УДК 519.816+519.237.8

Мулеса О.Ю., викладач

Послідовний аналіз варіантів в нечітких задачах кластеризації та ідентифікації

Розглядається задача виявлення у заданій множині близьких між собою за деякою ознакою елементів у випадку, коли не задано метрику або ознаки об'єктів є якісними. Пропонується модифікація методу послідовного аналізу варіантів до розв'язання поставлених задач. Розглядається прикладна задача та схема її розв'язання.

Ключові слова: кластеризація, ідентифікація, послідовний аналіз варіантів, нечітка база знань, група ЖСБ.

Ужгородський національний університет,
88000, м. Ужгород, вул. Підгірна, 46,
e-mail: mulesa.oksana@gmail.com

O. Y. Mulesa, teacher

A sequential analysis of variants in clustering and identification

The problem of detecting in a given closely related set for some sign of the elements when not specified is metric, or features of objects are qualitative. A modification of the sequential analysis of variants to solve tasks is proposed. An application tasks and scheme to solve it is considered..

Key words: clustering, identification, sequence analysis of variants, fuzzy knowledge base, a group of female sex workers.

National University of Uzhgorod, 88000,
Uzhgorod, Pidgirna str., 46,
e-mail: mulesa.oksana@gmail.com

Статтю представив д.т.н., проф. Волошин О.Ф.

Вступ

Задачі виявлення в заданій множині підмножини близьких між собою за деякою ознакою чи сукупністю ознак елементів та задача оцінки кількості елементів серед заданих, які належать певній групі, виникають в багатьох прикладних задачах. Зокрема, в соціології, медицині, політології виникає проблема виявлення в суспільстві певних груп населення, які об'єднані між собою спільними ознаками (родом занять, інтересами, звичками, особливостями здоров'я тощо) [1]. Існує декілька підходів до побудови математичної моделі та пошуку розв'язків для такого класу задач. До прикладу, такі задачі можуть бути сформульовані як задачі кластеризації [2] та розв'язуватися одним з класичних методів. Проте зустрічаються випадки, коли застосування відомих методів є недоцільним або утрудненим через специфічні особливості природи прикладної задачі (неможливість введення метрики, переважна більшість ознак об'єктів якісні тощо) [1]. Тому доцільним є розробка підходів до розв'язання згаданого класу задач у випадку, коли відомими методами знайти розв'язок, який є прийнятним для особи, що приймає рішення, є неможливим.

Наприклад, з метою профілактики поширення ВІЛ-інфекції та проведення оцінки кількості жінок, які належать до групи ЖСБ (жінок секс-бізнесу) в різних містах України проводяться соціологічні дослідження, робляться спроби побудувати соціально-демографічний портрет жінки, що надає платні сексуальної послуги [3]. В [4] запропоновано підхід до оцінки кількості жінок секс-бізнесу, який базується на використанні статистичної інформації. В основі даного підходу лежить використання методики RDS [4], теоретичною основою якої є теорія ланцюгів Маркова та теорія малих мереж та методики TLS [4] – вибіркового методу, відповідно до якого вибір учасників дослідження відбувається на спеціальних територіях (в місцях збору "цільової" групи).

Застосування такого підходу вимагає великих фінансових витрат із залученням міжнародних фондів, що не завжди є можливим та доцільним. До того ж, при такому підході здійснюються грубі узагальнення, які нівелюють вплив на вибір життєвого шляху таких важливих чинників як "релігійність" того чи іншого регіону, рівень освіченості населення місцевості тощо. Тому необхідною є побудова моделі, яка дала б змогу скоротити трудові та фінансові витрати при побудові соціально-демографічного портрету особи та при цьому максимально узагальнювала би всі фактори, що є складовими соціально-демографічного портрету, могла б врахувати не тільки статистичні дані, а й знання експертів.

Постановка задачі

Виділимо дві взаємопов'язані задачі.

Задача визначення оціночної кількості елементів із заданої множини, які належать певній її підмножині. Нехай задано множину об'єктів та множину ознак (критеріїв), за якими ці об'єкти оцінюються. Для кожного з об'єктів відомі значення за кожною з ознак. Необхідно оцінити скільки об'єктів із заданої множини належать певній її підмножині. Вказана підмножина описується правилами (співвідношеннями), побудованими на основі експертних опитувань щодо характеристик об'єктів, які належать підмножині. В такому представленні задачу можна віднести до задачі кластеризації [2], формальна постановка якої є такою: задано множину об'єктів *X*, необхідно розбити їх на групи (кластери) близьких між собою об'єктів.

Задача визначення міри належності об'єкта до заданої множини (кластеру). Нехай для заданого об'єкта, який характеризується набором ознак, необхідно оцінити міру його належності до деякої групи. Група задається правилами, побудованими на основі експертних опитувань щодо характеристик об'єктів, які їй належать. В такій постановці дана задача може бути віднесеною до задач ідентифікації [5].

Алгоритми кластеризації та ідентифікації є ефективними у випадку, коли задана метрика, за якою встановлюється відстань між об'єктами (для задач кластеризації), або вхідні та вихідні величини мають кількісний характер (для задач ідентифікації). В інших випадках застосування згаданих алгоритмів є утрудненим або вимагає допоміжних прийомів для приведення початкових даних до необхідної форми.

В роботі пропонується підхід до розв'язання поставлених задач, який базується на методі послідовного аналізу варіантів [6, 7] та не вимагає введення метрики між об'єктами.

Формальна постановка задачі та схема її розв'язування

Нехай задано множину об'єктів $O=\{O_1,O_2,\ldots,O_n\}$. Необхідно оцінити кількість об'єктів з заданої множини, які належать певній її підмножині G. Оцінка проводиться на основі відомих значень ознак (критеріїв) для кожного об'єкта з початкової множини. Значення критеріїв можуть бути як числовими результатами експериментів, так і нечисловими характеристиками об'єктів.

Характеристики виділеної групи визначаються експертами шляхом формування нечіткої бази знань [8] для заданих наперед лінгвістичних змінних [9].

На множині об'єктів задається нечітка множина належності до підмножини G. Позначимо через $\mu_\Omega(\mathrm{O_j})$ - міру належності об'єкта O_j підмножині G, тобто $\Omega=\{(O_j,\mu_\Omega(O_j))\}$. Об'єкт O_j належить G з певним ступенем належності, якщо його міра належності $\mu_\Omega(\mathrm{O_j})$ більша за деякий наперед визначений "допуск" [6] : $O_j \overset{\mu_j}{\in} G$, якщо $\mu_\Omega(\mathrm{O_j})>\Delta$, інакше, $O_j \notin G$, якщо $\mu_\Omega(\mathrm{O_j})\le\Delta$.

Нехай задано множину об'єктів $O=\{O_1,O_2,\ldots,O_n\}$. Позначимо через $J=\{1,2,\ldots,n\}$ - множину індексів об'єктів, що розглядаються. Нехай також задано множину ознак (критеріїв) $K=\{K_i\}$, $i=\overline{1,m}$, за якою необхідно провести оцінку.

Для кожного критерію експертами формується лінгвістична змінна [9] та проводиться розбиття множини критеріїв на групи рівнозначних взаємонезалежних критеріїв відповідно до їх впливу на оцінку. Нехай в результаті було отримано L груп критеріїв. Позначимо множину критеріїв рівня l через I_l.

Для кожної групи критеріїв, на основі інформації, отриманої від експертів, формуємо нечітку базу знань типу:

якщо $x_1\in A_{l1}$, $x_2\in A_{l2}$, …, $x_k\in A_{lk}$, то $y\in C$,

де $l\in\{1,\ldots,L\}$, k – кількість критеріїв на відповідному рівні ієрархії.

Далі кожним двом сусіднім рівням ієрархії поставимо у відповідність правило (функцію, згортку тощо) u_l $(l\in\{1,\ldots,L-1\})$, за допомогою якого буде проводитися узгодження оцінки об'єкта між рівнями ієрархії.

Алгоритм послідовного аналізу варіантів для задачі кластеризації

Опишемо алгоритм відсіву $\tilde{W}_2$ об'єктів, який базується на послідовному аналізі варіантів [6,7].

Будуємо таблицю $V=\{v_{ji}\}$, в якій i – стовпцю відповідає i – тий об'єкт з групи O, а j-ому рядку – j-тий критерій, за яким проводиться оцінка.

Елементами таблиці є значення ознак для об'єктів. Позначимо $J^{(0)} = J$, $V^{(0)} = V$.

Перша ітерація процедури $\tilde{W}_2$. *Крок 1.* На основі інформації, отриманої від експертів, будуємо базу знань для групи ознак першого рівня [9]:

Π_1: якщо $x_1 \in A_{11}^{(1)}, \ldots, x_{T_1} \in A_{1T_1}^{(1)}$, то $y \in C^{(1)}; \cdots$

Π_{τ_1} :якщо $x_1 \in A_{11}^{(\tau_1)}$, …, і $x_{T_1} \in A_{1T_1}^{(\tau_1)}$, то $y \in C^{(\tau_1)}$;

де τ_1 - кількість правил на 1-му рівні ієрархії.

Крок 2. Для передумов кожного правила знаходимо степені істинності: $A_{1j}^{(\xi j)}(v_{ji})$, $\xi_1 = \overline{1,\tau_1}$, $i = \overline{1,m}$, $j \in J^{(0)}$.

Застосовуємо логічний вивід та композицію нечітких множин [8]. В результаті за деяким правилом для кожного об'єкта отримуємо інтервал оцінок $\left[a_1^{(j)}, b_1^{(j)}\right]$.

Крок 3. Виключаємо з розгляду ті об'єкти, для яких $b_1^{(j)} \le \Delta$. З таблиці $V^{(0)}$ формуємо таблицю $V^{(1)}$, шляхом виключення стовпців, що відповідають відсіяним об'єктам. Відповідно отримуємо множину індексів об'єктів $J^{(1)}$, для яких продовжується обчислення оцінок.

Якщо множина $J^{(1)}$ порожня, то або застосовуємо діалогову процедуру для уточнення експертом величини Δ і повертаємося до третього кроку, або робимо висновок, що жоден з заданих об'єктів не належить до групи G.

В результаті виконання першої ітерації відсіюються об'єкти за ознаками першого рівня.

На другій ітерації для ознак другого рівня будується база знань вказаного типу. Для всіх об'єктів, що залишилися в розгляді, обчислюються інтервали оцінок за відповідними ознаками $\left[a_2^{(j)}, b_2^{(j)}\right]$, $j \in J^{(1)}$.

Визначаємо правило, за яким проводиться узгодження інтервалів оцінок отриманих на двох попередніх рівнях. Позначимо узгоджений інтервал $\left[a^{(j)}, b^{(j)}\right]$, $j \in J^{(1)}$. Виключаємо з розгляду ті об'єкти, для яких $b^{(j)} \le \Delta$.

Аналогічно до першої ітерації, формуємо таблицю $V^{(2)}$ та множину індексів $J^{(2)}$.

Якщо $J^{(2)} = \varnothing$, то, або експертом уточнюється допуск Δ і процедура повертається до першої ітерації, або робимо висновок, що жоден з заданих об'єктів не належить до підмножини G, і завершуємо роботу.

L-а заключна ітерація процедури $\tilde{W}_2$.

Отримавши узгоджені інтервали оцінок $\left[a^{(j)}, b^{(j)}\right]$, $j \in J^{(L-1)}$, за заданим правилом u_{L-1}, проводимо їх узгодження з інтервалами $\left[a_L^{(j)}, b_L^{(j)}\right]$, $j \in J^{(L-1)}$. В результаті отримуємо інтервали оцінок $\left[\bar{a}^{(j)}, \bar{b}^{(j)}\right]$. Виключаємо з розгляду ті об'єкти O_j, для яких $\bar{b}^{(j)} \le \Delta$. Формуємо множину індексів $J^{(L)}$. Якщо $J^{(L)} = \varnothing$, то робимо висновок, що жоден з об'єктів з множини O не належить до підмножини G, і завершуємо роботу. В протилежному випадку, проводимо дефазифікацію [8] інтервалу оцінок.

Позначимо дефазифіковані оцінки через μ_j, $j \in J^{(L)}$. Для об'єктів, для яких $\mu_j > \Delta$, робимо висновок, що $O_j \overset{\mu_j}{\in} G$, для об'єктів з $\mu_j \le \Delta$ робимо висновок, що $O_j \notin G$. Тобто:

$$\mu_\Omega(O_j) = \begin{cases} \mu_j, & j \in J^{(L)} \\ 0, & j \in J^{(0)} \setminus J^{(L)} \end{cases}.$$

Алгоритм визначення міри належності об'єкта до заданої групи

Не втрачаючи загальності, припустимо, що утворена ієрархія критеріїв складається з 2-х рівнів. Тоді процес обчислення оцінки об'єкта можна представити у виді такого алгоритму:

Крок 1. Формуємо нечітку базу знань зазначеного виду для групи критеріїв верхнього рівня ієрархії.

Крок 2. Застосовуємо логічний вивід та композицію нечітких множин [8]. В результаті отримуємо інтервал $[a_1, b_1] \subseteq [0,1]$ для оцінки ступеня належності об'єкта певній групі.

Крок 3. Проводимо дефазифікацію отриманого результату. Якщо отримана чітка оцінка задовольняє особу, що приймає рішення (ОПР), то алгоритм закінчено. Якщо ні – крок 4.

Крок 4. Формуємо нечітку базу знань зазначеного виду для групи критеріїв нижнього рівня ієрархії.

Крок 5. Застосовуємо логічний вивід та композицію нечітких множин. В результаті отримуємо інтервал $[a_2, b_2] \subseteq [0,1]$ для оцінки ступеня належності об'єкта певній групі.

Крок 6. Визначаємо правило u_1, за яким проводимо узгодження результатів, отриманих на обох рівнях. В результаті отримуємо новий

інтервал $[a, b] \subseteq [0, 1]$ для оцінки ступеню належності об‘єкта певній групі.

Крок 7. Виконуємо дефазифікацію отриманого результату .

Задача оцінювання чисельності представниць групи ЖСБ

Розглядаючи проблему оцінювання чисельності групи ризику ВІЛ-інфікування, можна виділити дві взаємопов’язані задачі, які випливають одна з одної [1]: задача визначення оціночної чисельності представників групи ЖСБ серед виокремленої групи та задача визначення міри належності особи до групи ЖСБ.

При ров’язуванні обох поставлених задач, вхідними даними можуть бути результати соціологічних досліджень, соціально-демографічні показники тощо.

Для прикладу розглянемо випадок, у якому експертами було виокремлено наступні ознаки: вік особи; сімейний стан; освітній рівень; рід занять; місце проживання. Також вищевказані ознаки було розбито на групи за їх впливом на оцінку:

Група 1. Рід занять.

Група 2. Вік, Сімейний стан.

Група 3. Освітній рівень.

Група 4. Місце проживання .

Введемо вектор характеристик особи $K = \{k_1, k_2, k_3, k_4, k_5\}$, компонентами якого є такі величини: k_1 – компонента, що вказує на вік особи, причому $k_1 \in [14, 52]$ і приймає цілочислові значення; k_2 – компонента, що містить дані про сімейний стан особи, причому $k_2 \in$ {незаміжня особа; розлучена особа; особа, що перебуває в громадянському шлюбі; офіційно заміжня особа}; k_3 –компонента, що містить дані про освітній рівень особи, причому $k_3 \in$ {неповна середня освіта; повна середня освіта; середня спеціальна освіта; вища освіта}; k_4 –компонента, що містить дані про місце проживання особи, причому $k_4 \in$ {особа проживає у тому ж місці, де і працює; особа проживає в сільській місцевості; особа проживає в іншому місті}; k_5 –компонента, що містить дані про рід занять особи, при чому $k_5 \in$ {безробітна особа; особа, що навчається; особа, що працює}.

Всі компоненти вектора характеристик є лінгвістичними змінними для яких сформовано наступні терм-множини:

T_1 = {неповнолітня особа; особа молодого віку; особа зрілого віку; особа старшого віку};

T_2 = {нестабільний сімейний стан, неорганізований сімейний стан, організований сімейний стан};

T_3 = {низький освітній рівень, середній освітній рівень, високий освітній рівень};

T_4 = {віддалене місце проживання, наближене місце проживання};

T_5 = {стабільний рід занять; нестабільний рід занять}.

Функції належності для відповідних нечітких множин доцільно будувати методом парних порівнянь побудови функцій належності, так як універсальні множини кожної змінної є скінченими [10].

Також опишемо лінгвістичну змінну, яка характеризує степінь належності особи до вказаної групи: C = {імовірність належності особи до групи ЖСБ} з наступною терм-множиною:

T_C = {Висока імовірність належності до групи ЖСБ; Низька імовірність належності до групи ЖСБ}.

На сонові даних, отриманих шляхом опитування експертів, представників різних професій, які володіють відповідними знаннями, будуємо таку базу знань:

Для ознак першої групи: “Якщо *рід занять стабільний*, то *імовірність* належності до групи ЖСБ *низька”* тощо.

Для ознак другої групи: “Якщо особа *молодого віку* і *сімейний стан неорганізований*, то *імовірність* належності до групи ЖСБ *висока*” тощо.

Для ознак третьої групи: “Якщо *освітній рівень особи високий*, то *імовірність* належності до групи ЖСБ *низька*” тощо.

Для ознак четвертої групи: “Якщо місце *проживання особи віддалене*, то *імовірність* належності особи до групи ЖСБ *висока*” тощо.

Базу знань доцільно будувати шляхом введення до неї правил з найбільшим рангом, який можна обчислювати як суму рангів експертів, які задали відповідне правило.

Коефіцієнти якості експертів обчислюються в таких задачах, як правило, з використанням документального методу, який дозволяє врахувати такі об’єктивні характеристики експертів, що можуть мати вплив на компетентність, як освіта, досвід роботи в проблемній області, наявність наукових здобутків в даній області тощо.

Дану задачу, в описаній математичній постановці, пропонується розв'язувати за допомогою алгоритму $\tilde{W}_2$.

Числовий експеримент

Нехай дано множину $O = \{O_j\}$, $j = \overline{1,20}$, що складається з 20 об'єктів з наступними характеристиками (Табл. 1):

Таблиця 1. Початкові дані

№	k_1	k_2	k_3	k_4	k_5
1	15	незам.	н.с.	місто	навч
2	16	незам.	н.с.	село	навч
3	17	незам.	н.с.	ін.місто	безроб.
4	18	незам.	с.	ін.місто	прац.
5	25	незам.	с.с.	село	прац.
6	19	незам.	н.с.	село	навч.
7	19	незам.	с.	місто	навч.
8	24	гр.шлюб	в.о.	місто	прац.
9	36	розл.	с.с.	місто	безроб.
10	24	зам.	в.о.	село	безроб
11	32	розл.	в.о.	ін.місто	прац.
12	25	незам.	с.с.	село	безроб.
13	19	незам.	н.с	місто	навч.
14	31	розл.	в.о.	село	прац.
15	25	зам.	н.с.	село	навч.
16	19	незам.	н.с.	село	безроб.
17	19	незам.	с.с.	ін.місто	безроб.
18	32	незам.	в.о	ін.місто	безроб.
19	20	зам.	с.с.	село	навч.
20	19	незам.	н.с	село	навч.

Необхідно визначити, які об'єкти із заданих належать до групи „ЖСБ" (G) та степені належності цих об'єктів групі при $\Delta = 0.75$.

Для застосування процедури формується база знань та функції належності нечітких множин, за допомогою методів описаних вище. Застосовуючи процедуру $\tilde{W}_2$ з чотирма ітераціями, отримаємо, що до вказаної групи, з множини заданих об'єктів, належать чотири об'єкти: O_3, O_{12}, O_{16}, O_{17}, причому ступені належності до групи у перерахованих об'єктів відповідно є такими: 0.80, 0.77, 0.77, 0.83. Як видно з отриманих результатів, на отримані оцінки суттєвий вплив має розбиття ознак на групи за значимістю, тобто за впливом на оцінку. Так як до першої групи експертами віднесений рід занять, то і серед об'єктів, що належать до вказаної груп всі мають одне і те ж значення цієї ознаки - „безробітна". Можна також порівняти значення ознак для O_5 та O_{12}. Як видно, дані об'єкти відрізняються тільки родом занять. Але перший, за результатами роботи процедури, не належить до групи G, а другий – належить зі ступенем належності 0.77.

Висновки

В роботі запропоновано підхід до розв'язування задач кластеризації та ідентифікації на основі експертної та статистичної інформації без використання навчальної вибірки. Кластеризація проводиться з використанням нечітких множин, що дозволяє використовувати нечислову вхідну інформацію. Робота запропонованого алгоритму проілюстровано на прикладній задачі.

Список використаних джерел

1. *Mironyuk I.S.* The results of the estimated population vulnerable to HIV infection groups (female sex workers) in the Transcarpathian region / I.S. Mironyuk, V.J. Shatylo, I.J. Hutsol, V. Brych. // Transcarpathian Centre for Prevention of AIDS– 2010. – P. 21-25. (in Ukrainian)
2. *R.Duda, P.Hart.* Pattern recognition and scene analysis. – M:Mir, 1976. (in Russian)
3. *Shvalahin O.Y., Mlavets Y.Y., Mironyuk Z.B.* Modeling social portrait of the person // Computational Intelligence. - Cherkasy: Maklaut, 2011. - P. 262. (in Ukrainian)
4. *Balakireva, T. Bondar, Sereda Y.* Monitoring the behavior of commercial sex workers, as a component of second generation surveillance. - K.: ICF "International HIV / AIDS Alliance in Ukraine", 2008. - 60 p. (in Ukrainian)
5. *Grop D.* Methods of identification systems. – M:Mir, 1979. - 302 p. (in Russian)
6. *Volkovich V.L., A.F. Voloshin.* The methods for automated design of complex systems. - K.: Nauk. Dumka, 1984. (in Russian)
7. *N.Malyar, O.Shvalagin.* Fuzzy procedures successive analysis of variants in combinatorial problems // Information technologies & knowledge, Intern. Journal, №1, Vol.6, 2012.- P. 81-87. (in Russian)
8. *Snytyuk V.E.* Prediction: Models & Methods. - K.: "Maklaut", 2008. (in Ukrainian)
9. *Zadeh L.* The concept of a linguistic variable. – M.:Mir, 1976. (in Russian)
10. *Melkumova E.M.* Construction methods of fuzzy set membership functions // Bulletin VSU, Series: System Analysis. 2009. №2. - P. 13-18. (in Russian)

Надійшла до редколегії 12.03.13