УДК 004.021

Данильчук Руслан Костянтинович
студент
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»
Данильчук Руслан Константинович
студент
Национального технического университета Украины
«Киевский политехнический институт имени Игоря Сикорского»
Danylchuk Ruslan
Student of
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

Жураковська Оксана Сергіївна
доцент кафедри автоматизованих систем обробки
інформації та управління
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Жураковская Оксана Сергеевна
доцент кафедры автоматизированных систем обработки
информации и управления
Национальный технический университет Украины
«Киевский политехнический институт имени Игоря Сикорского»
Zhurakovska Oksana
Associate Professor of the Department ASOIU
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

ЗАДАЧА КЛАСТЕРИЗАЦІЇ АДРЕС В МЕРЕЖІ БЛОКЧЕЙН

ЗАДАЧА КЛАСТЕРИЗАЦИИ АДРЕСОВ В СЕТИ БЛОКЧЕЙН

BLOCKCHAIN TRANSACTIONS ANALYSIS SYSTEM

Анотація. *У даній статті розглянуто практичне застосування методу кластеризації адрес в мережі блокчейн на прикладі задачі визначення кількох адрес одного користувача. Результати дослідження показують доцільність пропонованого підходу до кластеризації адрес Bitcoin. Користувачам може бути корисно уникнути небезпечних моделей використання Bitcoin, а дослідникам провести більш розширений аналіз анонімності.*

Ключові слова: *технологія blockchain, блокчейн, збереження даних, блок, майнер, учасники, записи, ключ, складність мережі, складність хешування.*

Аннотация. *В данной статье рассмотрено практическое применение метода кластеризации адресов в сети блокчейн на примере задачи определения нескольких адресов одного пользователя. Результаты исследования показывают целесообразность предлагаемого подхода к кластеризации адресов Bitcoin. Пользователям может быть полезно избежать опасных моделей использования Bitcoin, а исследователям провести более расширенный анализ анонимности.*

Ключевые слова: *технология blockchain, блокчейн, хранения данных, блок, майнер, участники, записи, ключ, сложность сети, сложность хэширования.*

Summary. *In this article, the practical application of the method of clustering of addresses in the blockade network is considered on the example of the task of determining the multiple addresses of one user. The results of the study show the*

appropriateness of the proposed approach to clustering Bitcoin addresses. It may be useful for users to avoid dangerous patterns of Bitcoin use, and for researchers to conduct a more advanced analysis of anonymity.

Key words: *blockchain technology, blocking, storage of data, block, miner, participants, records, key, compatibility of the network, compatibility of hashing.*

Вступ. Blockchain (з англ. block — блок, chain — ланцюг) — це ланцюжок блоків транзакцій, які зберігаються на комп'ютерах учасників ланцюжка. Кожен наступний блок пов'язаний з попереднім і складається з набору записів. Нові блоки завжди додаються лише в кінець цього ланцюжка [1].

Ланцюжок даних має три основні принципи:
- захищеність;
- розподіленість;
- відкритість.

Всі учасники блокчейну об'єднуються в комп'ютерну мережу. На кожному сервері зберігається копія всіх даних блоку. Це і є основою надійності blockchain.

Адже, щоб зламати ланцюжок, потрібно отримати доступ до бази даних всіх комп'ютерів мережі.

Всі дані, що з'являються в блоках відкриті (користувачі бачать їх) і зашифровані (користувачі не знають, кому вони належать).

Приклад запису в мережі блокчейн: «Користувач з ключем К отримав у кредит телефон з ключем S».

Кожен користувач може мати декілька різних ключів. Тобто, навіть знаючи ключ власника телефону, не можна дізнатися про наявність у нього штрафу за порушення правил дорожнього руху.

Підходи до вирішення задачі кластеризації

На рисунку 1 представлена класифікація алгоритмів та методів кластерного аналізу.

Сутність ієрархічних агломеративних методів полягає у тому, що на першому кроці кожний об'єкт вибірки розглядається як окремий кластер. Процес об'єднання кластерів відбувається послідовно, на підставі матриці відстаней або матриці подібності поєднуються найбільш близькі об'єкти. Послідовність об'єднання легко піддається геометричній інтерпретації й може бути представлена у вигляді графа-дерева. Основною передумовою ієрархічних дивізивних методів є те, що спочатку всі об'єкти належать до одного кластеру. У процесі класифікації за певними правилами поступово від цього кластера відокремлюються групи схожих між собою об'єктів [2]. Так, на кожному кроці кількість кластерів зростає, а міра відстані між кластерами зменшується. Складнощі ієрархічних методів кластеризації наступні:
- обмеження обсягу набору даних;
- вибір міри близькості;
- негнучкість отриманих класифікацій.

Перевага цієї групи методів порівняно з неієрархічними методами полягає у їх наочності і можливості отримання детального уявлення про структуру даних. При використанні ієрархічних методів існує можливість досить легко ідентифікувати викиди в наборі даних і в результаті підвищити якість даних. Велика кількість методів ієрархічного кластерного аналізу різниться не тільки використаними мірами подібності (розходження), але й алгоритмами класифікації.

Неієрархічні методи виявляють більш високу стійкість по відношенню до викидів, невірного вибору метрики, включення незначущих змінних в базу для кластеризації та інше. Необхідно заздалегідь фіксувати результуючу кількість кластерів, правило зупинки і, якщо на те є підстави, початковий центр кластеру, що суттєво впливає на ефективність роботи алгоритму. Якщо немає підстав штучно задавати ці умови, рекомендується використовувати ієрархічні методи.

Алгоритм кластеризації адресів

Розглянемо мережі на основі блоків, які допомагають об'єднати групи адрес блокчейн в одну суцільну систему. Ці показники засновані на певних моделях, які є загальними для багатьох транзакцій в мережі. Однак вони не завжди задовольняються для всіх транзакцій, а отже схильні до помилок. Це означає, що деякі адреси можуть бути помилково пов'язані між собою.

Для аналізу транзакцій, окрема транзакція розглядається як упорядкована $t=(A,B,c)$ та складається з:
- кінцевого багатоступеневого транзакційного входу A, де кожен вхід $(a_i, A_i) \in A$ — упорядкована пара адреси A_i і значення вхідного $a_i > 0$.
- кінцевого багатоступеневого транзакційного виходу B, де кожен вихід $(b_j, B_j) \in B$ — це упорядкована пара адреси B_j і значення вихідного $b_j \geq 0$.
- плати за транзакцію $c = \sum_{(a_i,)\in A} a_i - \sum_{(b_j,)\in B} b_j \geq 0$.

Для довільної множини транзакційних входів або виходів A позначаємо мультимережний адрес в A, як Addr(A).

Найбільш очевидною ідеєю для кластеризації адрес блокчейну є з'єднання всіх вхідних адрес однієї транзакції. Якщо дві або більше адрес є входами однієї транзакції з одним виходом, то всі ці адреси керуються тим самим користувачем [3].

Розглянемо транзакцію $t = (A, B, c)$, що задовольняє умови одноразової зміни.
- #Addr (B) = 2, тобто транзакція t має рівно два виходи.
- #Addr(A) = 6 ≠ 2, тобто кількість входів t не є рівною двом. Якщо #Addr(A) = #Addr(B) = 2,

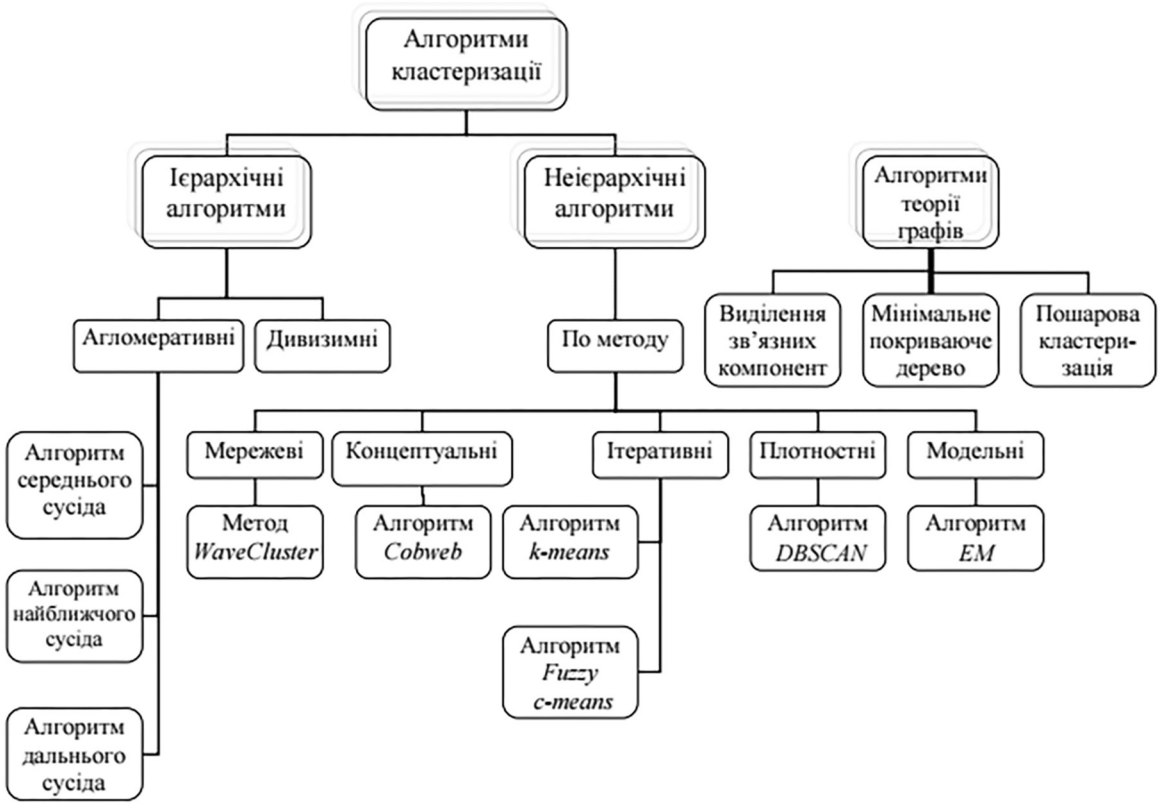

Рис. 1. Класифікація алгоритмів та методів кластерного аналізу

транзакція, швидше за все, поділиться міткою передачі.

- Обидва виходи транзакції t, B_1 та B_2 не є обмінними адресами, тобто $B_1, B_2 \notin \text{Addr}(A)$.
- Один вихід транзакції B_1 не існував до транзакції t, а десяткове подання значення b_1 має більше ніж 4 цифри після крапки.
- Інший вихід транзакції B_2 раніше був частиною мережі, і в попередніх транзакціях він не був адресований поза обліковим записом.

Розглянемо алгоритм кластеризації адрес на прикладі мережі Bitcoin, який регулює баланс інформації, що надходить безпосередньо з блоків Bitcoin (CS та OTC) та додаткову інформацію, зібрану з Інтернету у вигляді тегів.

Нехай $T = \{t_j\}$ і є набором всіх транзакцій в блоці біткойн, тоді як A є набором всіх адрес, що присутні в транзакції з T.

Кластеризація адрес Bitcoin — це розбивка $A = A_1 \cup A_2 \cup \ldots \cup A_N$ на непересічні підмножини $A_l \cap A_l = \varnothing$ для $l \neq j$. За допомогою $T_H \subset T$ позначимо сукупність всіх транзакцій, які задовольняють CS, або OTC. Для транзакції $t \in T_H$, через $\text{Addr}_H(t)$ позначимо множину всіх адрес, які слід віднести до одного користувача відповідно.

Інформація про теги представлена як сукупність негативних пар $L = \{(a_i, a_j,)\}$. Пара адрес $(a_i, a_j) \in L$, якщо у нас є частина інформації про те, що ці адреси не контролюються одним і тим самим користувачем.

Слід зазначити, що як CS і OTC, так і позабіржова евристика і набір негативних пар L можуть містити помилкову інформацію.

Розглянемо різні типи спостережень:

- у випадку, якщо всі адреси $\text{Addr}_H(t)$ для деяких $t \in T_H$ дійсно належать одному і тому ж користувачу з ймовірністю p;
- у випадку, якщо дві адреси $(a_i, a_j) \in L$ контролюються тим самим користувачем з ймовірністю q.

В інших випадках інформація про негативне об'єднання між будь-якою парою адрес в L перевіряється шляхом $1 - q$.

Нехай ймовірність $P(A, T_H, L \mid p, q)$ буде функцією від кластеризації A, транзакції T_H та негативних пар L:

$$P(A, T_H, L|p,q) = \prod_{t \in T_H} p^{\mathbb{I}(\text{Addr}_H(t) \subset Cl(A))} \times (1-p)^{\mathbb{I}(\text{Addr}_H(t) \not\subset Cl(A))} \times \times \prod_{\{a,a,\} \in L} (1-q)^{\mathbb{I}(\{a,a,\} \not\subset Cl(A))} \times q^{\mathbb{I}(\{a,a,\} \subset Cl(A))},$$

де для деякого набору Bitcoin адреси S позначення $S \subset Cl(A)$ означає, що існує кластер A_l, такий, що $S \subseteq A_l$.

Отже, log-правдоподібність співвідноситься як

$$\ln P(A, T_H, L|p,q) = \sum_{T \in T_H} \mathbb{I}(\text{Addr}_H(t) \subset Cl(A)) \ln(1-p) +$$

$$+\sum_{T\in T_H}\mathbb{I}\left(\mathrm{Addr}_H(\mathrm{t})\not\subset Cl(A)\right)\ln(p)+$$
$$+\sum_{\{a,\,a,\}\in \mathrm{L}}\mathbb{I}\left(\{a,\,a'\}\not\subset Cl(A)\right)\ln(1-p)+$$
$$+\sum_{\{a,\,a,\}\in \mathrm{L}}\mathbb{I}\left(\{a,\,a'\}\subset Cl(A)\right)\ln(p)$$

Слід зазначити, що запропонована модель не призначена для використання імовірнісної структури реального світу, а лише дає більш розгорнутий підхід до систематичного вивчення довіри між різними джерелами інформації. Більше того, це дозволяє ефективно оптимізувати параметри.

Максимізація log-правдоподібності — це задача дискретної оптимізації, яка фактично NP-повна.

Розглянемо ретроспективно всі транзакції в мережі Bitcoin, які задовольняють одну евристику. На кожному етапі вирішується, чи приєднуються кластери, що відповідають адресі $\mathrm{Addr}_H\ (t_J)$ до розглянутої транзакції t_J.

Нехай $A_J = A_{k1}\cup\ldots\cup A_{km}$ — об'єднання всіх кластерів, представники яких належать $\mathrm{Addr}_H\ (t_J)$.

Знайдемо зміни кількості негативних пар, що відповідають $\mathrm{Addr}_H\ (t_J)$ в один кластер A_J:

$$\Delta_{t_j}\left(\sum_{\{a,\,a,\}\in \mathrm{L}}\mathbb{I}\left(\{a,\,a'\}\not\subset Cl(A)\right)\right)=$$
$$=\sum_{\{a,\,a'\}\in\widehat{A}_J}\mathbb{I}\left(\{a,\,a,\}\in A_l\right)-\sum_{i=1}^{m_j}\sum_{\{a,\,a'\}\in A_{k_i}}\mathbb{I}\left(\{a_i,\,a_j,\}\in A_l\right)=$$
$$=\Delta_{\widehat{A}_j}-\sum_{i=1}^{m_j}\Delta_{A_j}$$

де ΔA_m — це кількість негативних пар в кластері A_m.

Тепер об'єднаємо всі кластери, що відповідають $\mathrm{Addr}_H\ (t_J)$, а отже зміна log-правдоподібності дорівнює

$$\Delta_P\left(t_j,A,L|p,q\right)=ln\left(\frac{p}{1-p}\right)+\left(\Delta_{\widehat{A}_j}-\sum_{i=1}^{m_j}\Delta_{A_j}\right)\ln\left(\frac{q}{1-q}\right).$$

Таким чином, якщо $\Delta_P\left(t_j,A,L|p,q\right)$ є позитивним, то ми зливаємо всі кластери, що відповідають $\mathrm{Addr}_H\ (t_J)$, в іншому випадку потрібно продовжувати наступну транзакцію.

Слід відзначити, що завдяки такому підходу зміна параметрів p і q може призвести до дуже немонотонної зміни кластеризації. Наприклад, можна зменшити параметр q, який повинен вести до менших кластерів, але з'ясується, що найбільший кластер стає ще більшим.

Висновки. У цій роботі було проаналізовано існуючі методи для розв'язку задачі кластеризації та запропоновано використати алгоритм групування адрес блокчейн для визначення множини адресів одного користувача. В роботі наведений даний алгоритм. Та проаналізовано його особливості: використання для кластеризації не тільки інформацію про блокчейни, а й інформацію з Інтернету поза мережі блокчейн, та розгляд деяких типів даних поза мережею як голоси проти адресного об'єднання в процесі кластеризації. Такий підхід дозволяє уникнути значної частини помилкових об'єднань кластерів.

Література

1. Nakamoto S. (2008) Bitcoin: A peer-to-peer electronic cash system.

2. Інформаційні технології УДК 004,825 к.т.н. Волосюк Ю. В. (ЄУ, м. Миколаїв) Аналіз алгоритмів кластеризації для задач інтелектуального аналізу даних.

3. Ron D. and Shamir A. (2012) Quantitative analysis of the full bitcoin transaction graph. Cryptology ePrint Archive, Report 2012/584.