Problems of identification in information systems

УДК 004.732.056 **doi:** https://doi.org/10.20998/2522-9052.2021.4.01

С. Ю. Гавриленко, О. А. Горносталь

Національний технічний університет "Харківський політехнічний інститут", Харків, Україна

РОЗРОБКА МЕТОДУ ІДЕНТИФІКАЦІЇ СТАНУ КОМП'ЮТЕРНИХ СИСТЕМ НА ОСНОВІ БЕГГІНГ-КЛАСИФІКАТОРІВ

Анотація. Предметом дослідження є методи та засоби ідентифікації стану комп'ютерної системи**. Метою статті** є підвищення якості ідентифікації стану комп'ютерної системи за рахунок розробки методу на основі ансамблевих класифікаторів. **Завдання**: дослідити методи побудови беггінг класифікаторів на основі дерев рішень, виконати їх налаштування та розробити метод ідентифікації стану комп'ютерної системи. Використовуваними **методами** є: методи штучного інтелекту, машинного навчання, ансамблеві методи. Отримано такі **результати**: досліджено використання беггінг-класифікаторів на основі мета-алгоритмів: *Pasting Ensemble*, *Bootstrap Ensemble*, *Random Subspace Ensemble*, *Random Patches Ensemble* та *Random Forest* для ідентифікації стану КС, виконано оцінку їх точності. Виконано дослідження параметрів налаштування окремих дерев рішень та знайдено їх оптимальні значення, а саме: максимальну кількість ознак, що використовуються при побудові дерева; мінімальну кількість розгалужень при побудові дерева; мінімальну кількість листків та максимальну глибину дерева. Визначено оптимальну кількість дерев рішень ансамблю. Запропоновано метод ідентифікації стану комп'ютерної системи, який відрізняється від відомих вибором мета-алгоритму класифікації та підбором оптимальних параметрів його налаштування. Проведено оцінку точності розробленого методу ідентифікації стану комп'ютерної системи. Розроблений метод реалізований програмно і досліджений під час розв'язання задачі ідентифікації аномального стану функціонування комп'ютерної системи. **Висновки**. Наукова новизна отриманих результатів полягає у розробці методу ідентифікації стану комп'ютерної системи за рахунок вибору мета-алгоритму класифікації та визначення оптимальних параметрів його налаштування.

Ключові слова: комп'ютерна система; події операційної системи; дерева рішень; ансамблеві методи; мета-алгоритм; беггінг; *Random Forest*.

Вступ

Сьогодні комп'ютерні системи (КС) використовуються практично у всіх галузях народного господарювання. Використання таких складних технічних систем, з однієї сторони, усе більшою мірою стикається з проблемами забезпечення їх функціональної та інформаційної безпеки. Разом із тим з іншої сторони, рівень розвитку засобів ідентифікації та діагностування 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 [1-3].

Об'єктом дослідження є процес ідентифікації стану комп'ютерної системи.

Предметом дослідження є методи ідентифікації стану комп'ютерної системи.

Постановка проблеми та огляд наукових публікацій. Комп'ютерна система характеризується великим обсягом показників її функціонування $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$. Одним із найбільш поширеним методів ідентифікації стану комп'ютерної системи є методи машинного навчання, які призначені безпосередньо працювати з величезними масивами даних, а саме: класичні методи [1, 2], методи навчання з підкріпленням [3], нейромережі і глибоке навчання [4], дерева рішень і ансамблеві методи [5, 6] та ін. Найбільш популярні алгоритми машинного навчання наведено в [7].

Оскільки показники функціонування КС є різнорідними, часто потребують затрат, пов'язаних з їх попередньою обробкою, то найбільш ефективним для вирішення завдання ідентифікації стану КС є використання дерев рішень (ДР) [8-9]. ДР використовують модель білого ящика, здатні працювати як з числовими, так і якісними даними, потребують малої підготовки даних, надають можливість переконатися в правильності моделі за допомогою статистичних тестів, тобто оцінити достовірність моделі. Разом із тим, точність моделей на основі ДР є не достатньою.

Одним з напрямків вдосконалення існуючих моделей є використання ансамблю класифікаторів на основі ДР. Найпопулярнішим видом ансамблів є сімейство беггінг-алгоритмів, що базується на ідеї поєднання незалежних слабких класифікаторів, які навчаються паралельно, використовуючи однаковий алгоритм навчання. В основі беггінгу лежить теорема Кондорсе :

$$\mu = \sum_{i=m}^{N} C_N^i p^i (1-p)^{N-i}$$

де μ – вірогідність прийняття вірного рішення класифікатором, N – кількість дерев рішень, m – мінімальне значення необхідної більшості вірних рішень класифікаторів, p – вірогідність прийняття вірного рішення класифікатора, C_N^i – число поєднань із N по i.

Ефективність беггінгу досягається завдяки тому, що помилки базових алгоритмів, навчених на різних підвибірках, взаємно компенсуються при голосу-

ванні, а також за рахунок того, що об'єкти-викиди можуть не потрапляти до деяких навчальних підвибірок. Беггінг спрямований на зменшення розкиду (*variance*) – різниці між помилкою класифікації тренувальної та тестової вибірок. За якістю одержуваних прогнозів, ансамблі з декількох моделей часто перевершують інші методи [10-11] .

У даній роботі досліджено використання беггінг-класифікаторів на основі мета-алгоритмів: *Pasting Ensemble*, *Bootstrap Ensemble*, *Random Subspace Ensemble* та *Random Patches Ensemble* для ідентифікації стану КС. Кожний із вищенаведених алгоритмів відрізняється алгоритмом формування вихідних даних класифікатору або підвибірок та вибором ознак [12].

Мета-алгоритм *Pasting* (склеювання) відомий як створення випадкових вибірок без заміни. Підвибірки містять усі вихідні ознаки $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$, формуються випадковим чином, є унікальними і не повторюються. При цьому близько 60% вихідних даних *X* використовують у якості вибірок для навчання, решта – тільки для тестування. Основним недоліком цього процесу є те, що кожна підвибірка не може бути повторено і це створює проблему, коли набір даних недостатньо великий.

Мета-алгоритм *Bootstrapping* (початкового завантаження) є найпоширенішим. Підвибірки містять усі вихідні ознаки $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$, формуються випадковим чином і можуть повторюватися.

Відповідно до мета-алгоритму *Random Subspaces* (випадкових підпросторів) підвибірки створюються шляхом випадкового вибору ознак, а не випадковим вибором підвибірок, як це було зроблено раніше, при цьому близько 60% вихідних даних X використовують у якості підвибірок для навчання, решта – тільки для тестування.

Відповідно до мета-алгоритму *Random Patches* (випадкових патчів) підвибірки створюються шляхом випадкового вибору як ознак, так і спостережень (зразків) без повторювання. При створенні вибірок, зазвичай, використовується приблизно 40-60% зразків та атрибутів для навчання.

Відповідно до мета-алгоритму *Random Forest* (випадковий ліс) підвибірки створюються шляхом випадкового вибору як ознак, так і спостережень (зразків) з повторюванням. Випадкові ліси ефективні, коли набір даних має дуже великий розмір, містить викиди, пропущені значення або класи є незбалансованими. Недоліком алгоритму є схильність до перенавчання, особливо за наявності великої кількості шумів у даних.

Жоден із вищенаведених мета-алгоритмів беггінгу не можна апріорі вважати найкращим або досконалим. Підтвердження доцільності використання конкретного мета-алгоритму має бути перевірено і підтверджено експериментом. Крім того, на якість використання кожного із алгоритмів впливає процедура налаштування як самих дерев рішень, так і кількості дерев, які входять до ансамблю.

Постановка завдання. Метою роботи є дослідження та розробка методу ідентифікації стану комп'ютерної системи на основі ансамблевих класифікаторів. Формальна постановка завдання може бути сформульована наступним чином. Нехай функціонування КС характеризується сукупністю її показників $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$, та існують розмічені пари об'єктів $\{(x_i, y_i\}_{i=1}^{N}$, де x_i – показник стану КС, а y_i – мітка стану КС (нормальний або аномальний). Крім того, існує відображення $\boldsymbol{f: X \to Y}$ значення якої відомі лише на об'єктах кінцевої навчальної вибірки $(X, Y) = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Потрібно сформувати мета-алгоритм ***f,*** який здатний класифікувати довільний об'єкт $x \in X$ та налаштувати значення його параметрів *w*: $F(f(w, x), y) \to opt$.

Розробка ансамблевого методу класифікації

В рамках дослідження розглянуто можливість використання беггінг-класифікаторів на основі мета-алгоритмів: *Pasting Ensemble*, *Bootstrap Ensemble*, *Random Subspace Ensemble* та *Random Patches Ensemble* для ідентифікації стану КС. У якості вихідних даних використано показники функціонування КС (завантаження центрального процесора, пам'яті, обсяг трафіку, кількість операцій зчитування/запису на диск, сигнатури вторгнень; статистичні дані аналізу системних подій (кількість операцій роботи із системним реєстром, файловою системою, кількість процесів та ін.).

Результати дослідження використання беггінг-класифікаторів наведено на рис. 1.

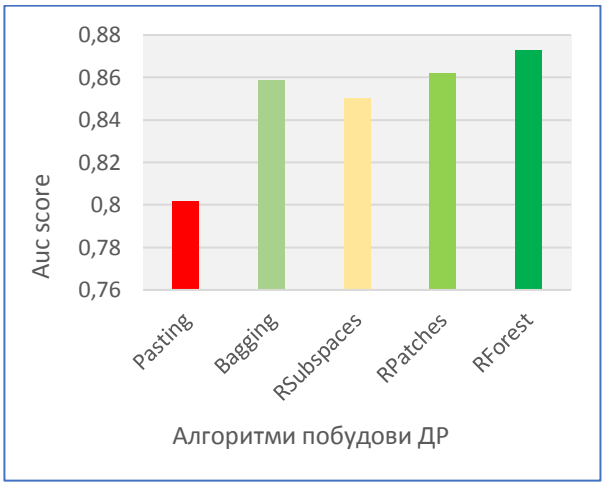

Рис. 1. Порівняння точності класифікації беггінг мета-алгоритмів на навчальній вибірці за допомогою метрики *Auc_score*

Отримано, що за умови стандартного налаштування найбільш якісним є класифікатор на основі алгоритму *Random Forest*, *ROC AUC* якого на навчальній вибірці складає 87%, при цьому точність на тестовій вибірці складає 82%.

Виконано дослідження параметрів налаштування ДР, які входять до ансамблю на основі мета-алгоритму *Random Forest*, а саме:

– максимальна кількість ознак, що використовуються при побудові дерева (*max features*).

– мінімальна кількість розгалужень при побудові дерева (*min sample split*).

– мінімальна кількість листків (*min sample leaf*).

– максимальна глибина дерева (*max depth*).

Результати дослідження налаштування на тестових вибірках наведено на рис. 2-5.

Отримано, що при наступних параметрах налаштування:

'max_depth' від 9; *'max_features'* від 7 до 12;

'min_samples_leaf" від 3 до 15;

'min_samples_split' від 6 до 12

точність класифікатору на основі дерев рішень є найкращою.

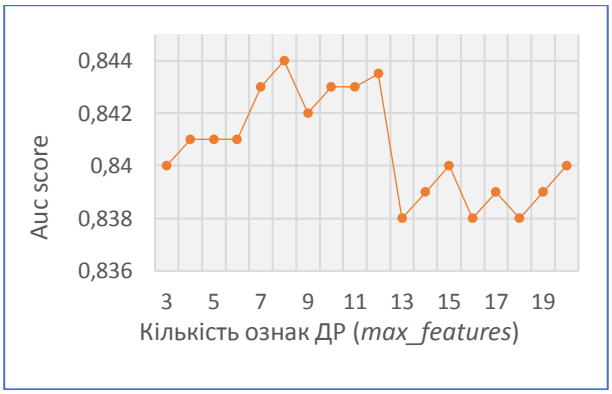

Рис. 2. Залежність точності класифікації ДР від кількості вихідних атрибутів (ознак)

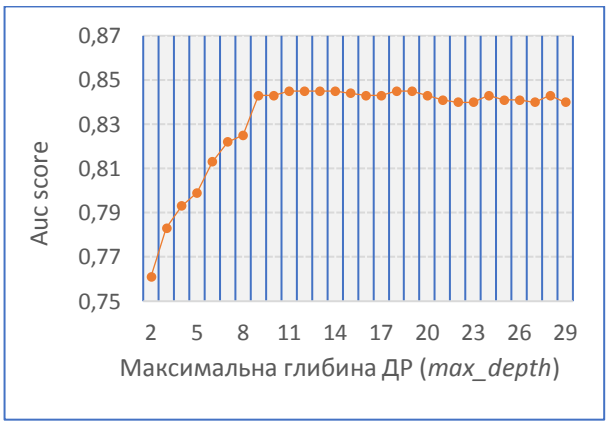

Рис.3. Залежність точності класифікації від максимальної глибини дерева рішень

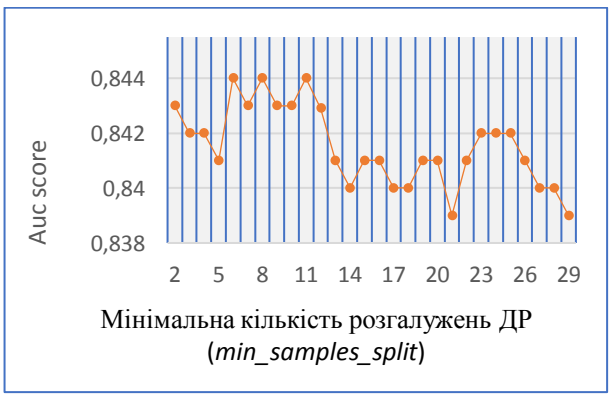

Рис. 4. Залежність точності класифікації ДР від мінімальної кількості розгалужень дерева рішень

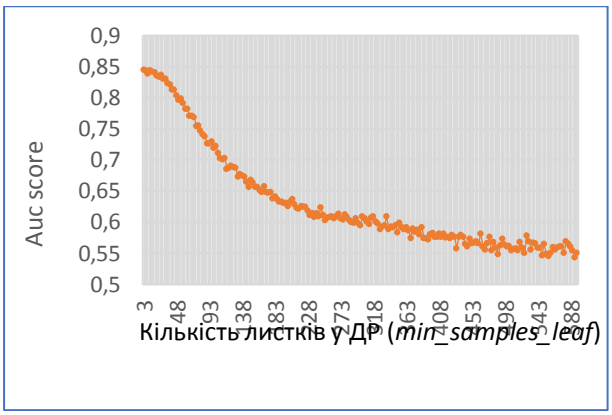

Рис. 5. Залежність точності класифікації ДР від кількості листків дерева рішень

Досліджено залежність точності класифікації від кількості дерев рішень ансамблю (рис. 6) на тестовій та навчальній вибірках.

Отримано, що оптимальна кількість класифікаторів у ансамблі (*max_number*) складає від 33 до 43 ДР.

При цьому налаштування класифікаторів дозволило збільшити точність класифікації на навчальній вибірці до 97%, точність класифікації на тестовій вибірці – до 85%.

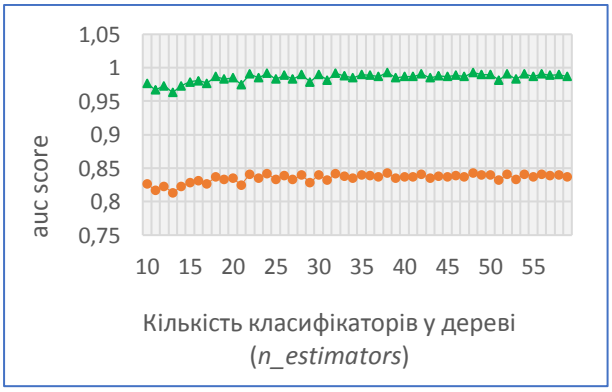

Рис. 6. Залежності точності класифікації від кількості дерев рішень ансамблю (круглі маркери знизу – для тестової вибірки, трикутні маркери вгорі – для тренувальної вибірки)

Висновки

У даній роботі досліджено використання беггінг-класифікаторів на основі мета-алгоритмів:

Pasting Ensemble,
Bootstrap Ensemble,
Random Subspace Ensemble,
Random Patches Ensemble,
Random Forest

для ідентифікації стану комп'ютерні системи.

У якості вихідних даних використано показники функціонування КС.

Точність роботи класифікаторів оцінено за допомогою *ROC*-аналізу.

Отримано, що за умови стандартного налаштування найбільш якісним є класифікатори на основі алгоритму *Random Forest*, *ROC AUC* яких складає 87% на навчальній вибірці.

Виконано налаштування класифікаторів. Отримано, що при наступних параметрах налаштування; '*max_depth*' від 9; '*max_features*' від 7 до 12; '*min_samples_leaf*' від 3 до 15; '*min_samples_split*' від 6 до 12 точність класифікатору на основі дерев рішень є найкращою.

Досліджено залежність точності класифікації від кількості дерев рішень ансамблю.

Отримано, що оптимальна кількість класифікаторів у ансамблі (*max_number*) складає від 33 до 43 дерев рішень. При цьому, налаштування дерев рішень та вибір кількості дерев рішень ансамблю дозволило збільшити точність ідентифікації на основі мета-алгоритму *Random Forest*, *ROC AUC* класифікатору на навчальній вибірці складає 97%, на тестовій вибірці – 85%.

За результатами досліджень запропоновано метод ідентифікації стану комп'ютерні системи, який відрізняється від відомих вибором мета-алгоритму класифікації та підбором оптимальних параметрів його налаштування.

Розроблений метод реалізований програмно і досліджений під час розв'язання задачі ідентифікації стану функціонування комп'ютерні системи.

Проведені експерименти підтвердили працездатність запропонованого методу, що надає можливість рекомендувати його для практичного використання з метою підвищення точності ідентифікації стану комп'ютерної системи.

Перспективи подальших досліджень можуть полягати в проведенні розробки ансамблю нечітких дерев рішень.

References

1. Alpajdin E. Mashinnoe obuchenie: novyj iskusstvennyj intellekt. Moskva: Izdatel'skaya gruppa «Tochka», 2017. 208 p.
2. Marmanis H. Algoritmy intellektual'nogo interneta. Peredovye metodiki sbora, analiza i obrabotki dannyh. Sb-P, M: Simvol, 2011. 468 p.
3. Flah P. Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannyh. Moskva: DMKPress, 2015. 400 p.
4. Tarhov D. A. Nejrosetevye modeli i algoritmy. Moskva: Radiotekhnika, 2014. 352 p.
5. Kaftannikov I. L., Parasich A. V. Osobennosti primeneniya derev'ev reshenij v zadachah klassifikacii. *Vestn. YUUrGU. Ser. «Komp'yuternye tekhnologii, upravlenie, radioelektronika»*. 2015, T. 15, № 3. PP. 26–32.
6. Cha Zhang. Ensemble Machine Learning. Methods and Applications. New York Dordrecht Heidelberg London: Springer, 2012. 329 p.
7. Vipin Kumar. The Top Ten Algorithms in DataMining. Taylor & Francis Group, LLC, 2009. 2006 p.
8. Metody postroeniya derev'ev reshenij v zadachah klassifikacii v Data. URL: https://ami.nstu.ru/~vms/lecture/data_mining/trees.htm (last accessed November 15, 2021).
9. Subbotin S.O. Postroenie derev'ev reshenij dlya sluchaya maloinformativnyh priznakov. *Radio Electronics, Computer Science, Control*. 2019. № 1. PP. 122-130.
10. Marcelo Bacher, Irad Ben-Gal, Erez Shmueli. An information theory subspace analysis approach with application to anomaly detection ensembles.', Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2017. V. 1, PP. 27–39.
11. Vinutha H.P., Basavaraju Poornia. Analysis of Feature Selection and Ensemble Classifier Methods for Intrusion Detection. *International Journal of Natural Computing Research*, 2018. N1. PP. 57-72.
12. Archish Rai Kapil. Bagging. URL: https://www.datavedas.com/bagging/ (last accessed November 15, 2021).

Список літератури

1. Алпайдин Э. Машинное обучение: новый искусственный интеллект. Москва: Издательская группа «Точка», 2017. 208 с.
2. Марманис Х. Алгоритмы интеллектуального интернета. Передовые методики сбора, анализа и обработки данных. Сб-П, М: Символ, 2011. 468 с.
3. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. Москва: ДМКПресс, 2015. 400 с.
4. Тархов Д. А. Нейросетевые модели и алгоритмы. Москва: Радиотехника, 2014. 352 с.
5. Кафтанников И. Л., Парасич А. В. Особенности применения деревьев решений в задачах классификации. *Вестн. ЮУрГУ. Сер. «Компьютерные технологии, управление, радиоэлектроника»*. 2015, Т. 15, № 3. С. 26–32.
6. Cha Zhang. Ensemble Machine Learning. Methods and Applications. New York Dordrecht Heidelberg London: Springer, 2012. 329 p.
7. Vipin Kumar. The Top Ten Algorithms in DataMining. Taylor & Francis Group, LLC, 2009. 2006 p.
8. Методы построения деревьев решений в задачах классификации. URL: https://ami.nstu.ru/~vms/lecture/data_mining/trees.htm (дата звернення: 15.11.2021).
9. Субботін С.О. Построение деревьев решений для случая малоинформативных признаков. *Radio Electronics, Computer Science, Control*. 2019. № 1. С. 122-130.
10. Marcelo Bacher, Irad Ben-Gal, Erez Shmueli. An information theory subspace analysis approach with application to anomaly detection ensembles, *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 2017. V. 1, PP. 27–39.
11. Vinutha H.P., Basavaraju Poornia. Analysis of Feature Selection and Ensemble Classifier Methods for Intrusion Detection. *International Journal of Natural Computing Research*, 2018. *N*1. *PP*. 57-72.
12. Archish Rai Kapil. Bagging. URL: https://www.datavedas.com/bagging/ (дата звернення: 15.11.2021).

Received (Надійшла) 23.09.2021
Accepted for publication (Прийнята до друку) 03.11.2021

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

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

Горносталь Олексій Андрійович – аспірант кафедри обчислювальної техніки та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Oleksii Hornostal – PhD Student of Department of "Computer Engineering and Programming", National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: gornostalaa@gmail.com; ORCID ID: https:// orcid.org/0000-0001-5820-9999.

Разработка метода идентификации состояния компьютерных систем на основе беггинг-классификаторов

С. Ю. Гавриленко, А. А. Горносталь

Аннотация. Предметом исследования есть методы и средства идентификации состояния компьютерной системы. **Цель статьи** – повышение качества идентификации состояния компьютерной системы за счет разработки метода на основе ансамблевых классификаторов. Задание: исследовать методы построения беггинг классификаторов на основе деревьев решений, выполнить их настройки и разработать метод идентификации состояния КС. **Используемые методы**: методы искусственного интеллекта, машинного обучения, ансамблевые методы. **Получены следующие результаты:** исследовано использование беггинг классификаторов на основе мета-алгоритмов: *Pasting Ensemble, Bootstrap Ensemble, Random Subspace Ensemble, Random Patches Ensemble* и *Random Forest* для идентификации состояния компьютерной системы, выполнена оценка их точности. Выполнены исследования параметров настройки отдельных деревьев решений и найдены их оптимальные значения, а именно: максимальное количество признаков, используемых при построении дерева; минимальное количество ветвлений при построении дерева; минимальное количество листьев и максимальную глубину дерева. Определено оптимальное количество деревьев решений ансамбля. Предложен метод идентификации состояния компьютерной системы, отличающийся от известных выбором мета-алгоритма классификации и подбором оптимальных параметров его настройки. Проведена оценка точности разработанного метода идентификации состояния компьютерной системы. Разработанный метод реализован программно и исследован при решении задачи идентификации аномального состояния функционирования компьютерной системы. **Выводы**. Научная новизна полученных результатов заключается в разработке метода идентификации состояния компьютерной системы за счет выбора мета-алгорита классификации и определения оптимальных параметров его настройки.

Ключевые слова: компьютерная система; события операционной системы; деревья решений; ансамблевые методы; мета-алгоритм; беггинг; *Random Forest*.

Development of a method for identification of the state of computer systems based on bagging classifiers

Svitlana Gavrylenko, Oleksii Hornostal

Abstract. The subject of the research is methods and means of identifying the state of a computer system . The purpose of the article is to improve the quality of computer system state identification by developing a method based on ensemble classifiers. **Task**: to investigate methods for constructing bagging classifiers based on decision trees, to configure them and develop a method for identifying the state of the computer system. **Methods used**: artificial intelligence methods, machine learning, ensemble methods. **The following results were obtained**: the use of bagging classifiers based on meta-algorithms were investigated: Pasting Ensemble, Bootstrap Ensemble, Random Subspace Ensemble, Random Patches Ensemble and Random Forest methods and their accuracy were assessed to identify the state of the computer system. The research of tuning parameters of individual decision trees was carried out and their optimal values were found, including: the maximum number of features used in the construction of the tree; the minimum number of branches when building a tree; minimum number of leaves and maximum tree depth. The optimal number of trees in the ensemble has been determined. A method for identifying the state of the computer system is proposed, which differs from the known ones by the choice of the classification meta-algorithm and the selection of the optimal parameters for its adjustment. An assessment of the accuracy of the developed method for identifying the state of a computer system is carried out. The developed method is implemented in software and investigated when solving the problem of identifying the abnormal state of the computer system functioning. **Conclusions**. The scientific novelty of the results obtained lies in the development of a method for identifying the state of the computer system by choosing a meta-algorithm for classification and determining the optimal parameters for its configuration.

Keywords: computer system; operating system events; decision trees; ensemble methods; meta-algorithm; bagging; Random Forest.