DOI 10.31891/2307-5732-2021-301-5-22-26
УДК 004.891

ДЖУЛІЙ В.М.
Хмельницький національний університет
ORCID http://orcid.org/0000-0003-1878-4301
e-mail:dg2303@ukr.net
КЛЬОЦ Ю.П.
Хмельницький національний університет
ORCID https://orcid.org/0000-0002-3914-0989
e-mail:sprklyots@gmail.com
МУЛЯР І.В.
Хмельницький національний університет
ORCID http://orcid.org/0000-0002-6659-605X
e-mail:iga2000@yahoo.com
ЖИЛЕВИЧ М.Л.
Хмельницький національний університет
e-mail:dg2303@ukr.net
ДЖУЛІЙ А.В.
Університет економіки і підприємництва, м.Хмельницький
ORCID ID: 0000-0001-5011-3052
e-mail: kksmkhnu@gmail.com

КОНТРОЛЬ ДОДАТКІВ ІНТЕРНЕТ-ТРАФІКА КОМП'ЮТЕРНИХ МЕРЕЖ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Розглянуто актуальне завдання контролю доступу до Інтернет-ресурсів має важливе прикладне значення: блокування доступу до нелегальної, екстремістської, антисоціальної інформації, запобігання розголошенню конфіденційної інформації через Інтернет та ін. Для вирішення подібних завдань широкого поширення набули методи машинного навчання. Одним з найбільш часто використовуваних і ефективних для класифікації мережевого трафіка методів машинного навчання є «випадковий ліс» (Random Forest), що представляє собою ансамблевий метод, який діє шляхом побудови множини вирішальних дерев. Для оцінки ефективності роботи алгоритму Random Forest при класифікації мережевого трафіка за типами прикладних протоколів, що працюють в мережі Інтернет, був здійснений збір трафіка в мережі. Досліджувалися додатки, які генерують пакети, що відносяться до різних протоколів прикладного рівня: BitTorrent, DNS, HTTP, SSL, Skype, Steam. Після відбору інформаційних ознак і попередньої обробки даних сформовані навчальна і тестова вибірки, одна з яких містила фоновий трафік. В результаті застосування алгоритму класифікації Random Forest до отриманих даних знайдені оцінки ефективності роботи даного алгоритму в умовах наявності і відсутності фонового мережевого трафіку.

Ключові слова: моделі, ефективність, алгоритми, *мережевий трафік, вирішальні дерева, машинне навчання, фоновий трафік.*

DZHULIY VOLODYMYR M., KLYOTS YURII P., MULYAR IHOR V., ZHILEVICH MYKHAILO L.
Khmelnytsky National University
DZHULIY ANDRII V.
University of Economics and Entrepreneurship, Khmelnytsky, Ukraine

CONTROL OF APPLICATIONS OF INTERNET TRAFFIC OF COMPUTER NETWORKS METHODS OF MACHINE LEARNING

The considered actual problem of controlling access to Internet resources has an important applied value: blocking access to illegal, extremist, antisocial information, preventing the disclosure of confidential information via the Internet, etc. For the development of a wide range of enterprises, the methods of machine technology have been developed. One of the most frequently victorious and effective methods for the classification of hedgehog traffic is the Random Forest, an ensemble method, which is a kind of tree path to inspire a multitude of virgins. To assess the effectiveness of the Random Forest algorithm in the classification of network traffic by types of application protocols operating on the Internet, the collection of network traffic was performed. Applications that generate packets related to different application layer protocols were studied: BitTorrent, DNS, HTTP, SSL, Skype, Steam. After selection of information features and preliminary data processing, training and test samples were formed, one of which contained background traffic. As a result of applying the Random Forest classification algorithm to the obtained data, estimates of the efficiency of this algorithm in the presence and absence of background network traffic were found. The presence of background traffic belonging to classes that did not participate in the training of the algorithm significantly impairs the accuracy of classification. It is shown that the number of attributes for traffic classification is not as important as the choice of classification algorithms. The results of the classification at the testing stage showed that the machine learning algorithms using the "decision trees" Random Forest and C4.5 are best suited for classification with a large number of classes. Classification accuracy indicators using AdaBoost and Bagging suggest that in most cases, combining multiple classifiers into an ensemble and making a decision based on "voting" can improve the results of the classification. To achieve classification accuracy, it is sufficient to calculate the classification attributes for a sample containing 5 .... 10 consecutive packets. Accuracy can be improved to 99% and higher if the statistics were calculated on the basis of 35 consecutive packets.

Keywords: models, efficiency, algorithms, network traffic, decision trees, machine learning, background traffic.

Вступ

Задача контролю доступу до Інтернет-ресурсів має важливе прикладне значення: блокування

доступу до нелегальної (екстремістської, антисоціальної та ін.) інформації, запобігання доступу до Інтернет-ресурсів в особистих цілях в навчальний або робочий час, запобігання витоку конфіденційної інформації через Інтернет, що не відповідає політиці або небажану поведінку користувачів, шкідливі програми і атаки, які зазвичай використовують неперевірений канал зашифрованого трафіку HTTPS [1].

Це можуть бути потенційно небезпечні програми. Вони можуть ставити під загрозу дані і системні активи, впливати на продуктивність праці співробітників і використовувати пропускну здатність мережі. На сьогоднішній день існує безліч як комерційних, так і некомерційних продуктів, які розв'язують подібні задачі. До найбільш поширених комерційних продуктів можна віднести: WebSense, NetNanny.

Кількісні показники при оцінюванні роботи систем фільтрації Інтернет-трафіка наступні: точність аналізу - відсоток правильно відфільтрованих Інтернет-ресурсів; зайве блокування або хибно позитивні помилки - відсоток «хороших» ресурсів, помилково заборонені системою фільтрації.

Для вирішення подібних завдань широкого поширення набули методи, засновані на технологіях математичної статистики і машинного навчання, за допомогою яких невідомі шкідливі програми можуть бути виявлені з визначеним ступенем ймовірності. Даний підхід дозволяє системі, що проектується, легко адаптуватися до постійно змінюваної природи Інтернет-ресурсів та враховувати специфіку аналізу мережевого трафіка [2].

Одним з найбільш часто використовуваних і ефективних для класифікації мережевого трафіка методів машинного навчання є вирішальне дерево. Вирішальне дерево (Decision tree) – розв'язок задачі навчання з учителем, засноване на тому, як вирішує завдання прогнозування людина. У загальному випадку дерево з вирішальними правилами в нелистових вершинах (вузлах) і деякому висновку цільової функції в листових вершинах (прогнозом) [3]. Вирішальне правило – деяка функція від об'єкта, що дозволяє визначити, в яку з дочірніх вершин потрібно помістити даний об'єкт. У листових вершинах можуть перебувати різні об'єкти: клас, який потрібно присвоїти об'єкту який потрапив туди (в задачах класифікації), безпосереднє значення цільової функції (задача регресії) [4].

Випадковий ліс (Random Forest) являє собою ансамблевий метод навчання для класифікації і регресії, який діє шляхом побудови множини вирішальних дерев. Випадковий ліс - є одним з небагатьох універсальних алгоритмів. Універсальність полягає, по-перше, в тому, що він хороший у багатьох задачах (70% в задачах, що зустрічаються на практиці, якщо не враховувати задачі з зображеннями), по-друге, в тому, що є випадкові ліса для вирішення задач класифікації, регресії, кластеризації, пошуку аномалій, селекції ознак і т.д. RF (random forest) – це множина вирішальних дерев. У задачі регресії їх відповіді усереднюються, в задачах класифікації приймається рішення голосуванням за більшістю. Всі дерева будуються незалежно за наступною схемою: вибирається підвибірка навчальної вибірки розміру samplesize (може бути з поверненням), за нею будується дерево (для кожного дерева своя підвибірка); для побудови кожного розгалуження в дереві переглядаємо max_features випадкових ознак (для кожного нового розгалуження свої випадкові ознаки); вибираються найкращі ознака і розгалуження по ньому (за заздалегідь заданим критерієм), дерево будується, як правило, до вичерпання вибірки (поки в листі не залишаться представники тільки одного класу), але в сучасних реалізаціях є параметри, які обмежують висоту дерева, число об'єктів в листі і число об'єктів підвибірки, при якому проводиться розгалуження. Така схема побудови відповідає головному принципу ансамблювання (побудови алгоритму машинного навчання на базі кількох, в даному випадку, вирішальних дерев): базові алгоритми повинні бути хорошими і різноманітними (тому кожне дерево будується на своїй навчальній вибірці і при виборі розщеплення є елемент випадковості).

Постановка задачі

Метою роботи є оцінка ефективності роботи алгоритму Random Forest (RF) в задачах класифікації додатків в умовах наявності і відсутності фонового мережевого трафіка. Однією з головних проблем, яку потрібно вирішити при розробці системи класифікації трафіка використовуючи алгоритми машинного навчання є формування вхідних даних і програмне забезпечення, яке використовується для цих цілей. Вхідні дані представляють собою зразки пакетів, класифікованих згідно з додатками, які їх створили. В даний час не існує єдиного набору вхідних даних, що є стандартом в області класифікації трафіку, як не існує і єдиного підходу до їх отримання. У той же час точність алгоритмів машинного навчання безпосередньо залежить від обсягу, якості та репрезентативності набору вхідних даних, що використовувалися в процесі навчання. Тому отримання якісного набору вхідних даних є важливим завданням. Для коректної реалізації машинного навчання слід попередньо вирішити завдання вибору оптимальних атрибутів і скорочення їх кількості, а також дослідити можливості застосування методів кластеризації для попереднього виділення окремих груп протоколів. Алгоритми машинного навчання з використанням «дерев рішень» найкращим чином підходять для класифікації при великій кількості класів [5].

Основна частина

Алгоритм RF опирається на техніку беггінга – використання композиції незалежно навчаючих алгоритмів. В результаті будується множина вирішальних дерев, кожне з окремої випадкової підмножини вхідної вибірки даних, розмір підвибірок співпадає з розміром вхідної вибірки і має повторення [1]. Для i-го дерева генерується випадковий вектор V_n, який не залежить від згенерованих раніше векторів $V_1, \dots, V_{n-1}$, але має такий же розподіл. Дерево «нарощується» із застосуванням тренувальної вибірки і

вектора V_n, в результаті утворюється класифікатор $g\{x, V_n\}$, де $x-$ вхідний вектор. Дерева будуються за допомогою стандартного алгоритму бінарного вирішального дерева.

Розглянемо навчальну вибірку $S^t = \{(x_1, y_1), ..., (x_t, y_t)\}$, де x_i – вектори інформаційних ознак об'єкта. Множина всіх можливих значень векторів ознак X - простір образів (об'єкт – точка в просторі образів). Стан залежною змінної (мітки класів) y_i можуть приймати тільки кінцеве число значень. З кожною вершиною s дерева пов'язана деяка підмножина простору образів $X_s \subset X$ (з кореневою вершиною зв'язується весь простір образів X). Підвибірка $S_s^t \subset S^t$ навчальної вибірки (з кореневою вершиною зв'язується вся навчальна вибірка S^t). Вирішальне правило $p_s : X \to \{0,1,...,n_s - 1\}$, де $n_s \geq 2$ – кількість дочірніх вершини s (для бінарного дерева $n_s = 2$), яке визначає розбиття множини X на n непересічних підмножин. Найчастіше в якості вирішального правила береться одна з ознак $x_{i(s)}$. Позначимо $s_{i(s)}$ вершину, що є i-м нащадком вершини s. Множина X_s і функція p_s задають множини: $X_{s_{i(s)}} = X_s \cap \{x \in X : p_s(x) = i\}$. В результаті кожному внутрішньому вузлу відповідає один з вхідних атрибутів, термінальним вершинам відповідають мітки класів. Метою побудови вирішального дерева є класифікація векторів x. Атрибути для кожного дерева вибираються з навчальної вибірки випадковим чином. Атрибути, що відносяться до більш, ніж двом класам, можуть бути відібрані більше одного разу для різних вузлів.

Розгалуження в проміжному вузлі при побудові дерева необхідно проводити так, щоб забрудненість була мінімальною.

Нехай $R(s)$ – деяка підвибірка, пов'язана з вершиною s. Забрудненість вершини $i(s) = 0$, якщо підвибірка $R(s)$ містить екземпляри тільки одного класу, і буде максимальною при однаковій кількості екземплярів кожного класу. В результаті, кількість екземплярів, що належать іншим класам, (домішок), в кожному класі після розбиття має прагнути до мінімуму. Існує кілька мір забрудненості вершини. Найпопулярнішими є ентропійний критерій і критерій Джині.

Ентропійний критерій заснований на понятті кількості інформації, яку містить розгалуження. Ентропію i-го вузла можна обчислити скориставшись співвідношенням $i(s) = -\sum_{i=1}^{c} E(w_i) \log_2 E(w_i)$, де $E(w_i)$ – частка екземплярів класу w_i в $R(s)$, c – кількість класів.

Критерій Джині оцінює "відстань" між розподілами класів. Ще одна, менш використовувана, міра забрудненості заснована на частоті помилок класифікації. Вона визначається як мінімальна ймовірність того, що екземпляр буде класифікований невірно: $i(s) = 1 - \max_j E(w_j)$.

Чим менше критерій розщеплення, тим краще якість розщеплення. На практиці важливо, щоб забрудненість зменшувалася при переході від вузла до його нащадків. Визначимо зменшення забрудненості на вузлі s як $\Delta i(s) = i(s) - E_L(s_L) - E_R(s_R),$, де E_L і E_R – частки екземплярів лівого (s_L) і правого (s_R) нащадків вузла s відповідно. Формула справедлива тільки для випадку бінарного дерева. Кращим розщепленням вважається те, при якому величина $\Delta i(s)$ максимальна. Як правило, при побудові дерева поняття максимальної $\Delta i(s)$ не є точним, оскільки при виборі оптимального розщеплення не провадиться повний перебір всіх можливих варіантів, а лише проводиться звуження набору до декількох варіантів, з яких потім і вибирається той, за якого отримаємо найбільше значення. Якщо вирощувати повне дерево, поки в листових вершинах не буде досягнута мінімально можлива забрудненість, відбудеться перенавчання моделі, тобто вона просто «запам'ятає» всі варіанти класифікації для тренувальної вибірки і не буде здатна до роботи на тестових даних. В результаті кожна листова вершина буде відповідати за один класифікуємий екземпляр. Якщо ж зупинити розщеплення занадто рано, виявиться досить високою помилка і буде знижена ефективність. Тому важливим завданням є побудова збалансованого дерева, для чого необхідний вибір правильного критерію зупинки розщеплення. Таких критеріїв існує кілька. Один з них – завдання мінімального порогу на число екземплярів у листових вершинах. Як тільки кількість екземплярів в даній вершині стає менше заданого порога, розщеплення зупиняється, і вершина вважається термінальною. Ще один критерій - вибір гранично допустимого значення параметра $\Delta i(s)$. Розщеплення вершини не відбувається, якщо після нього зменшення забрудненості менше заданого порогу. Можна також просто обмежувати глибину дерева. В цьому випадку побудова дерева закінчується, якщо досягнута задана глибина. Поширеним методом є крос-валідація, за якої відбувається оцінка помилки класифікації на тренувальній множині і тестовій. Розщеплення відбувається до тих пір, поки помилка мінімальна.

Використання ранньої зупинки розщеплення має істотний недолік: рішення про зупинку розщеплення приймається без урахування випадку, при якому воно могло б бути продовжено. Тобто, навіть

якщо продовженні розщеплення сильно підвищили б ефективність класифікації, така ситуація не розглядається. Існує інший підхід до зменшення дерев – прунінг (pruning, відсікання гілок) повних вирішальних дерев, коли проміжні вузли замінюються на термінальні і відносяться до більш пріоритетного в піддереві класу. У RF прунінг не використовується, так як має високу обчислювальну складність.

Після побудови всіх дерев ліс організовується як найпростіший ансамблевий класифікатор. Кожне дерево голосує за очікуваний клас і екземпляр визначається в клас, який набрав найбільшу кількість голосів по всіх деревах у лісі.

Алгоритм RF має ряд переваг: низька кількість керуючих параметрів і параметрів моделі; стійкість до перенавчання; не потрібний відбір ознак. Одним з важливих переваг RF є те, що дисперсія моделі зменшується зі збільшенням кількості дерев в лісі, в той час як зміщення залишається незмінним. До недоліків RF можна віднести інтерпретованість, втрати продуктивності через корельованих змінних, і залежність від генератора випадкових чисел.

Проведено побудову випадкового лісу і оцінка якості класифікації на заданій вибірці. Дослідним шляхом були підібрані найбільш прийнятні параметри алгоритму.

Для визначення ефективності алгоритму використовуються наступні метрики: точність, повнота, F1-міра, значення яких легко визначити на підставі матриці помилок класифікації, яка складається для кожного класу окремо. У матриці відображається кількість правильних і не правильних рішень по заданому класу.

Графічне представлення даних метрик отриманих експериментально для всіх проаналізованих класів наведено на рис. 1.

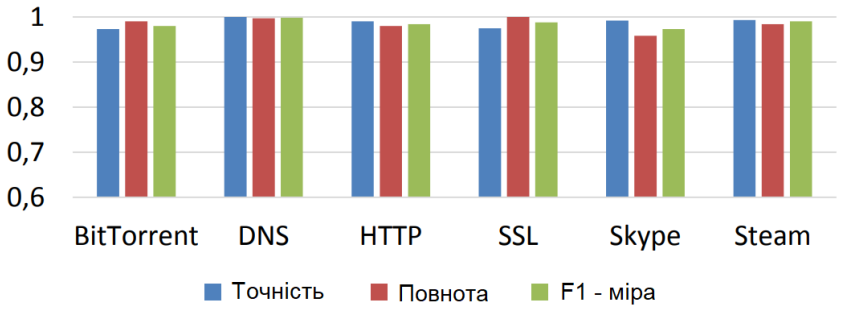

Рис. 1. Точність, повнота, F-міра для тестової вибірки

Видно, що найбільшу ефективність алгоритм має для даних, що відносяться до DNS трафіку. Крім перевірки роботи алгоритму на тестовій вибірці, що має такий же класовий склад, як і навчальна, оцінка його якості проводилася також в умовах присутності фонового трафіку, тобто в разі, коли в тестовій вибірці були присутні екземпляри класів, відсутніх в навчальній вибірці.

Така ситуація, коли в даних, що класифікуються присутній фоновий трафік, більш наближена до дійсності, в силу різноманіття використовуваних в мережі Інтернет протоколів. Такий набір даних дозволяє оцінити роботу алгоритму в реальних умовах. Розглянемо, як змінилися показники якості класифікації (рис. 2).

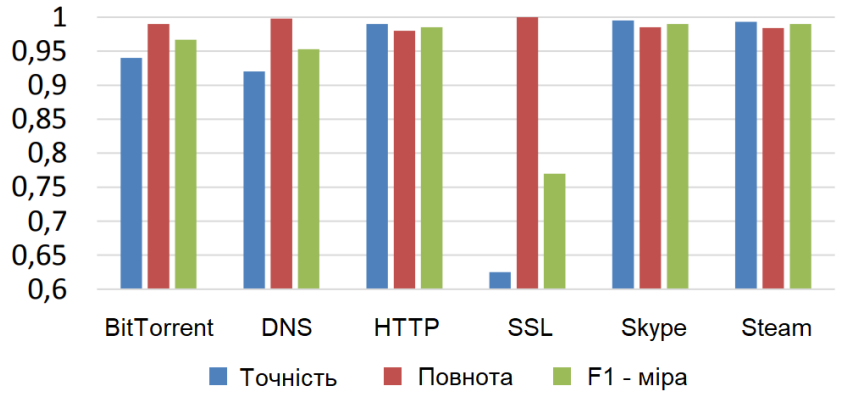

Рис. 2. Точність, повнота, Fl-міра при наявності фонового трафіку

Як видно, наявність фонового трафіку практично не вплинуло на значення повноти, але значно погіршило значення точності класифікації, оскільки збільшилася кількість False Positive екземплярів поява яких викликано наявністю фонового трафіку, що належить до класів, які в навчанні не брали участь. Разом з тим, алгоритм RF продемонстрував високу ефективність у режимі offline, про що, зокрема, свідчить Fl-міра рівна відповідно 0,987 і 0,759 за відсутності і наявності фонового трафіку.

Висновки

Для проведення аналізу ефективності алгоритму RF в задачах класифікації мережевого трафіку був зібраний набір даних, що містить потоки, що відносяться до різних протоколів прикладного рівня: BitTorrent, DNS, HTTP, SSL, Skype, Steam [5]. Проведена оцінка роботи алгоритму RF як на «чистій» тестової вибірці, так і на вибірці, що містить фоновий трафік у вигляді екземплярів, які відносяться до класів невідомих навченому алгоритму. Показано, що наявність домішок істотно впливає на точність класифікації виконаної за допомогою алгоритму RF. Алгоритм RF продемонстрував високу ефективність в режимі offline, про що, зокрема, свідчить F1-міра, рівна відповідно 0,987 і 0,759 за відсутності і наявності фонового трафіку.

Наявність фонового трафіку, що належить до класів, які в не брали участь в навчанні алгоритму, значно погіршує точність класифікації. Алгоритм RF мало придатний для класифікації в режимі реального часу через часову складність обробки, що оцінюється співвідношенням, $\left(Mmn\log\left[\left(n\right)\right]\right)$, де n – кількість екземплярів, m – кількість інформаційних признаків, M – кількість дерев.

Показано, що кількість атрибутів для класифікації трафіку не так важливо, як вибір алгоритмів класифікації. Тим не менш, важливо уникати атрибутів, які містять конкретні значення тільки для невеликої кількості випадків, що належать до конкретного класу, що може привести до зайвого навчання класифікатора і також не буде можливості ідентифікувати невідомі випадки.

Літертура

1. Шелухин О.И. Сетевые аномалии. Обнаружение, локализация, прогнозирование / О.И. Шелухин. – М. : Горячая линия -Телеком, 2019. – 448 с.

2. Шелухин О.И. Классификация IP-трафика методами машинного обучения / О.И. Шелухин, С.Д. Ерохин. – М. : Горячая линия -Телеком, 2018. – 284 с.

3. Батурин Ю.М. Компьютерная преступность и компьютерная безопасность / Ю.М. Батурин, А.М. Жодзинский. – М. : Юридическая литература, 2006. – 160 с.

4. Нестеров С.А. Основы информационной безопасности : учебник / С. А. Нестеров. – СПб : Лань, 2017. – 423 с.

5. Олифер В.Г. Безопасность компьютерных сетей / В. Г. Олифер, Н. А. Олифер. – М. : Горячая линия-Телеком, 2017. – 644 с.

References

1. Sheluhin O.I. Setevye anomalii. Obnaruzhenie, lokalizaciya, prognozirovanie / O.I. Sheluhin. – M. : Goryachaya liniya -Telekom, 2019. – 448 s.

2. Sheluhin O.I. Klassifikaciya IP-trafika metodami mashinnogo obucheniya / O.I. Sheluhin, S.D. Erohin. – M. : Goryachaya liniya -Telekom, 2018. – 284 s.

3. Baturin Yu.M. Kompyuternaya prestupnost i kompyuternaya bezopasnost / Yu.M. Baturin, A.M. Zhodzinskij. – M. : Yuridicheskaya literatura, 2006. – 160 s.

4. Nesterov S.A. Osnovy informacionnoj bezopasnosti : uchebnik / S. A. Nesterov. – SPb : Lan, 2017. – 423 s.

5. Olifer V.G. Bezopasnost kompyuternyh setej / V. G. Olifer, N. A. Olifer. – M. : Goryachaya liniya-Telekom, 2017. – 644 s.

Рецензія/Peer review : 17.09.2021 р. Надрукована/Printed :10.10.2021 р.