УДК 004.056.53

Світлана УНІЧЕНКО[1], Галина ЛИСИК[2], Мар'яна ЗАКРЕВСЬКА[3]

[1] Західноукраїнський національний університет

[2] Бродівська гімназія імені Івана Труша Львівської області

[3] Ситненський ліцей Крупецької сільської ради Рівненської області

МЕТОД ВИПАДКОВОГО ЛІСУ ДЛЯ ІДЕНТИФІКАЦІЇ КОРИСТУВАЧА НА ОСНОВІ СТИЛІСТИЧНИХ ХАРАКТЕРИСТИК ЕЛЕКТРОННИХ ТЕКСТІВ

Вступ. Поширення Інтернету та засобів масової комунікації створює ситуацію, коли кількість електронної текстової інформації зростає, а разом із цим збільшується потреба в методах ідентифікації користувачів [1].

Існуючі методи ідентифікації, такі як ідентифікація за технічними характеристиками пристроїв або поведінковими ознаками користувачів на веб-ресурсах, не завжди дозволяють визначити особистість конкретної людини. Це обмежує їхню ефективність, особливо у випадках, коли користувачі залишають короткі повідомлення або діють під вигаданими іменами.

Одним із перспективних напрямів розвитку технологій є використання біометричної ідентифікації, зокрема методів, що базуються на аналізі лінгвістичних та стилістичних характеристик тексту. Цей підхід дозволяє ідентифікувати користувача за унікальними рисами його письма, що є своєрідним відбитком автора.

Мета: розробити метод випадкового лісу для ідентифікації користувача на основі стилістичних характеристик електронних текстів.

1. Метод випадкового лісу для ідентифікації користувача на основі стилістичних характеристик електронних текстів

Алгоритм випадкового лісу раніше не використовувався для розв'язання задачі ідентифікації англомовних Інтернет-користувачів, хоча під час вирішення схожих задач в інших предметних областях було доведено, що алгоритм має ряд переваг.

Зокрема, простота алгоритму для розуміння, наочність результатів побудови, відсутність необхідності вибору вхідних ознак (під час побудови дерева вибираються найбільш значущі ознаки, які й використовуються для побудови), швидкість навчання та висока точність класифікації, ефективна робота з великими обсягами даних, а також здатність обробляти велику кількість ознак, можливість роботи з усіма типами ознак (дискретні, неперервні, бінарні, символічні тощо) та можливість роботи з даними з пропущеними значеннями.

Основна ідея алгоритму RF полягає в побудові ансамблю (лісу) випадкових дерев ухвалення рішень. Структура дерева представляє собою деревоподібний граф, що має в своєму складі ребра (гілки) та вузли двох типів: листя та внутрішні вузли (рисунок 1).

У листях дерева містяться значення цільової функції, у нашому випадку клас - користувач, на ребрах записані значення ознак, від яких залежить значення цільової функції, у вузлах - самі ознаки. Щоб класифікувати повідомлення,

необхідно спуститися від кореня дерева до листа та отримати значення класу, що в ньому міститься.

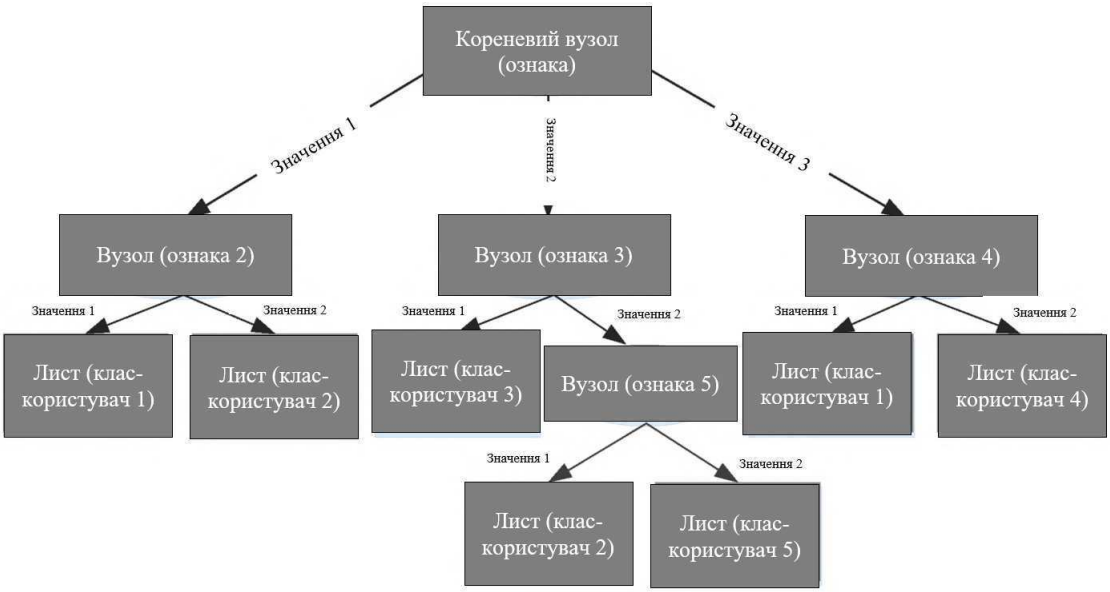

Рисунок 1 - Дерево рішень із довільною кількістю нащадків

Нехай дано навчальну вибірку $T_{tr} \in T$, що складається з q повідомлень:

$$T_{tr} = \begin{Bmatrix} t_{us}1 \\ \dots \\ t_{usq} \end{Bmatrix} = \begin{bmatrix} f_11 & \dots & f_1n \\ \dots & \dots & \dots \\ f_q1 & \dots & f_qn \end{bmatrix},$$

розмірність ознакового простору - n; n' - параметр, який визначає кількість випадково відібраних ознак із числа n (використовується $n' = \sqrt{n}$); $U_{потенц}$ - множина потенційних користувачів (g - кількість потенційних користувачів). Для кожного повідомлення $t_{us} \in T_{tr}$ відома його належність до певного користувача із множини $U_{потенц} = \{U_1, \dots, U_g\}$.

Дерева ансамблю будуються за наступним принципом:

1) генеруються випадкові підвибірки з повторенням розміром q' з навчальної вибірки - $T_{tr} = \{T'_1, T'_2, \dots, T'_h\}$, де h - кількість згенерованих підвибірок. Деякі тексти потрапляють у підвибірку кілька разів, а деякі не потрапляють у неї взагалі. Також випадковим чином відбираються n' ознак (рисунок 2);

2) на основі кожної такої підвибірки будується дерево рішень:

а) на першому кроці роботи алгоритму є корінь, множина повідомлень T', яку необхідно розбити на підмножини, та множина ознак $F' = \{f_1, \dots, f_{m'}\}$ довжиною n'. Під час побудови обирається одна з ознак як критерій, що забезпечує найкраще розділення за n'. Для найкращого вибору використовується підхід на основі мінімізації ентропії $H(S) = -(\sum_{u \in U} P(u) \times \log_2 P(u))$, на відміну від виразу $IG(S, f_i) = H(S) - \sum_{x \in values(f_i)} \frac{|S_x|}{|S|} \times H(S_x)$, що використовується в класичному алгоритмі. В цих формулах $H(S)$ - ентропія до розділення за ознакою, $values(f_i)$ - всі можливі значення ознаки f_i, S_x - підмножина набору даних, де $f_i = x$, $|S|$ - кількість елементів у множині.

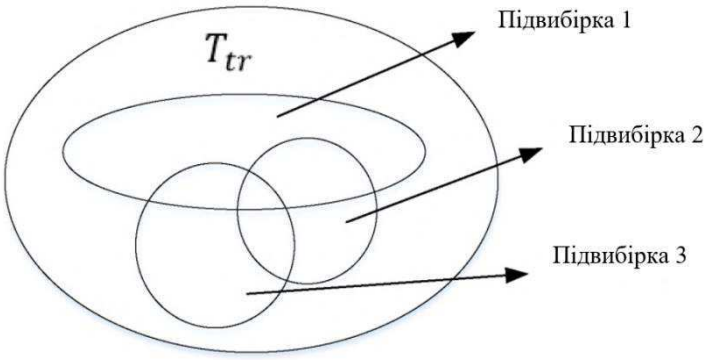

Рисунок 2 - Генерація підвибірок для побудови дерев рішень

б) обрана ознака f_i має k значень, що дає розбиття на k підмножин. Далі створюються k нащадків вузла, кожному з яких відповідає підмножина, отримана при розбитті початкової множини T'. Вибір ознаки та розбиття за нею на підмножини рекурсивно застосовуються до всіх k нащадків. Умовою виходу з рекурсії є випадки, коли, після розгалуження в вузлі опиняються тексти одного користувача (тоді вузол стає листом), або вузол виявився асоційованим з порожньою множиною (тоді вузол також стає листом, але в якості користувача обирається користувач, який найчастіше зустрічається у безпосередніх предків цього вузла);

в) дерева будуються до вичерпання вибірки, гілки не відсікаються;

3) класифікація здійснюється голосуванням: кожне дерево лісу класифікує текст до одного з користувачів, тобто голосує за певного користувача. Далі текст відноситься до того користувача, за якого проголосувало найбільше дерев (рисунок 3).

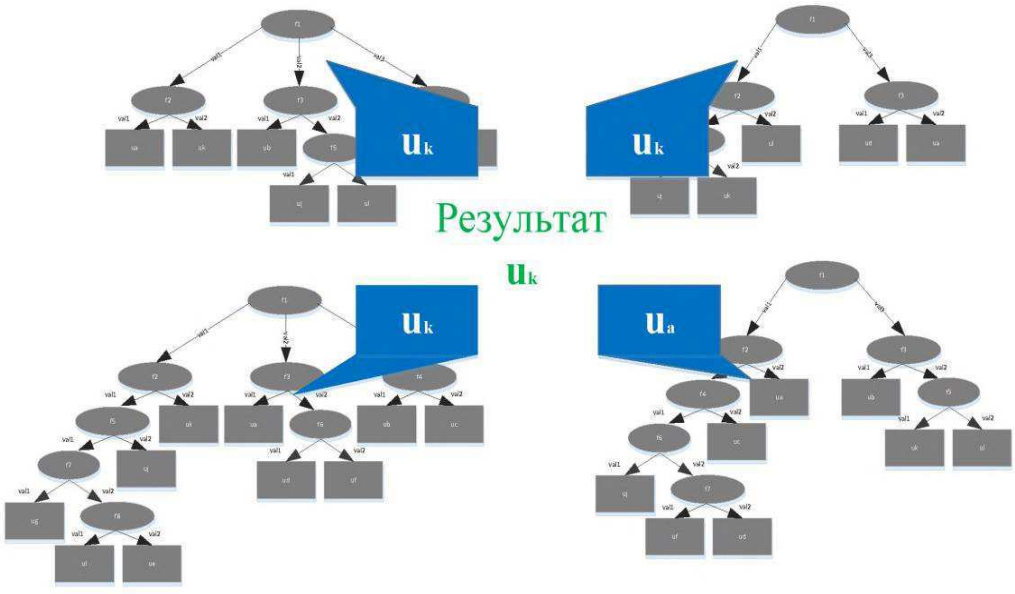

Рисунок 3 - Класифікація повідомлення за принципом голосування ансамблю дерев рішень

Передбачається, що більшість згенерованих дерев самі по собі правильно передбачають користувача, і що дерева, які помиляються, видають різні результуючі класи.

Висновок. Розроблено метод випадкового лісу для ідентифікації користувача на основі стилістичних характеристик електронних текстів.

Перелік використаних джерел.

1. Rosenblum N., Zhu X., Miller B.P. Who Wrote This Code? Identifying the Authors of Program Binaries. Computer Security. 2011. Vol. 6879. P. 172-189.