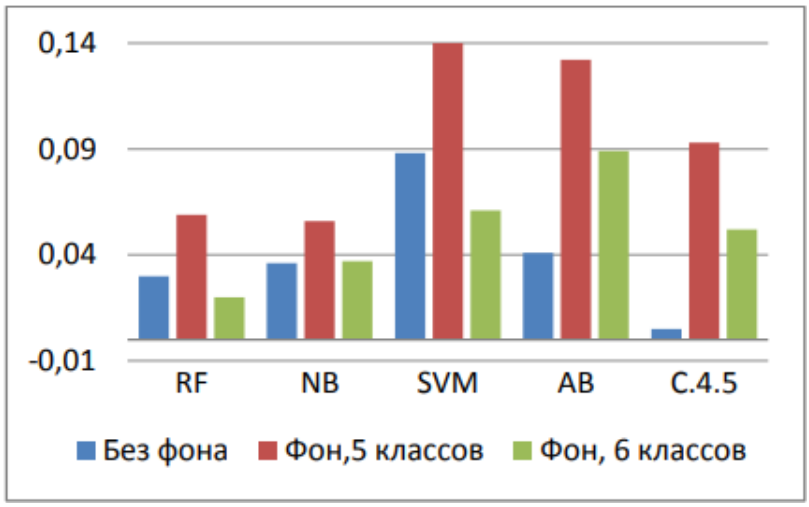


Рис. 3. Сравнительная гистограмма параметра False Positive Rate для алгоритмов классификации

Для преодоления проблемы априорной неопределенности вызванной присутствием фонового трафика в КС, построенных с использованием телекоммуникационных технологий, введен в рассмотрение дополнительный класс «Неизвестное приложение». показавший результаты незначительно уступающие по качественным показателям алгоритмам классификации «с учителем» и являющийся альтернативой методам кластеризации k-Means и DBSCAN. Показано, что для наиболее эффективного алгоритма классификации

RF введение дополнительного класса позволило снизить величину ложной классификации (FPR) на 33%, при этом снижение достоверности правильной классификации (TPR) составляет не более 2,5%.

Библиографический список

1. Muhammad, S. Machine learning based intelligent system for IP traffic classification / S. Muhammad, A. Kashif, K. Jebran, I. Faisal // Sindh University Research Journal. – 2013. – № 45. – P. 731-736.

2. Muhammad, S. Network traffic classification techniques and comparative analysis using machine learning algorithms / S. Muhammad, Y. Xiangzhan, A. L. Asif, Y. Lu, N.K. Karn, F. Abdessamia // 2nd IEEE International Conference on Computer and Communications. – 2016. – P. 2451-2455.

УДК 311.42

ОЦЕНКА ВЛИЯНИЯ ОТНОСИТЕЛЬНЫХ ПОКАЗАТЕЛЕЙ ПРОМЫШЛЕННОГО ПРОИЗВОДСТВА РЕГИОНОВ РОССИЙСКОЙ ФЕДЕРАЦИИ НА СОВРЕМЕННОЕ СОСТОЯНИЕ ПРОМЫШЛЕННОСТИ

Перегудова Вероника Сергеевна, *магистрант Института экономики и управления АПК ФГБОУ ВО РГАУ - МСХА имени К.А. Тимирязева, veronika.peregudova-0598@mail.ru*

Аннотация: *В статье анализируется степень влияния тех или иных факторов на индекс промышленного производства.*

Ключевые слова: *промышленность, регрессионный анализ, факторы, зависимые переменные, индекс промышленного производства.*

Промышленный сектор имеет большое значение для развития страны в целом. Доказанный факт, что страны с сильным промышленным сектором показали больший экономический рост, улучшили национальный доход и повысили уровень жизни людей. Индустриализация сыграла важную роль в улучшении экономических условий различных стран.

Преимущества индустриализации заключаются в следующем:

1) Экономическая стабильность.

Страна, которая зависит только от сельского хозяйства, не может достичь стабильности. Существует дисбаланс, используется только человеко-сила, то есть трудоемкая технология. Следовательно, индустриализация обеспечивает экономическую стабильность страны там, где все зависит не только от одного сектора. Существует баланс между вкладом обоих секторов в экономику.

2) Увеличение денежных резервов.

С появлением все большего числа отраслей промышленности произойдет рост денежных поступлений. Экспорт будет расти, а импорт начнет сокращаться. Будет больше притока денежных средств, увеличится самообеспеченность.

3) Использование природных ресурсов.

Существует много неиспользуемых ресурсов, таких как бесплодные земли и полезные ископаемые, которые могут быть бесполезны для сельскохозяйственного или финансового секторов страны. Поэтому промышленное развитие увеличило бы использование таких ресурсов, которые в противном случае были бы полностью растрачены впустую, и их вклад в денежном выражении был бы равен нулю.