Інформаційні технології

УДК 004.732.056 **doi: 10.26906/SUNZ.2023.4.070**

С. Ю. Гавриленко, В. О. Полторацький

Національний технічний університет “Харківський політехнічний інститут”, Харків, Україна

МЕТОД ПІДВИЩЕННЯ ОПЕРАТИВНОСТІ КЛАСИФІКАЦІЇ ДАНИХ ЗА РАХУНОК ЗМЕНШЕННЯ КОРЕЛЯЦІЇ ОЗНАК

Анотація. *Об'єктом* дослідження є процес ідентифікації стану комп'ютерної мережі. *Предметом* дослідження є методи ідентифікації стану комп’ютерних мереж. *Метою* статті є підвищення оперативності виявлення вторгнень у комп’ютерні мережі за рахунок зменшення кореляції ознак. *Методи, що використовуються*: методи штучного інтелекту, машинного навчання, методи зменшення кореляції ознак. *Отримано такі результати*: Досліджено ефективність використання підходів, які зменшують кореляцію даних: метод головних компонентів (PCA), незалежних компонентів (ICA), L1 та L2 регуляризацію, обгрунтовано метод для подальшого дослідження. За результатами досліджень запропоновано спеціальну процедуру зменшення кореляції вихідних даних. Для оцінки якості та оперативності запропонованої процедури, розроблено програмні моделі на основі: Gradient Boosting, Random Forest, повнозв’язної нейронної мережі (FCNN) та згорткової нейронної мережі (CNN). У якості вихідних даних використано набір UNSW-NB 15, який містить інформацію про нормальне функціонування мережі та під час вторгнень. Виконано порівняльний аналіз якості та оперативності розроблених моделей. *Висновки.* Наукова новизна отриманих результатів полягає у розробці методу виявлення вторгнень в комп’ютерні мережі, який відрізняється від відомих наявністю спеціальної процедури зменшення кореляції вихідних даних, що дозволило підвищити оперативність процесу ідентифікації.

Ключові слова: машинне навчання, класифікація даних, попередня обробка даних, кореляція даних, комп’ютерні мережі, нейронні мережі, ансамблеві класифікатори, системи виявлення вторгнень.

Вступ

На сьогодні, виявлення вторгнень в комп’ютерні системи та мережі є одним із пріоритетних завдань та потребує удосконалення та розробки нових методів. Одним із важливих кроків побудови моделі виявлення вторгнень є етап попередньої обробки даних. Важливою складовою цього етапу є виявлення ознак, які корелюють між собою.

Наявність корелюючих ознак негативно впливає на якість моделі, оскільки дані містять надмірну інформацію. Це призводить до того, що модель враховує ту саму інформацію кілька разів, що робить її менш ефективною і менш інтерпретованою. Корельовані ознаки роблять також модель менш стійкою і більш чутливою до шуму в даних. Зміна значень однієї ознаки може призвести до неправильних висновків щодо впливу на цільову змінну через кореляцію з іншою ознакою. Крім того, коли ознаки сильно корелюють, стає складніше інтерпретувати вплив кожної ознаки на цільову змінну. Це може бути проблемою у випадках, коли важливо зрозуміти, саме які ознаки впливають на передбачення моделі. Наявність кореляції в даних може призвести до проблеми мультиколлінеарності, коли матриця ознак стає близькою до сингулярної або незворотної. Це може ускладнити навчання лінійних моделей, оскільки вони стають нестійкими. Також корельовані ознаки збільшують складність моделі, підвищують час навчання моделі, без суттєвого покращення її продуктивності. Це може призвести до перенавчання, особливо у випадках, коли даних мало.

Об’єктом дослідження є процес ідентифікації стану комп'ютерної мережі.

Предметом дослідження є методи ідентифікації стану комп’ютерних мереж.

Постановка проблеми та огляд наукових публікацій

Кореляція між ознаками оцінюється з використанням різних статистичних метрик та методів. Найбільш поширені способи оцінки кореляції ознак включають:

– коефіцієнт кореляції Пірсона

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sqrt{S_x^2 S_y^2}},$$

де $\bar{x}, \bar{y}$ – вибіркові середні x^m, y^m, S_x^2, S_y^2 – вибіркові дисперсії, $r_{xy} \in [-1, 1]$.

– коефіцієнт кореляції Спірмена

$$r_s = 1 - \frac{6\sum(x_i - y_i)^2}{N(N^2 - 1)},$$

де $x_i - y_i$ – різниця між спряженими значеннями рангів змінних x і y.

На відміну від коефіцієнта Пірсона, який вимірює лінійний зв'язок, коефіцієнт Спірмена може виявляти інші типи зв'язку, наприклад, монотонний нелінійний зв'язок.

Для боротьби з проблемою кореляції ознак, використовують:

– видалення однієї з корелюючих ознак;

– застосування методів відбору ознак, таких як, взаємна інформація (mutual information)[1,2] або використання моделей, що дозволяють оцінити інформативність ознак [3-5];

– перетворення ознак, наприклад, використання методу головних компонентів (Principal Component Analysis, PCA)[6] або методу незалежних компонентів (Independent Component Analysis, ICA)[7].

– Використання регуляризації (у разі лінійних моделей), наприклад, L1 або L2 регуляризація, щоб стримувати вплив кореляційних ознак[8].

Разом із тим, видалення або трансформація корелюючих ознак може призвести до погіршення якості моделі з кількох причин. Корелюючі ознаки можуть містити корисну інформацію щодо залежності між змінними. При видаленні однієї з них, модель втрачає доступ до цієї інформації та може стати менш інформативною.

Крім того, ознака, яка корелює, може мати високу інформативність і її видалення вплине на якість моделі. Модель може стати менш здатною до виявлення та врахування залежності у даних, що знижує точність та здатність моделі до передбачення цільової змінної. Видалення корелюючих ознак може зменшити мультиколінеарність, що у деяких випадках призводить до перенавчання моделі, особливо якщо ознака, що видаляється, містить важливу інформацію для моделювання. Також видалення ознак, які корелюють, зазвичай, ускладнює інтерпретацію моделі та пояснення результатів.

Щоб вирішити, дане протиріччя, потрібно оцінити, як видалення ознак, що корелюють, впливає на якість, оперативність та інтерпретованість моделі. У деяких випадках це може бути корисним для поліпшення продуктивності моделі, а в інших – може негативно вплинути на її здатність робити точні прогнози.

На сьогодні, найбільш популярними методами трансформації даних, які корелюють є: регуляризація, метод головних компонентів (PCA)[9], метод незалежних компонентів (ICA).

Регуляризація, може зробити модель більш простою та менш гнучкою. Це може бути небажаним, якщо дані мають складні взаємозв'язки, які модель має враховувати. Занадто сильна регуляризація, спрямована на зменшення кореляції та може призвести до недонавчання моделі, особливо якщо дані мають суттєві залежності між ознаками. Крім того, визначення оптимального рівня регуляризації для зменшення кореляції може бути складним завданням. Неправильний вибір параметрів регулювання може призвести до небажаних ефектів.

Метод незалежних компонентів передбачає, що приховані компоненти є статистично незалежними та мають безперервні розподіли. Однак у реальних даних може бути безліч винятків із цього припущення. Крім того, якщо дані містять складні нелінійні залежності, ICA може працювати менш ефективно. Визначення правильної кількості прихованих компонентів у методі ICA також є складним завданням. Неправильно вибрана кількість компонентів може призвести до втрати інформації або небажаних ефектів. Метод чутливий до шуму даних.

Метод головних компонентів також може бути неефективним для даних із нелінійними залежностями між ознаками та потребує масштабування даних. PCA чутливий до викидів у даних, тому при використанні цього методу необхідно проаналізувати дані та видалити викиди. Незважаючи на ці обмеження, PCA є ефективним інструментом для зменшення розмірності та зменшення кореляції даних, особливо для даних із лінійною залежністю між ознаками.

При виборі методу, важливо враховувати його обмеження та вибирати метод залежно від характеристик даних та мети аналізу. Для подальшого дослідження обрано метод головних компонент, якому передує етап попередньої обробки даних.

Метод головних компонент (Principal Component Analysis, PCA) виконує проекцію даних у новий набір змінних, які називаються головними компонентами[10], зберігаючи, при цьому, найбільшу кількість інформації, зменшує розмірність даних та прибирає кореляцію між ознаками.

Метод PCA працює так:

Крок 1. Виконується стандартизація ознак (середнє значення ознаки зводиться до нуля, стандартне відхилення – до одиниці)

$$x_{ij} = \frac{y_{ij} - \overline{y_j}}{s_j}, i = 1, n; j = 1, m,$$

де y_{ij} – елемент матриці вхідних даних, $\overline{y_j}$ – середнє значення для j-го фактора (ознаки), s_j – стандартне відхилення для j-го фактора.

Крок 2. Обчислюється коваріаційна матриця, яка показує, які ознаки корелюють між собою. Коваріаційна матриця є симетричною матрицею, де кожен елемент показує ступінь кореляції між двома ознаками

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - X')(Y_i - Y')}{n-1},$$

де X' - арифметичне середнє даних X, Y' - арифметичне середнє даних Y, n – розмір вибірки.

Крок 3. Знаходяться власні значення та власні вектори коваріційної матриці (власні значення показують, яка дисперсія притаманна кожній головній компоненті, а власні вектори є напрямами у просторі ознак).

Крок 4. Виконується сортування основних компонентів за зменшенням власних значень, тобто перші компоненти мають найбільшу дисперсію даних.

Крок 5. Вирішується, скільки основних компонентів залишити. Чим більше компонентів залишено, тим більше інформації зберігається, але розмірність даних залишається високою.

Крок 6. Виконується проекція вихідних даних на новий простір ознак. Цей новий простір матиме меншу розмірність і буде лінійними комбінаціями вихідних ознак.

Розробка методу підвищення оперативності моделі класифікації даних за рахунок зменшення кореляції ознак

У цій роботі у якості вихідних даних використано набір UNSW-NB 15 , який був розроблений у лабораторії Cyber Range Австралійського центру кібербезпеки (ACCS) та містить інформацію про нормальне функціонування мережі та під час синтетичних вторгнень.

Набір даних містить атрибути, які характеризують наявність або відсутність таких типів атак:

Analysis, Backdoors, DoS, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode and Worms.

Для дослідження ефективності моделі у середовищі GOOGLE Colab Python розроблено їх програмні моделі.

Якість моделі оцінено за допомогою таких характеристик:

точність (Accuracy),
влучність (Precision),
повнота (Recall),
міра F1 (F1 score).

У даній роботі запропоновано метод виявлення вторгнень в комп'ютерні мережі, який відрізняється від відомих наявністю спеціальної процедури зменшення кореляції вихідних даних, що дозволило підвищити оперативність процесу ідентифікації.

Процедура зменшення кореляції вихідних даних базується на наступному алгоритмі:

Крок 1. Всі категоріальні ознаки перетворюємо на числові значення методом factorize(). Заповнюємо пропуски даних.

Крок 2. Аналізуємо ознаки та видаляємо з набору даних неінформативні ознаки.

Крок 3. Будуємо кореляційну матрицю.

Крок 4. Якщо є ознаки що корелюються між собою більш ніж заданий параметр кореляції, (наприклад, 90%), то обробляємо їх методом головних компонент (PCA). Для цього формуємо датафрейми з двох ознак, що максимально корелюють між собою та застосовуємо метод головних компонент. Кожний набір перетворюємо на нову ознаку. Після формування нових ознак видаляємо старі та додаємо нові ознаки до основного набору даних.

Крок 5. Будуємо модель та оцінюємо її якість. Якщо якість моделі суттєво не змінилася та вище заданого порогового значення, то повертаємося до кроку 4, інакше до кроку 6.

Крок 6. Якщо точність моделі суттєво знизилась, то виконуємо аналіз ознак які були оброблені методом головних компонент на кроці 4 і приймаємо рішення щодо їх відновлення. Завершуємо процес.

Відповідно до вищенаведеного алгоритму виконано попередню обробку даних набору UNSW-NB 15. За результатами аналізу видалено з набору даних такі ознаки: «id, attack_cat, label» оскільки вони не містять корисної інформації. Попередній аналіз даних показав наявність в даних ознак, що корелюють (рис. 1). Дані які мають кореляцію вище 90% були спроектовані у новий набір даних за допомогою методу головних компонент (PCA).

Дослідним шляхом було виявлено, що якщо обробити методом головних компонент або видалити будь яку з ознак «ct_srv_src» або «ct_dst_src_ltm», то точність класифікації помітно знижується. Тому вирішено залишити ці ознаки в датасеті попри їх кореляцію у 94 відсотки. Таким чином, за результатами експертного аналізу та використання вищенаведеного алгоритму 42 ознаки датасету трансформовано в 31 ознаку.

Надалі виконано нормалізацію даних. Крім того, оскільки в цьому наборі класи розподілені приблизно порівну (з 82332 загальної кількості 45332 це атаки), то балансування класів не виконувалась.

Для оцінки якості розробленого методу досліджено ансамблеві моделі на основі алгоритмів: Gradient Boosting, Random Forest (за умови стандартного налаштування параметрів моделі) та нейронні мережі.

Результати дослідження ансамблевих моделей наведено в табл. 1.

```
corr = pre_drop_df.corr()
corr.style.background_gradient(cmap='coolwarm')
```

	dur	proto	service	state	spkts	dpkts	sbytes	dbytes	rate	sttl	dttl	sload	dload	sloss
dur	1.000000	-0.053138	-0.104148	0.079980	0.280239	0.217507	0.225432	0.172492	-0.118032	-0.000990	0.090048	-0.076344	-0.047033	0.240113
proto	-0.053138	1.000000	-0.239996	-0.228909	-0.033177	-0.040910	-0.012132	-0.023490	0.208858	0.199680	-0.221595	0.139353	-0.071932	-0.019758
service	-0.104148	-0.239996	1.000000	-0.135552	-0.026797	-0.046527	0.004752	-0.033948	0.286647	0.112926	-0.378092	0.002966	-0.143624	-0.004532
state	0.079980	-0.228909	-0.135552	1.000000	0.050412	0.045430	0.033448	0.025496	-0.394435	-0.537155	0.295439	-0.270630	0.076395	0.041305
spkts	0.280239	-0.033177	-0.026797	0.050412	1.000000	0.369554	0.965750	0.198324	-0.068249	-0.092536	0.054601	-0.044194	0.074440	0.973644
dpkts	0.217507	-0.040910	-0.046527	0.045430	0.369554	1.000000	0.175834	0.976419	-0.083173	-0.163830	0.036483	-0.054145	0.133835	0.189060
sbytes	0.225432	-0.012132	0.004752	0.033448	0.965750	0.175834	1.000000	0.010036	-0.025102	-0.017866	0.049891	-0.015228	-0.006428	0.995027
dbytes	0.172492	-0.023490	-0.033948	0.025496	0.198324	0.976419	0.010036	1.000000	-0.047978	-0.114537	0.012537	-0.031266	0.100923	0.014561
rate	-0.118032	0.208858	0.286647	-0.394435	-0.068249	-0.083173	-0.025102	-0.047978	1.000000	0.388155	-0.453913	0.550104	-0.138441	-0.040139
sttl	-0.000990	0.199680	0.112926	-0.537155	-0.092536	-0.163830	-0.017866	-0.114537	0.388155	1.000000	-0.033338	0.252901	-0.386224	-0.038088
dttl	0.090048	-0.221595	-0.378092	0.295439	0.054601	0.036483	0.049891	0.012537	-0.453913	-0.033338	1.000000	-0.293939	-0.139491	0.061249
sload	-0.076344	0.139353	0.002966	-0.270630	-0.044194	-0.054145	-0.015228	-0.031266	0.550104	0.252901	-0.293939	1.000000	-0.092772	-0.025938
dload	-0.047033	-0.071932	-0.143624	0.076395	0.074440	0.133835	-0.006428	0.100923	-0.138441	-0.386224	-0.139491	-0.092772	1.000000	0.009210
sloss	0.240113	-0.019758	-0.004532	0.041305	0.973644	0.189060	0.995027	0.014561	-0.040139	-0.038088	0.061249	-0.025938	0.009210	1.000000
dloss	0.171182	-0.030430	-0.037502	0.029645	0.198683	0.981506	0.007091	0.997109	-0.062073	-0.137737	0.021966	-0.040456	0.117303	0.014661
sinpkt	0.079840	-0.036653	-0.086310	-0.061994	-0.014501	-0.017141	-0.005399	-0.010201	-0.065681	-0.179270	-0.075620	-0.041718	-0.032094	-0.008124
dinpkt	0.150801	-0.025898	-0.059731	0.076754	-0.003309	-0.007181	-0.001432	-0.007266	-0.052206	-0.006154	0.090734	-0.033787	-0.024481	-0.000470
sjit	0.146598	-0.030106	-0.067714	0.036344	-0.002407	-0.003862	-0.002675	-0.005182	-0.061961	0.030062	0.140634	-0.040018	-0.028759	-0.000422
djit	0.165419	-0.039621	-0.059706	0.032881	0.010481	0.034276	-0.003050	0.029201	-0.081591	-0.084072	0.113799	-0.052820	-0.031342	0.002057
swin	0.083993	0.281584	0.497066	0.240661	0.111828	0.144478	0.041450	0.083045	0.524075	0.270458	0.749894	0.361542	0.222838	0.070170

Рис. 1. Кореляційна матриця для даних UNSW-NB 15

Таблиця 1 – **Показники якості та оперативності моделей на основі алгоритмів: Gradient Boosting та Random Forest**

Показники якості та оперативності моделі	Дані після обробки запропонованим методом		Дані без обробки	
	Gradient Boosting	Random Forest	Gradient Boosting	Random Forest
Accuracy	0.959	0.975	0.962	0.978
Precision	0.960	0.975	0.960	0.975
Recall	0.960	0.975	0.960	0.980
F1-score	0.960	0.975	0.965	0.980
Час навчання, с	26.8	11	32	12
Час розпізнавання, с	0.04	0.3	0.06	0.6

Як видно із таблиці, зменшення кореляції ознак надало можливість зменшити час навчання моделі та підвищити час розпізнавання до 2 разів. При цьому точність моделі практично не змінилась.

У якості нейронних методів досліджено: повнозв'язну нейронну мережу (FCNN) та згорткову нейронну мережу (CNN).

Розробку моделі класифікації на основі повнозв'язної нейронної мережі виконано з використанням бібліотек "tensorflow" та "keras" GOOGLE COLAB Python. Нейронна мережа має один прихований шар, який містить 128 нейронів. При налаштування моделі використано функцію активації "relu", вихідний шар на два виходи з функцією "softmax, оптимізатор – "adam", функцію втрат – "categorical_crossentropy". Роботу моделі оцінюємо метриками якості: "accuracy", "precision", "recall", "f1-score".

Процес навчання містить десять епох з розміром батчу (вибірки) 64. Результати дослідження моделі на основі повнозв'язної нейронної мережі з одним прихованим шаром наведено в табл. 2.

Таблиця 2 **– Показники якості та оперативності моделі на основі повнозв'язної нейронної мережі FCNN**

Показники якості та оперативності моделі	Дані після обробки запропонованим методом	Дані без обробки
Accuracy	0.949	0.951
Precision	0.949	0.951
Recall	0.949	0.951
F1-score	0.948	0.951
Час навчання, с	13	24
Час розпізнавання, с	0.9	1.3

Як видно із таблиці, оперативність моделі збільшилась на етапі навчання до 1,9 разів, на етапі розпізнавання – до 1,4 разів. Згорткова нейронна мережа(Convolutional Neural Network, CNN) працює з багатовимірними даними, що потребує, у нашому випадку, перетворення та нормалізації даних. Так для датасету, попередньо обробленого за допомогою методу головних компонент (PCA), створено масив розмірністю (31, 31, 3), де перші два значення це висота та ширина кадру в пікселях а останнє значення це три канала кольору (RGB). Тобто кожен піксель представлений трьома значеннями каналів. Для необробленого датасету розмірність буде (42, 42, 3).

При побудові моделі використано такі параметри налаштування: тип оптимізатор – "adam", функція втрат – "categorical_crossentropy".

Процес навчання містить десять епох з розміром батчу (вибірки) 128.

Результати дослідження моделі на основі згорткової нейронної мережі (CNN) наведено в табл. 3.

Таблиця 3 **– Показники якості та оперативності моделі на основі згорткової нейронної мережі (CNN)**

Показники якості та оперативності моделі	Дані після обробки запропонованим методом	Дані без обробки
accuracy	0.990	0.993
precision	0.990	0.993
Recall	0.990	0.993
F1-score	0.989	0.993
Час навчання, с	28	37
Час розпізнавання, с	0.3	0.5

Як видно із таблиці, оперативність моделі збільшилась на етапі навчання до 1,3 разів, на етапі розпізнавання – до 1,6 разів.

Висновки

У рамках дослідження виконано аналіз різних підходів до підвищення оперативності моделей виявлення вторгнень у комп'ютерні мережі за умови наявності ознак, які корелюють.

Розглянуто найбільш поширені способи оцінки кореляції ознак. Досліджено такі підходи зменшення кореляції ознак: метод головних компонентів (Principal Component Analysis, PCA), метод незалежних компонентів (Independent Component Analysis, ICA), L1 та L2 регуляризація.

Запропоновано спеціальну процедуру зменшення кореляції вихідних даних. Розроблено програмні моделі ідентифікації стану комп'ютерних мереж на основі алгоритмів: Gradient Boosting, Random Forest, повнозв'язної нейронної мережі (FCNN) та згорткової нейронної мережі (CNN). У якості вихідних даних використано набір UNSW-NB 15, який містить інформацію про нормальне функціонування мережі та під час вторгнень. Виконано їх порівняльний аналіз.

Отримано, що зменшення кореляції ознак надало можливість підвищити оперативність ідентифікації стану комп'ютерних мереж до 1,9 разів на етапі навчання та до 2 разів на етапі розпізнавання. При цьому точність моделі практично не змінилась

Таким чином, в даній роботі запропоновано метод виявлення вторгнень в комп'ютерні мережі, який відрізняється від відомих наявністю спеціальної процедури зменшення кореляції вихідних даних, що дозволило підвищити оперативність процесу ідентифікації стану комп'ютерної мережі.

Подальші дослідження будуть спрямовані на дослідження моделей ідентифікації стану комп'ютерної мережі на основі моделі "Vision Transformer for Small-Size Datasets".

СПИСОК ЛІТЕРАТУРИ

1. Vergara, J.R., Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput & Applic* **,** 2014, Vol.24, pp.175–186. https://doi.org/10.1007/s00521-013-1368-0
2. Hoque, N., Bhattacharyya, D.K., & Kalita, J.K. MIFS-ND: A mutual information-based feature selection method. *Expert Syst. Appl.,* 2014, Vol.41, 6371-6385.https://www.researchgate.net/publication/262526444_MIFS-ND_A_mutual_information-based_feature_selection_method
3. Smita Chormunge, Sudarson Jena. (). Correlation based feature selection with clustering for high dimensional data. *Journal of Electrical Systems and Information Technology*, 2018, Vol. 5 (3), pp.542-549. https://doi.org/10.1016/j.jesit.2017.06.004
4. Hall, M.A. (). Correlation-based feature selection of discrete and numeric class machine learning, *Working paper*. Hamilton, New Zealand: University of Waikato, Department of Computer Science., 2000, pp.1-10. https://hdl.handle.net/10289/1024
5. Krzysztof Michalak, Halina Kwasnicka. Correlation–based feature selection strategy in classification problems. Int. J. Appl. Math. Comput. Sci., 2006, Vol. 16(4), pp.503–511. https://bibliotekanauki.pl/articles/908379.pdf
6. Ibrahim, S.; Nazir, S.; Velastin, S.A. Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis. J. Imaging, 2021, Vol.7, pp. 225-241. https://doi.org/10.3390/jimaging7110225
7. F. Vrins, J. A. Lee, M. Verleysen, V. Vigneron and C. Jutten, "Improving independent component analysis performances by variable selection," 2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718), Toulouse, France, 2003, pp. 359-368, doi: https://10.1109/NNSP.2003.1318035 .
8. Ng. Andrew. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004, pp 78-84. https://dl.acm.org/doi/10.1145/1015330.1015435
9. Karen Garate-Escamilla A, Hassani AHE, Andres E, Classification models for heart disease prediction using feature selection and PCA, *Informatics in Medicine Unlocked* , 2020, pp. doi: https://doi.org/10.1016/j.imu.2020.100330
10. Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., & Saikhom, R. et al. Multivariate Statistical Data Analysis- Principal Component Analysis (PCA). International Journal of Livestock Research, 2017, Vol.7(5), pp. 60-78. http://doi.org/10.5455/ijlr.20170415115235

Received (Надійшла) 11.09.2023
Accepted for publication (Прийнята до друку) 22.11.2023

Method of increasing the efficiency of data classification at the account of reducing the correlation of the sign

Svitlana Gavrylenko, Vadym Poltoratskyi

Abstract. The object of the study is the process of identifying the state of the computer network. The subject of research is methods of identifying the state of computer networks. The purpose of the article is to increase the efficiency of detecting intrusions into computer networks by reducing the correlation of features. Methods used: methods of artificial intelligence, machine learning, methods of reducing the correlation of features. The following results were obtained: The effectiveness of using approaches that reduce data correlation was investigated: the method of principal components (PCA), independent components (ICA), L1 and L2 regularization, the method was justified for further research. According to the research results, a special procedure for reducing the correlation of the initial data is proposed. To evaluate the quality and efficiency of the proposed procedure, software models based on: Gradient Boosting, Random Forest, fully connected neural network (FCNN) and convolutional neural network (CNN) were developed. The UNSW-NB 15 set, which contains information on normal network functioning and during intrusions, was used as the source data. A comparative analysis of the quality and efficiency of the developed models was performed. Conclusions. The scientific novelty of the obtained results lies in the development of a method for detecting intrusions into computer networks, which differs from known methods by the presence of a special procedure for reducing the correlation of the output data, which made it possible to increase the efficiency of the identification process.

Keywords: machine learning, data classification, data preprocessing, data correlation, computer networks, neural networks, ensemble classifiers, intrusion detection systems.