Problems of identification in information systems

УДК 004.732.056 **doi:** https://doi.org/10.20998/2522-9052.2021.2.01

С. Ю. Гавриленко, І. В. Шевердін, Г. В. Гейко

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

ОЦІНКА ІНФОРМАТИВНОСТІ ТА ВИБІР ОЗНАК ПРИ ІДЕНТИФІКАЦІЇ СТАНУ КОМП’ЮТЕРНОЇ СИСТЕМИ

Анотація. Предметом статті є дослідження методів визначення інформативності ознак. **Метою статті** є підвищення якості класифікації стану комп’ютерної системи за рахунок вибору найбільш інформативних ознак. **Завдання**: дослідити методи вибору оптимальних інформаційних ознак для ідентифікації стану комп’ютерної системи на основі аналізу подій операційної системи *Windows*. Використовуваними **методами** є: методи машинного навчання, ансамблеві методи, методи вибору оптимальних інформаційних ознак. Отримано такі **результати**: виконано аналіз подій операційної системи *Windows*, досліджено методи вибору оптимальних інформаційних ознак: методи-обгортки (*Wrappers*), вбудовані методи (*Embedded*) і методи-фільтри (*Filters*). Виконано оцінку інформативності та вибір ознак при ідентифікації стану комп’ютерної системи. Для оцінки ефективності вибраних ознак було використано ансамблевий метод класифікації стану комп’ютерної системи на основі беггінгу та дерева рішень J48. Досліджено залежність точності класифікації стану комп’ютерної системи від обраних ознак та визначено набір атрибутів, які забезпечують максимальну точність класифікації стану комп’ютерної системи. **Висновки**. Наукова новизна отриманих результатів полягає у аналізі подій операційної системи *Windows*, оцінці їх інформативності та вибору ознак при ідентифікації стану комп’ютерної системи.

Ключові слова: комп’ютерна система; події операційної системи; інформативність ознак; дерева рішень; ансамблеві методи; беггінг.

Вступ

Сьогодні комп’ютерні системи використовуються практично у всіх галузях народного господарювання. Експлуатація таких складних технічних систем усе більшою мірою стикається з проблемами забезпечення їх інформаційної та функціональної безпеки. Разом із тим, рівень розвитку засобів діагностування та ідентифікації деструктивних змін режимів функціонування і внутрішніх характеристик таких систем на сьогодні не можуть гарантувати необхідний рівень захисту інформації. Саме тому дослідження методів та засобів ідентифікації стану комп’ютерних систем є актуальним завданням [1].

Комп’ютерна система характеризується великим обсягом показників її функціонування. Для цього широко використовуються алгоритми машинного навчання, які збудовані таким чином, щоб безпосередньо працювати з величезними масивами інформації [2].

Разом із тим, на етапах постановки задачі машинного навчання і формування даних не завжди зрозуміло, які ознаки важливі для побудови оптимального алгоритму. Крім того, дані можуть містити багато надлишкової (шумової) інформації, що погіршує якість роботи алгоритму і уповільнює його роботу. Тому в більшості випадків перед вирішенням завдання класифікації, необхідно вибрати ті ознаки, які найбільш інформативні. Відбір важливих ознак також може допомогти розшифрувати механізми, що лежать в основі проблематики дослідження [3].

Постановка проблеми та огляд наукових публікацій. Ознаки (*feature*), що використовуються для побудови моделі, істотно впливають на якість результатів.

Неінформативні або слабо інформативні ознаки можуть значно знизити ефективність моделі.

Відбір ознак – це процес вибору ознак, що мають найбільш тісні взаємозв’язки з цільової змінної.

Метою відбору ознак є:

- спрощення моделей;
- зменшення ймовірності перенавчання, тобто чим менше надлишкових даних, тим менше можливостей для моделі приймати рішення на основі «шуму»;
- підвищення точності, а саме чим менше суперечливих даних, тим вище точність;
- скорочення часу навчання, мається на увазі, чим менше даних, тим швидше навчається модель [4].

Передумовою застосування методики обирання ознак є те, що вихідні дані містять надлишкові або недоречні деякі ознаки, які можуть бути усунені без спричинення значної втрати інформації [5]. Завдання вибору оптимального набору ознак полягає в тому, щоб вибрати таку підмножину ознак з вихідного набору ознак, щоб **точність класифікатора, навченого на цій підмножині ознак, була максимальною** (по всій підмножини вихідної множини ознак).

Отже, нехай *I(T)* – алгоритм навчання, *T(X)* – навчальна вибірка, X – множина ознак навчальної вибірки *T(X)*, $X' \subseteq X$ – підмножина множини ознак, $T(X')$ – навчальна вибірка, побудована з використанням підмножини ознак X', D – класифікатор: $D = I(T(X'))$, $Q(D)$ – точність класифікатора. Тоді оптимальний набір ознак визначається так:

$$X_{opt} = \arg\max_{X' \subseteq X} Q(D), \quad D = I(T(X')).$$

Існує кілька підходів до вибору оптимальних інформаційних ознак. За поширеною класифікацією [6], існує три основні категорії методів вибору оптимальних інформаційних ознак: *методи-обгортки* (*Wrappers*), *вбудовані методи* (*Embedded*) і *методи-фільтри* (*Filters*).

Методи-обгортки є універсальними та якісними однак вимагають великих обчислювальних витрат і особливих зусиль по оцінці класифікатора і вибору найкращої стратегії пошуку [7]. *Вбудовані методи* виконують відбір ознак під час процедури навчання класифікатора, є швидкими. Однак, дані алгоритми не є універсальними [8]. *Методи-фільтри* є найбільш поширеними, мають найменшу обчислювальною складністю серед розглянутих підходів, а також масштабність і простоту застосування [9]. Вони засновані на деяких показниках, які не залежить від методу класифікації (наприклад, такі як кореляція ознак з цільовим вектором, критерії інформативності). Крім того, подібні методи показують досить хороші результати на практиці, вони не взаємодіють з алгоритмом навчання і вибирають оптимальну підмножину ознак, використовуючи тільки інформацію, отриману з навчальної вибірки. Методи-фільтри виконуються на етапі **попередньої обробки**, до виконання алгоритму навчання. Вони можуть як незалежно оцінювати інформативність ознак для навчання, так і оцінювати підмножину ознак **в сукупності**. У першому випадку знадобиться визначити значення порогової константи (потрібної для того, щоб відкинути ті ознаки, інформативність яких для алгоритму навчання нижче значення порога). У другому випадку знадобиться проводити пошук по простору підмножин ознак [10].

Постановка завдання. Метою дослідження є підвищення якості класифікації стану комп'ютерної системи за рахунок вибору найбільш інформативних ознак.

Вибір підходу та критеріїв інформативності ознак для класифікації стану комп'ютерної системи

Аналіз подій операційної системи надає можливість оцінити стан комп'ютерної системи. Для збору подій операційної системи *Windows* 10 було використано програмний додаток *ProcessMonitor* (рис. 1).

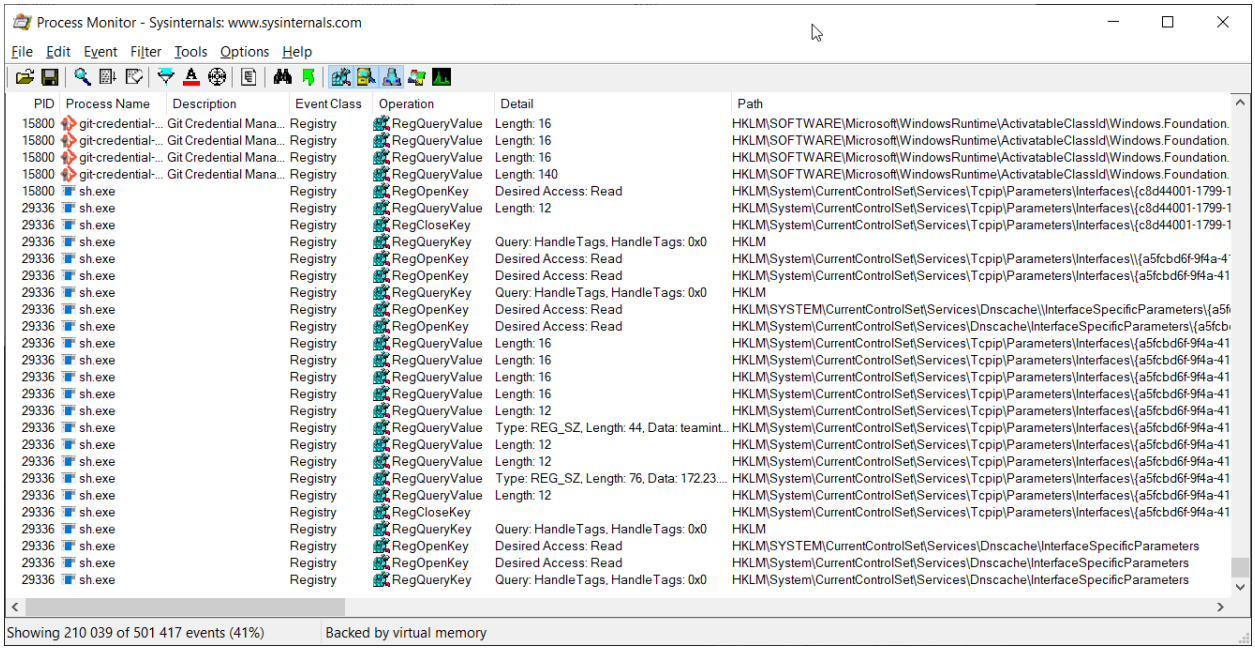

Рис. 1. Використання *ProcessMonitor* для збору подій (**Fig. 1.** Use ProcessMonitor to collect events)

Отримано, що КС характеризується великою кількістю подій та атрибутів. Основні труднощі класифікації її стану полягають в дуже великому розмірі простору ознак, яке носить назву прокляття розмірності. Це призводить до зниження ефективності оцінки її стану та потребує вилучення нерелевантних, або надлишкових ознак із заданого вектору атрибутів [11].

У якості підходу до вибору оптимальних інформаційних ознак використано метод-фільтрації на базі інформаційної ентропії, який має невелику обчислювальну складністю, масштабність і простоту застосування.

Інформаційна ентропія – міра невизначеності деякої системи (в статистичній фізиці або теорії інформації), зокрема непередбачуваність появи будь-якого символу первинного алфавіту. В останньому випадку при відсутності інформаційних втрат ентропія чисельно дорівнює кількості інформації на символ переданого повідомлення. Ентропія характеризує чистоту довільного набору. Це лежить в основі методів ранжування атрибутів. Міра ентропії розглядається як міра непередбачуваності системи [12].

Інформаційна ентропія визначається як:

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 (p(y)),$$

де $p(y)$ – гранична функція щільності ймовірності для випадкової величини Y.

Якщо спостережувані значення Y у навчальному наборі даних S розподіллено відповідно до значень

другої ознаки *X*, а ентропія *Y* відносно частин індукованих ознакою *X* є меншою за ентропію ознаки *Y* до розділення, тоді існує зв'язок між ознаками *Y* і *X*. Інтуїтивно, ентропія відповідає ступеню хаосу в системі. Чим вище ентропія, тим менше впорядкована система і навпаки.

Тоді відносна ентропія *Y*/*X* знаходиться так:

$$H(Y \mid X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x) \log_2 (p(y \mid x)),$$

де *p(y|x)* – умовна ймовірність *y* заданого *x*.

Враховуючи ентропію, як критерій домішки в навчальному наборі *S*, ми можемо визначити міру, що відображає додаткову інформацію про *Y*, яку надає *X* та представляє величину, на яку зменшується ентропія *Y* [13]. Тобто ми отримаємо інформацію (*InfoGainAttributeEval*) – існує кореляція між значеннями *X* і *Y* чи ні.

InfoGainAttributeEval (*IG*) визначає приріст інформації і знаходиться як:

$$IG = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y) = \\ = H(Y) + H(X) - H(X, Y).$$

Приріст інформації (*IG*) для ознаки *X* визначається як різниця ентропії вибірок, отриманих без використання інформації ознаки *X* і з використанням цієї інформації.

Значення приросту інформації для ознаки говорить про різницю біт інформації, необхідних для того, щоб класифікувати об'єкт з використанням ознаки *X* і без її використання.

Чим більше параметр *IG* – тим сильніше кореляція між значеннями.

SymmetricalUncertAttributeEval (*SU*), а саме приріст інформації є симетричним показником – тобто кількість інформації, отриманої про *Y* після спостереження *X*, дорівнює кількості інформації, отриманої про *X* після спостереження *Y*. Симетрія є бажаною властивістю для міри взаємозв'язку ознак. На жаль, отримання інформації є упередженим на користь функцій із більшою кількістю значень. Симетрична невизначеність *SU* компенсує схильність інформаційного приросту *IG* до атрибутів з більшими значеннями і нормалізує його значення до діапазону [0,1]:

$$SU = 2.0 * \left[\frac{IG}{H(Y) + H(X)} \right].$$

GainRatioAttributeEval (*GR*) або коефіцієнт приросту інформації *GR* – це модифікація приросту інформації *IG*, яка зменшує її упередженість. Росс Квінлан запропонував зменшити упередження до багатозначних атрибутів, беручи до уваги кількість та розмір гілок при виборі атрибута:

$$GR = \frac{H(Y) - H(Y \mid X)}{H(X)}.$$

Коефіцієнт посилення долає проблему з приростом інформації, беручи до уваги кількість гілок, які виникли б до розбиття. Він коригує приріст інформації, беручи до уваги внутрішню інформацію *X* про розбиття. Таким чином, він вирішує недолік отримання інформації, а саме отримання інформації, що застосовується до атрибутів, які можуть приймати велику кількість різних значень [15].

CorrelationAttributeEval (*R*), а саме коефіцієнт кореляції Пірсона між двома змінними дорівнює коваріації двох змінних, або сумі добутків відхилень, поділеній на добуток їх стандартних відхилень [16]. Нехай, є дві вибірки

$$x^m = (x_1, ..., x_m), \; y^m = (y_1, ..., y_m).$$

Коефіцієнт кореляції Пірсона розраховують за формулою:

$$R = r_{xy} = \frac{\sum_{i=1}^{m} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \overline{x})^2 \sum_{i=1}^{m} (y_i - \overline{y})^2}} = \frac{\mathrm{cov}(x, y)}{\sqrt{s_x^2 s_y^2}},$$

де $\overline{x}, \overline{y}$ – вибіркові середні, x^m, y^m, s_x^2, s_y^2 – вибіркові дисперсії [14].

Таким чином, у якості критеріїв інформативності ознак для класифікації стану комп'ютерної системи надалі досліджено:

InfoGainAttributeEval (приріст інформації),

SymmetricalUncertAttributeEval (оцінка симетричності приросту інформації),

GainRatioAttributeEval (коефіцієнт приросту інформації),

CorrelationAttributeEval (коефіцієнт кореляції Пірсона).

Вибір оптимального набору інформативних ознак та оцінка точності класифікації стану комп'ютерної системи

Результати дослідження (рис. 1) дозволили оцінити інформативність ознак на базі алгоритмів:

CorrelationAttributeEval (*R*),

InfoGainAttributeEval (*IG*),

GainRatioAttributeEval (*GR*),

SymmetricalUncertAttributeEval (*SU*).

Як видно із рис. 2 найбільш інформативними атрибутами є:

Process ID – ім'я процесу;

Operation – тип операції (наприклад: (*RegOpenKey*, *CloseFile and etc.*);

Image Path – шлях до реєстру чи файлу, наприклад: *C:\Users\VirtualUser\Desktop\ZipFileSystemZipper.CSV* або *HKCU\Software\Classes\CLSID\{56AD4C5D-B908-4F85-8FF1-7940C29B3BCF}\Instance*;

Result – результат виконання операції, наприклад: *SUCCESS*, *REPARSE*, *NAME NOT FOUND*, *BUFFER OVERFLOW* та ін.;

Event Class – тип події (реєстр, міжпроцесна комунікація, інтернет комунікація, виведення на носії, наприклад: *File System*, *Registry* та ін.);

Image Path – шлях до виконуваного файлу, який ініціював подію (наприклад: *C:\Windows\Explorer.EXE*);

Company – розробник програмного продукту та процесу, який ініціював цю подію, наприклад: *Microsoft Corporation*;

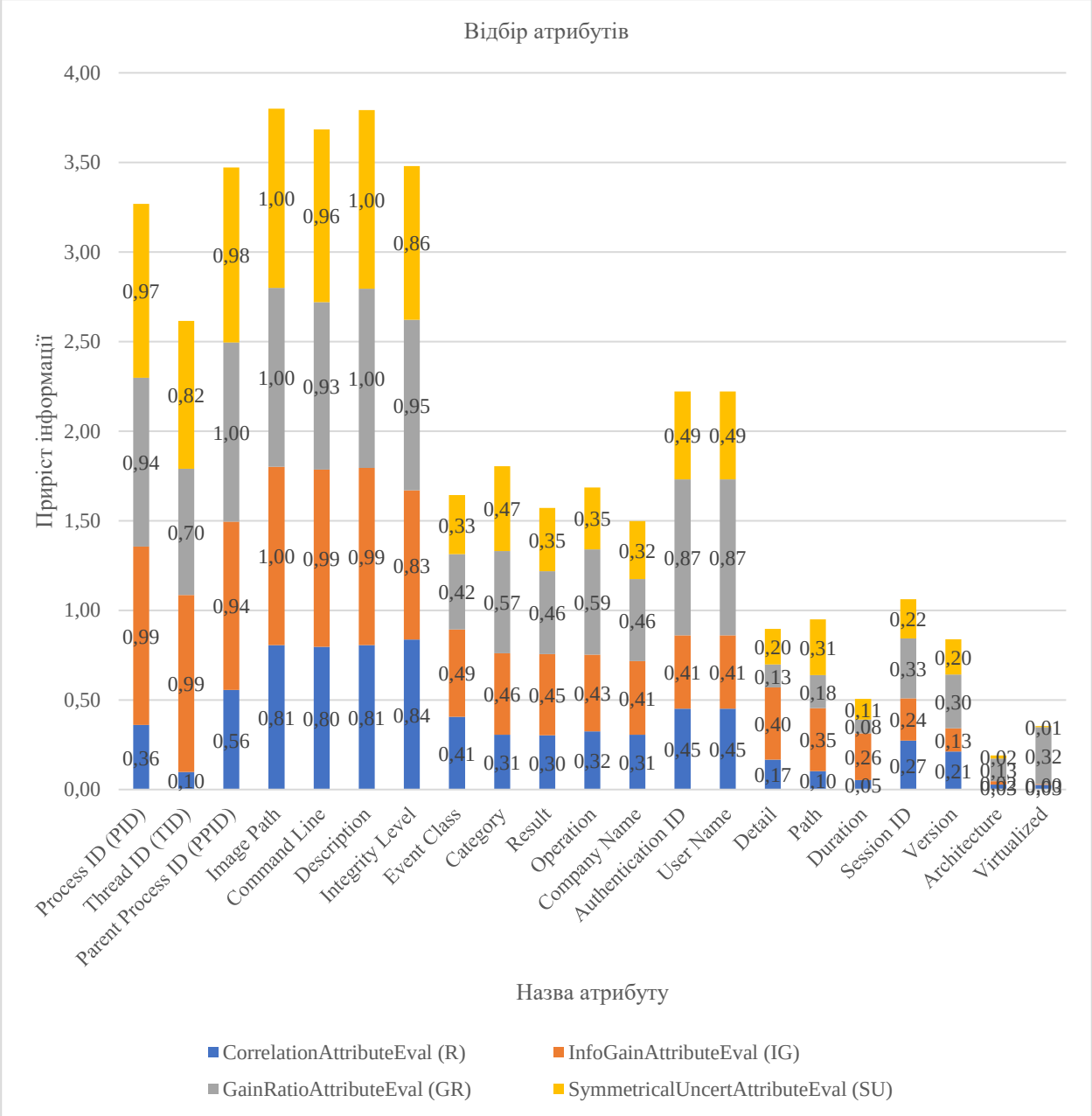

Рис. 2. Інформативність атрибутів (ознак) (**Fig. 2.** Informativeness of attributes (signs))

Description – опис програмного компоненту, наприклад: *Windows Explorer*;

User – имя користувача, який ініціалізував процес, наприклад: *DESKTOP*-159*T*3*OE**VirtualUser*;

Command Line – параметри командного рядка, наприклад: *C*:\ *Windows**System*32*svchost.exe -k LocalServiceNetworkRestricted -p*;

Integrity – пріоритет і важливість виконуваної події, наприклад: *System*, *Medium*, *High*, *Low*;

Category – тип операції, наприклад: *read*, *write*, *read metadata*, *write metadata* та ін.;

Authentication ID – *ID* користувача з метою виявлення перейменування користувачів і груп, наприклад: 00000000:000270*cb*;

Крім того, використовуючи параметри *Process ID*, *Process Name* і *Image Path* можна з точністю ідентифікувати ініціатора події, наприклад, конкретний виконуваний файл і процес [17].

Дослідження впливу вибору ознак на точність класифікації стану комп'ютерної системи

Подальші дослідження були пов'язані з оцінкою впливу вибору ознак на якість класифікації стану комп'ютерної системи.

У якості інструменту для аналізу стану КС та оцінки ансамблевих класифікаторів було обрано ПЗ "*Weka* (*Waikato Environment for Knowledge Analysis*)" [17], яке містить набір засобів віртуалізації і компонентів для інтелектуального аналізу даних та вирішення завдань прогнозування. У якості класифікатору використано ансамблевий метод класифікації стану комп'ютерної системи на основі алгоритму побудови дерева рішень *J*48 [12].

Для експерименту було сформовано базовий набір атрибутів з найбільшими значеннями приросту

інформації, а саме більше 80% на базі результатів показників описаних раніше функцій R, GR, SU, IG.

Результати першого експерименту (рис. 3) надали можливість оцінити точність класифікації Accurancy (A) при переборі значень атрибутів. При цьому менш інформативні атрибути додавалися до базових окремо один від одного і наділі визначалася точність класифікації стану КС.

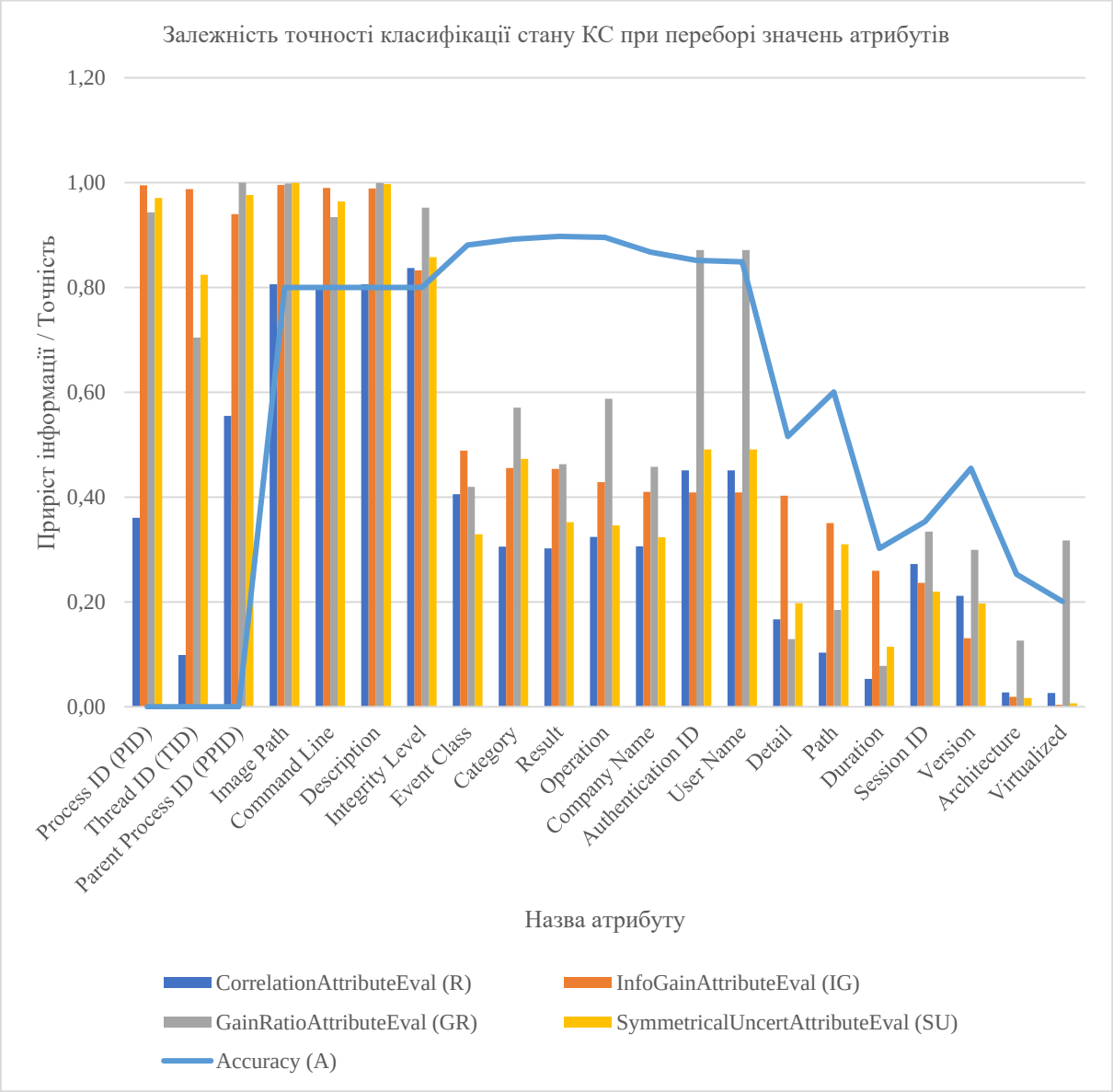

Рис. 3. Результати першого експерименту (**Fig. 3**. The results of the first experiment)

Як видно із рис. 3, точність класифікації при додаванні майже кожної ознаки зростала. Але починаючи з ознаки *Detail* точність класифікації стала зменшуватися.

Результати другого експерименту дозволили отримати акумулятивну точність класифікації Accumulated Accurancy (AA), де на кожному кроці додавали крок за кроком відібрані у першому експерименті ознаки та визначали точність класифікації залишаючи вибрані атрибути у наборі (рис. 4).

Аналіз інформаційного навантаження атрибутів *Process ID* (*PID*), *Thread ID*, *Parent Process ID* (*PPID*), є великими та потенційно вони можуть зменшити інформаційну ентропію, однак, вони не мають інформаційного навантаження, так як за цими значеннями атрибутів не можливо ідентифікувати стан комп'ютерної системи. Тому вищенаведені атрибути вилучені із набору інформативних ознак для ідентифікації стану КС.

Таким чином, за результатами експериментів для ідентифікації стану КС було виділено наступні атрибути: Process ID, Operation, Image Path, Result, Event Class, Image Path, Company, Description, User, Command Line, Integrity, Category, Authentication ID. Сукупна точність при використанні обраних атрибутів має значення 98,32%.

Висновки

В даній роботі виділено події функціонування операційних системах сімейства *Windows. Р*озглянуто підходи до вибору оптимальних інформаційних ознак для ідентифікації стану КС.

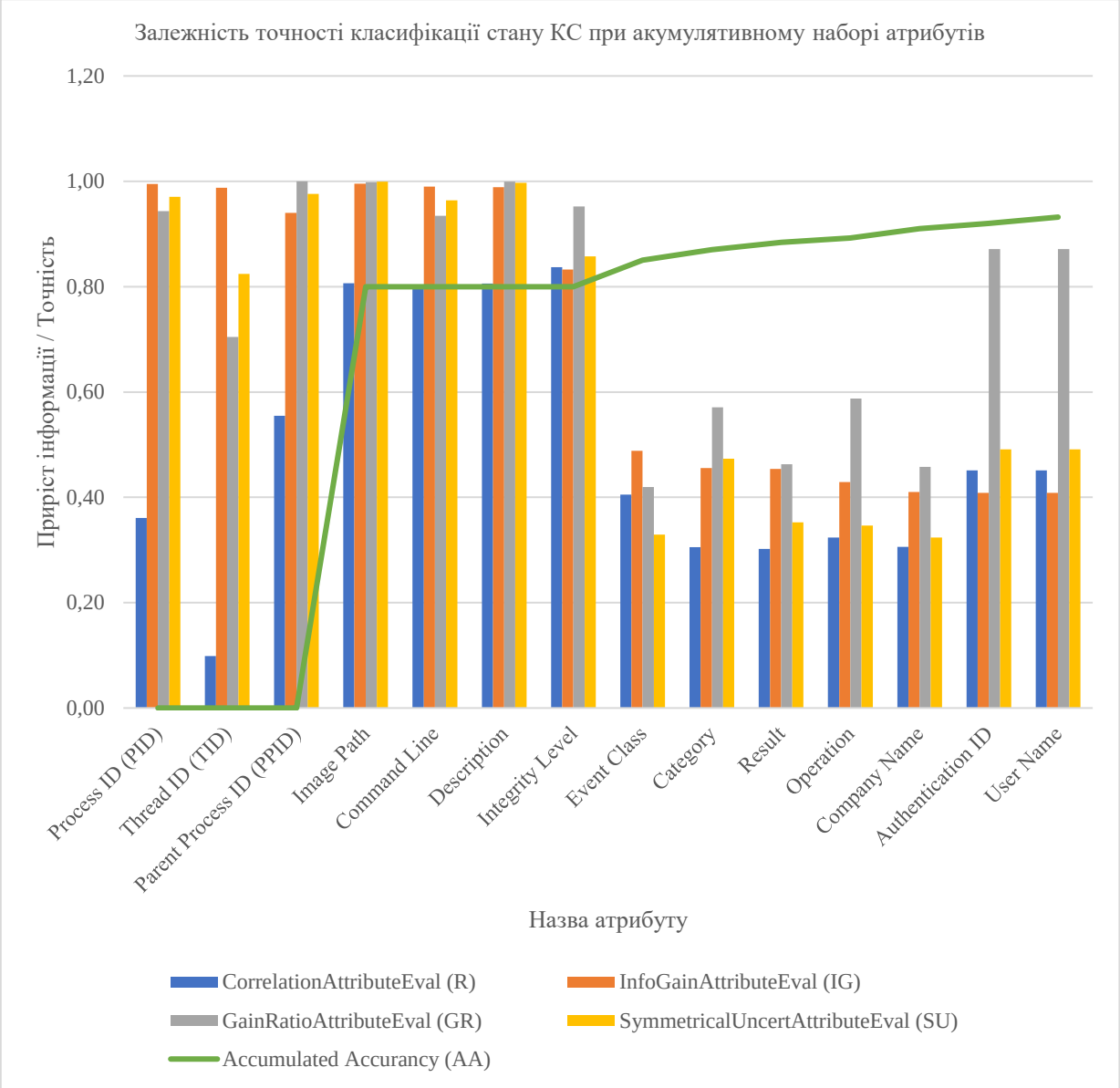

Рис. 4. Результати другого експерименту (**Fig. 4**. The results of the second experiment)

Досліджено методи вибору оптимальних інформаційних ознак:

методи-обгортки (Wrappers),

вбудовані методи (Embedded),

методи-фільтри (Filters).

Отримано, що найбільш поширеними є методи-фільтри, які мають найменшу обчислювальною складність, масштабованість, простоту застосування, не взаємодіють з алгоритмом навчання і вибирають оптимальну підмножину ознак, використовуючи тільки інформацію, отриману з навчальної вибірки.

Інформативність виділених події функціонування операційних системах сімейства *Windows* була оцінена з використанням алгоритмів:

CorrelationAttributeEval (*R*),

InfoGainAttributeEval (*IG*),

GainRatioAttributeEval (*GR*),

SymmetricalUncertAttributeEval (*SU*).

Виконано дослідження впливу вибраних ознак на точність класифікації стану комп'ютерної системи.

У якості класифікатору використано ансамблевий метод класифікації стану комп'ютерної системи на основі алгоритму побудови дерева рішень *J*48. За результатами дослідження визначено набір атрибутів, які забезпечують максимальну точність класифікації стану комп'ютерної системи.

Подальші дослідження технологій ідентифікації стану комп'ютерної системи можуть бути виконані в системах протидії комп'ютерного вторгнення.

СПИСОК ЛІТЕРАТУРИ

1. Andreea Bendovschi. Cyber-Attacks – Trends, Patterns and Security Countermeasures. *7th International conference on financial criminology 2015,* 13-14 April 2015,Wadham College, Oxford, United Kingdom. P. 24-31.
2. Кульбак С. Теория информации и статистика. М.: Наука, 1967. 408 с.

3. R Kohavi, G. John. Wrappers tor feature selection. *Artificial Intelligence*. 91(1-2): 1997. P. 273-324.
4. Isabelle Guyon, Andre Elissceff. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 3(2003). 2003. P.1157-1182.
5. Molina L.C., Belanche L., Ncbot A. Feature Selection Algorithms: A Survey And Experimental Evaluation. *Proceedings of the 2002 IEEE International Conference on Data Mining*, IEEE Computer Society: 2002. P. 306-313.
6. Stańczyk U. Feature Evaluation by Filter, Wrapper, and Embedded Approaches. In: Stańczyk U., Jain L. (eds) Feature Selection for Data and Pattern Recognition. Studies in Computational Intelligence. 2015. Springer, Berlin, Heidelberg. Vol. 584. 568 p.
7. Phuong T. M., Lin Z., Altman R. B. Choosing SNPs using feature selection. Archived at the Wayback Machine Proceedings. *IEEE Computational Systems Bioinformatics Conference*, CSB. 2016. P. 301-309, DOI: https://doi.org/10.1142/s0219720006001941
8. Saghapour, E.; Kermani, S.; Sehhati, M. A novel feature ranking method for prediction of cancer stages using proteomics data. *PLoS ONE*, 2017, 12 (9). P. 24-29.
9. Hamon Julie. Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale (Thesis) (in French). Lille University of Science and Technology. 2013.
10. Yiming Yang, Jan O. Pedersen. A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning* (ICML' 97) 1997. P. 412-420.
11. Гавриленко С.Ю., Шевердін І.В. Ідентифікація стану комп'ютерної системи на основі ансамблевого методу класифікації. *Системи управління навігації та зв'язку*. Полтава: ПНУ, 2020. Вип. 3 (61). С. 75-79, DOI: https://doi.org/10.26906/SUNZ.2020.3.075
12. Gavrylenko S., Sheverdin I. The ensemble method development of classification of the computer system state based on decisions trees. *Advanced Information Systems*. Vol. 4, No. 2. 2020. P. 5-10. DOI: https://doi.org/10.20998/2522-9052.2020.3.01
13. Tom Carter. An introduction toinformation theory andentropy. *Complex Systems Summer School*. Santa Fe, September 3, 2014. P. 34-39.
14. David J. C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge: Cambridge University Press, 2003. 629 p.
15. Narendra P., Fukunaga K. A Branch and Bound Algorithm for Feature Subset Selection. *IEEE Transactions on Computer*, 26(9): 1977. P. 917-922.
16. Lai, Chun Sing; Tao, Yingshan; Xu, Fangyuan; Ng, Wing W.Y.; Jia, Youwei; Yuan, Haoliang; Huang, Chao; Lai, Loi Lei; Xu, Zhao; Locatelli, Giorgio. A robust correlation analysis framework for imbalanced and dichotomous data with uncertainty. *Information Sciences*. 2019. P. 58-77.
17. Гавриленко С.Ю., Шевердін І.В. Розробка методу оцінки стану комп'ютера на базі аналізу системних подій. *Методи та прилади контролю якості*. Івано-Франківськ, 2018. С. 108-114.

References

1. Andreea, Bendovschi (2015), "Cyber-Attacks – Trends, Patterns and Security Countermeasures", *7th International conference on financial criminology*, 13-14 April 2015,Wadham College, Oxford, United Kingdom, pp. 24-31.
2. Kulbak, S. (1967), *Information Theory and Statistics*, Science, Moscow, 408 p.
3. Kohavi, R. and John, G. (1997), "Wrappers tor feature selection", *Artificial Intelligence*, 91(1-2), pp. 273-324.
4. Isabelle, Guyon and Andre, Elissceff (2003), "An introduction to variable and feature selection", *Journal of Machine Learning Research*, 3'(2003), pp. 1157-1182.
5. Molina, L.C., Belanche, L. and Ncbot, A. (2002), "Feature Selection Algorithms: A Survey And Experimental Evaluation", *Proceedings of the 2002 IEEE International Conference on Data Mining*, IEEE Computer Society, pp. 306-313.
6. Stańczyk, U. (2015), "Feature Evaluation by Filter, Wrapper, and Embedded Approaches", Stańczyk U., Jain L. (eds), *Feature Selection for Data and Pattern Recognition. Studies in Computational Intelligence*, Springer, Berlin, Heidelberg, vol 584, 568 p.
7. Phuong, T.M., Lin, Z. and Altman, R.B. (2016), "Choosing SNPs using feature selection. Archived at the Wayback Machine Proceedings", *IEEE Computational Systems Bioinformatics Conference*, CSB, pp. 301-309, DOI: https://doi.org/10.1142/s0219720006001941
8. Saghapour, E.; Kermani, S. and Sehhati, M. (2017), "A novel feature ranking method for prediction of cancer stages using proteomics data", *Lille University of Science and Technology*, 12 (9), pp. 24-29.
9. Hamon, Julie (2013), "Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale" (Thesis) (in French), *Lille University of Science and Technology*.
10. Yiming, Yang and Jan O., Pedersen (1997), "A comparative study on feature selection in text categorization", *Proceedings of the Fourteenth International Conference on Machine Learning* (ICML' 97), pp. 412-420.
11. Gavrilenko, S.Yu. and Sheverdin, I.V. (2020), "Identification of the state of a computer system based on the ensemble method of classification", *Navigation and communication control systems*, Vol. 3 (61), PNTU, Poltava, pp. 75-79, DOI: https://doi.org/10.26906/SUNZ.2020.3.075
12. Gavrylenko, S. and Sheverdin, I. (2020), "The ensemble method development of classification of the computer system state based on decisions trees", *Advanced Information Systems*, Vol. 4, No. 2, pp. 5-10, DOI: https://doi.org/10.20998/2522-9052.2020.3.01
13. Tom, Carter (2014), "An introduction to information theory andentropy", *Complex Systems Summer School*, Santa Fe, September 3, pp. 34-39.
14. David J. C., MacKay (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge: Cambridge University Press, 629 p.
15. Narendra, P. and Fukunaga, K. (1977), "A Branch and Bound Algorithm for Feature Subset Selection", *IEEE Transactions on Computer*, 26(9), pp. 917-922.

16. Lai, Chun Sing; Tao, Yingshan; Xu, Fangyuan; Ng, Wing W.Y.; Jia, Youwei; Yuan, Haoliang; Huang, Chao; Lai, Loi Lei; Xu, Zhao; Locatelli, Giorgio (2019), "A robust correlation analysis framework for imbalanced and dichotomous data with uncertainty", *Information Sciences*, pp. 58-77.
17. Gavrilenko, S.Y. and Sheverdin, I.V. (2018), "Development of a method for assessing the state of the computer based on the analysis of system events", *Methods and devices of quality control*, Ivano-Frankivsk, pp. 108-114.

Received (Надійшла) 17.02.2021
Accepted for publication (Прийнята до друку) 16.04.2021

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Гавриленко Світлана Юріївна – доктор технічних наук, доцент, професор кафедри обчислювальної техніки та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Svitlana Gavrylenko – Doctor of Technical Sciences, Associate Professor, Professor of Computer Engineering and Programming Department, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: gavrilenko08@gmail.com; ORCID ID: https://orcid.org/0000-0002-6919-0055.

Шевердін Ілля Валентинович – аспірант, кафедра обчислювальної техніки та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Illia Sheverdin – PhD Student of Computer Engineering and Programming Department, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: illia.sheverdin@gmail.com, ORCID ID: https://orcid.org/0000-0002-7881-0658.

Гейко Геннадій Вікторович – кандидат технічних наук, доцент кафедри обчислювальної техніки та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Hennadii Heiko – Candidate of Technical Sciences, Associate Professor of Computer Engineering and Programming Department, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: gennady1752@gmail.com; ORCID ID: https://orcid.org/0000-0001-6958-8306.

Оценка информативности и выбор признаков при идентификации состояния компьютерной системы

С. Ю. Гавриленко, И. В. Шевердин, Г. В. Гейко

Аннотация. Предметом статьи является исследование методов определения информативности признаков. **Целью статьи** является повышение качества классификации состояния компьютерной системы за счет выбора наиболее информативных признаков. **Задача:** исследовать методы выбора оптимальных информационных признаков для идентификации состояния компьютерной системы на основе анализа событий операционной системы Windows. **Используемыми методами** являются: методы машинного обучения, ансамблевые методы, методы выбора оптимальных информационных признаков. Получены **следующие результаты**: выполнен анализ событий операционной системы Windows, исследованы методы выбора оптимальных информационных признаков: методы-обертки (Wrappers), встроенные методы (Embedded) и методы-фильтры (Filters). Выполнена оценка информативности и выбор признаков при идентификации состояния компьютерной системы. Для оценки эффективности выбранных признаков были использованы ансамблевый метод классификации состояния компьютерной системы на основе беггингу и дерева решений J48. Исследована зависимость точности классификации состояния компьютерной системы от выбранных признаков и определен набор атрибутов, которые обеспечивают максимальную точность классификации состояния компьютерной системы. **Выводы.** Научная новизна полученных результатов заключается в анализе событий операционной системы Windows, оценке их информативности и выбора признаков при идентификации состояния компьютерной системы.

Ключевые слова: компьютерная система; события операционной системы; информативность признаков; деревья решений; ансамблевые методы; беггинг.

Informativity assessment and attributes selection in a computer system state identification

Svitlana Gavrylenko, Illia Sheverdin, Hennadii Heiko

Abstract. The **subject** of the article is a study of methods of determining the informativeness of attributes. **The aim of the article** is improvement of the classification quality of a computer system state by selecting the most informative features. **Objective:** To explore methods for selecting optimal information features to identify a computer system state based on an analysis of the Windows operating system events. **The methods used are**: machine learning methods, ensemble methods, methods of selecting the optimal information features. **The following results were obtained**: analysis of the Windows operating system events was performed, methods of selection the optimal information features were investigated: wrapper methods (Wrappers), embedded methods (Embedded) and filter methods (Filters). The informativeness assessment and selection features were performed for identifying a computer system state. An ensemble method for classifying a computer system state based on a bagging and J48 decision tree was developed to evaluate the effectiveness of selected features. The dependency of the classification accuracy of a computer system state on the selected features was investigated, and the attributes set that provides the maximum classification accuracy of a computer system state was determined. **Conclusions.** The scientific novelty of the results is in the analysis of the Windows operating system events, assessment of their informativeness and selection of features in the identification a computer system state.

Keywords: computer system; operating system events, informative features, decision trees; ensemble methods; bagging.