

Зосіменко В.В.

КЛАСТЕРИЗАЦІЯ, ЯК МЕТОД ОБ'ЄКТИВІЗАЦІЇ ЕЛЕМЕНТІВ КРИМІНАЛІСТИЧНОЇ ХАРАКТЕРИСТИКИ ЗЛОЧИНІВ

У даний час існує загальноприйняте визначення терміну криміналістична характеристика злочину (КХЗ) – це система даних про криміналістичні значимі ознаки певного виду або групи злочинів, які мають закономірні зв'язки між ними, і які мають мету створення та перевірки криміналістичних версій [1].

Значення КХЗ перед усім полягає в тому, що на початковому етапі розслідування злочинів, коли існує певний дефіцит інформації, за рахунок стійких кореляційних зв'язків між різними елементами КХЗ є можливість розгляду нових версій щодо невстановлених обставин, при цьому саме трирівнева структура КХЗ дозволяє встановити взаємозв'язок між структурними елементами КХЗ [2]. В попередніх роботах ми наводили обрану нами структуру як елементів КХЗ, так і ознак елементів, зазначали про наявність та силу статистичного зв'язку між певними показниками [3, 4].

Метою даного дослідження було підвищення об'єктивності при визначенні «таксономічної формули» злочинця шляхом оптимізації і систематизації статистичних зв'язків між різними показниками КХЗ.

При вирішенні завдання знаходження характеристик передбачуваного злочинця, на першому етапі нам необхідно виділити загальні класи злочинів і зіставити їх з класами злочинців. Оскільки класифікація відносить кожен об'єкт до однієї з заздалегідь визначених груп, для побудови оптимальної розбивки об'єктів на групи, зробимо кластеризацію (кластер англ. *cluster* – скупчення) – об'єднання декількох однорідних елементів, яке може розглядатися як самостійна одиниця, що володіє певними властивостями) наявної бази.

Необхідно виділяти класифікацію та кластеризацію. Класифікацією називається віднесення кожного елемента в певний клас із заздалегідь відомими параметрами, отриманими на етапі навчання. При цьому кількість класів чітко обмежена. Кластеризація – це розбиття безлічі даних на кластери, параметри яких заздалегідь невідомі. Кількість кластерів може бути довільною або фіксованою.

Кластеризацію можна визначити як процес об'єднання даних у групи за схожими ознаками. Оскільки кластеризація розбиває безліч об'єктів на групи, які визначаються тільки її результатом, оптимальність може бути визначена як вимога мінімізації середньоквадратичної помилки розбиття:

$$e^2(X, L) = \sum_{j=1}^K \sum_{i=1}^{n_j} \|x_i^{(j)} - c_j\|^2,$$

де C_j – «центр мас» кластера j . «Центр мас» кластера – точка в просторі характеристичних векторів із середніми для даного кластера значеннями характеристик. У нашому випадку будемо використовувати кластеризацію для спрощення подальшої обробки даних і розуміння структури множини об'єктів X^i за допомогою дивізації на групи схожих об'єктів. При цьому ми можемо істотно скоротити обсяг збережених даних у разі великої вибірки, залишивши по одному найбільш типовому представнику від кожного

кластера. Також можливе виділення нетипових об'єктів, які не підходять до жодного з кластерів.

У першому випадку кількість кластерів будемо мінімізувати. У другому випадку важливіше забезпечити високу ступінь подібності об'єктів усередині кожного кластера, а кластерів може бути скільки завгодно. У третьому випадку найбільший інтерес представляють окремі об'єкти, які не вписуються ні в один з кластерів. Для цього наведемо визначення тих понять, з якими будемо оперувати. *Об'єкт* – елементарна група даних, з якою оперують алгоритми кластеризації. Кожному об'єкту ототожнюється *вектор характеристик*: $x = (x_1 \dots, x_d)$. Компоненти x_i є окремими характеристиками об'єкта. Кількість характеристик d визначає *розмірність* простору характеристик. Множинність, що складається з усіх векторів характеристик, будемо позначати X . $X = (x_1, \dots, x_n)$, де $x_i = (x_{i1}, \dots, x_{id})$. *Кластер* – підмножина «близьких один до одного» об'єктів із X . Відстань $R(x_i, x_j)$ між об'єктами x_i та x_j – результат застосування обраної метрики в просторі характеристик.

У загальному випадку кластеризація даних включає в себе такі етапи:

1. Виділення характеристик.
2. Визначення метрики.
3. Розбивка об'єктів на групи.
4. Наведення результатів.



Рис.1. Алгоритм послідовної кластеризації.

Застосуємо схему виділення кластерів у нашому випадку. Для виділення вектора характеристик спочатку необхідно вибрати властивості, які характеризують ці об'єкти. Ними можуть бути кількісні характеристики (вік, кількість ран), якісні характеристики (день тижня, стать, локалізація ран) тощо. Потім необхідно зменшити розмірність простору характеристичних векторів, тобто виділити найбільш важливі властивості об'єктів. Зменшення розмірності прискорює процес кластеризації, і в деяких випадках дозволяє візуально оцінювати її результати. Виділені характеристики варто нормалізувати. Далі всі об'єкти представляються у вигляді характеристичних векторів. Ми будемо повністю ототожнювати об'єкт з його характеристичним вектором. У всіх цих випадках може застосовуватися ієрархічна кластеризація, коли великі кластери дробляться на більш дрібні, ті в свою чергу – ще дрібніші і т.д.

Такі завдання називаються завданнями таксономії (таксон *лат. Taxon*, – «порядок, пристрій, організація») – група в класифікації, що складається з дискретних об'єктів, що об'єднуються на підставі спільних властивостей і ознак) і будують не одну розбивку вибірки на класи, які не перетинаються, а

систему вкладених розбивок. Результат таксономії зазвичай представляється не як проста розбивка множини об'єктів на кластери, а у вигляді таксономічного дерева – дендрограми. Дендрограма дозволяє навести кластерну структуру у вигляді плаского графіку незалежно від того, яка розмірність початкового простору. Замість номеру кластеру об'єкт характеризується перерахуванням всіх кластерів, яким він належить, від великого до дрібного. Введемо поняття «таксономічної формули», як способу опису кластерної структури, де буквено-цифрові позиції у формулі позначають відповідні градації предиктора згідно з таблицями 1, 2.

Таблиця 1

Позиція предиктора в таксономічній формулі

Позиція	X	X	XXXXX	Y	X	X	X
Предиктор	Стать	Вік	Зони ушкоджень	Загальна кількість ушкоджень	Місце вчинення злочину	День тижня	Характер дня тижня
Приклад	1	6	345	с	1	7	2

Опис злочину: потерпілий – чоловік, віком понад 60 років, локалізація ушкоджень – тулуб, верхні і нижні кінцівки, кількість ушкоджень – від 3 до 10, злочин вчинений в житловому приміщенні в неділю, не в робочий день. Таксономічна формула потерпілого: 16345с172.

Таблиця 2

Градація предиктора в таксономічній формулі

Предиктор	Градація
Стать	чоловік – 1; жінка – 2
Вік	до 18 – 1; 18-29 – 2; 30-39 – 3; 40-49 – 4; 50-59 – 5; 60 і більше – 6
Зона ушкоджень	голова – 1; шия – 2; тулуб – 3; верхні кінцівки – 4; нижні кінцівки – 5
Загальна кількість ушкоджень	1- А; 2- В; 3-10 – С; 11-20 – D; 21-50 – Е; більш 50 – F;
Місце вчинення злочину	Житлове приміщення - 1; Нежитлове приміщення – 2; Вулиця – 3; Не визначене - 4
День тижня	Понеділок – 1; ...; Неділя – 7
Характер дня тижня	Робочий день – 1; Вихідний день – 2

Формула може бути розширена в разі збільшення розмірності простору (район, час року і т.д.) або зменшена за рахунок об'єднання кластерів. Серед алгоритмів ієрархічної кластеризації виділяють два основних типи. *Дівізімні* або *низхідні* алгоритми розбивають вибірку на все більш і більш дрібні кластери. Більш поширені *агломеративні* або *висхідні* алгоритми, в яких об'єкти об'єднуються у все більш і більш великі кластери. Реалізація цієї ідеї представлена в такому алгоритмі. Спочатку кожен об'єкт вважається окремим кластером. Для одноелементних кластерів природним чином визначається функція відстані

$$R(\{x\}, \{x'\}) = \rho(x, x').$$

Потім запускається процес злиття. На кожній ітерації замість пари найближчих кластерів U і V утворюється новий кластер.

$$W = \bar{U} \cup V$$

Відстань від нового кластера W до будь-якого іншого кластера S обчислюється по відстанях $R(U, V)$, $R(U, S)$, $R(V, S)$, які до цього моменту повинні бути відомі.

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|,$$

де α , β , γ – числові параметри.

Ця універсальна формула узагальнює практично всі розумні способи визначити відстань між кластерами. Вона була запропонована Лансом і Вільямсом у 1967 році [5].

Визначення числа кластерів найпростіше робити шляхом відсікання фінальної ділянки дендрограми. На горизонтальній осі знаходиться інтервал максимальної довжини $|R_{t+1} - R_t|$, і в якості результуючої кластеризації видається безліч кластерів St . Число кластерів одно $K = l - t + 1$. При необхідності можна задати обмеження на мінімальне і максимальне число кластерів $K_0 < K < K_1$.

Найбільш часто зустрічаються кластери при 7-вимірному просторі характеристик (характеристичний просторі): 123a111- 5(0,6%); 123a172- 7(0,8%); 123a341-6(0,7%); 123a351 - 7(0,8%); 123a362- 12(1,4%); 123a372- 8(0,9%); 133a111- 9(1,1%); 133a141 - 8(0,9%); 133a162 - 12(1,4%); 133a172 - 8(0,9%); 143a111 - 11(1,3%); 143a121 - 7(0,8%); 143a131 - 7(0,8%); 143a172 - 7(0,8%); 153a121 - 6(0,7%); 153a162 - 7(0,8%).

У нашому випадку інтерес представляє не тільки таксономічне древо цілком, але й визначення оптимального числа кластерів для подальшої класифікації та побудови системи класифікаційного співставлення.

Кластери, які найбільш частіше зустрічаються при 3-вимірному просторі характеристик (характеристичному просторі): a11- 124(14,6%); a12 - 100(11,8%); c11 - 85(10,0%); c12 - 60(7,1%); a31 - 54(6,4%); b11-43(5,1%); c31 - 40(4,7%); a32- 37(4,4%); a21- 28(3,3%); b12- 27(3,2%); c32- 24(2,8%); d11 - 23(2,7%); c21- 20(2,4%); b31- 18(2,1%); e11-18(2,1%); b32- 15(1,8%); d12-15(1,8%); a22- 13(1,5%); b21- 11(1,3%); c22- 10(1,2%).

Отже, наведені результати з визначенням та обґрунтуванням вибору предикторів (показників), дозволили встановити таке.

1. Кластеризація може застосовуватися при аналізі як перший етап симпліфікації даних. Після виділення схожих груп необхідно застосовувати інші методи.

2. На даний час немає будь-якого універсального алгоритму кластеризації, який був би ефективним на даних різного походження. В основному, використовуються ітеративні методи, при цьому, результат їх застосування істотно залежить від правильності оцінки кількості кластерів і розмірності простору характеристик.

3. Систематизація кластерів показників трирівневої структури КХЗ (елементи, ознаки, значення) в різних характеристичних просторах (3-

вимірному, 7-вимірному тощо) у вигляді «таксономічної формули», в подальшому дозволить (що і є кінцевою метою наших досліджень) на початковому етапі розслідування злочинів, коли існує певний дефіцит інформації, отримати можливість розгляду версій щодо невстановлених обставин, а саме – визначення «таксономічної формули» злочинця.

Список використаних джерел

1. Балугина Т.С. Криминалистика. Общие положения методики расследования
<http://hghltd.yandex.net/yandbtm?fmode=inject&url=http%3A%2F%2Fwww.ref.by%2Frefs%2F37%2F39377%2F1.html&text=%D0%9A%D1%80%D0%B8%D0%BC%D0%B8%D0%BD%D0%B0%D0%BB%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F%20%D1%85%D0%B0%D1%80%D0%B0%D0%BA%D1%82%D0%B5%D1%80%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0%20%D0%BF%D1%80%D0%B5%D1%81%D1%82%D1%83%D0%BF%D0%BB%D0%B5%D0%BD%D0%B8%D0%B9%3A%20%D1%81%D1%83%D1%89%D0%BD%D0%BE%D1%81%D1%82%D1%8C%20%D1%81%D1%82%D1%80%D1%83%D0%BA%D1%82%D1%83%D1%80%D0%B0%20%D0%B8%20%D1%81%D0%B> -
YANDEX_597пре ступлений / Т.С. Балугина. -
<http://hghltd.yandex.net/yandbtm?fmode=inject&url=http%3A%2F%2Fwww.ref.by%2Frefs%2F37%2F39377%2F1.html&text=%D0%9A%D1%80%D0%B8%D0%BC%D0%B8%D0%BD%D0%B0%D0%BB%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F%20%D1%85%D0%B0%D1%80%D0%B0%D0%BA%D1%82%D0%B5%D1%80%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0%20%D0%BF%D1%80%D0%B5%D1%81%D1%82%D1%83%D0%BF%D0%BB%D0%B5%D0%BD%D0%B8%D0%B9%3A%20%D1%81%D1%83%D1%89%D0%BD%D0%BE%D1%81%D1%82%D1%8C%20%D1%81%D1%82%D1%80%D1%83%D0%BA%D1%82%D1%83%D1%80%D0%B0%20%D0%B8%20%D1%81%D0%B> -
YANDEX_599Лекции. КубГУ, 2002-2003 г.
2. Фесик П.Ю. Технология использования криминалистической характеристики в раскрытии убийств / П.Ю. Фесик // Дис. ... канд. юридических наук: 12.00.09. – Нижний Новгород, 2011. – 239 с.
3. Зосіменко В.В. Статистичне моделювання ознак, які індивідуалізують особу за наслідками її агресивних дій з використанням гострих предметів / В.В. Зосіменко // Судово-медична експертиза. – 2014. – №2. – С. 23-27.
4. Зосіменко В.В. Експертний аналіз випадків вбивств із застосуванням гострих предметів як основа криміналістичної характеристики злочину / В.Д. Мішалов, В.В. Зосіменко // Інтегративна антропологія. – 2014. – №2(24). – С. 33-37.
5. Уиллиамс У.Т., Ланс Д.Н. Методы иерархической классификации / У.Т. Уиллиамс, Д.Н. Ланс // Статистические методы для ЭВМ (под ред. М.Б. Малютова). – М.: Наука, 1986. – С. 269-301.