УДК 510.22

Білик А.С.

Порівняння масивів якісних даних на прикладі не ототожнених явищ

У статті розглядається загальна методика порівняння масивів якісних даних, коли одному масиву треба послідовно зіставити сукупність інших, розподілених за однаковою ієрархічною структурою із урахуванням можливої присутності невизначеності через неповність або брак даних.

Постановка проблеми. Останнім часом м'які обчислення та аналіз якісних даних знаходить усе більш широке застосування, зокрема у прикладній інформатиці, доказовій медицині, економіці тощо. Як правило, для обробки масивів якісних даних використовуються переважно статистичні та комбінаторні методи із наділенням якісним даним кількісних характеристик. Проте якісні дані лінгвістичного характеру можуть бути неповними або не визначними, що ускладнює їх урахування. При аналізі умовного об'єкту, що складається із якісних даних наведеного характеру, іноді важливо знати сукупність об'єктів, що найбільше відповідають заданому.

Аналіз останніх досліджень. Систематизовані загальні методи вибіркових оцінок для баз даних у повному обсязі викладені у праці [4]. Найближча за умовами постановки проблеми методика викладена у роботі [3]. Джерела [1] і [2] містять загальні теоретичні відомості, використані у даній роботі.

Постановка завдання. Нехай ми маємо деяке природне або штучне явище *a*, що володіє множиною характерних якісних ознак, або ж, *проявів*. Приймемо, що явище *a* загальновідоме, і всі його прояви можуть бути досконало вивчені та описані у результаті багаторазових прямих спостережень, вимірювань, експериментів та інших достовірних даних.

Нехай ми маємо також явище *a',* котре також володіє множиною проявів, тотожних або відмінних від проявів явища *a*. Проте явище *a'* на початковому етапі досліджень не є відомим, і методи отримання даних щодо його проявів в цілому обмежені одноразовими візуальними спостереженнями або вимірюваннями, а інші дані не завжди доступні.

Явища, що володіють наведеними властивостями явища *a* будемо називати *ототожненими*. Відповідно, явища, що володіють на початковому етапі досліджень властивостями явища *a' ,* будемо називати *не ототожненими.* Зіставлення явищ типу *a* явищам типу *a'* і оцінка розбіжностей між ними, інакше кажучи, їх *ототожнення* є однією з приоритетних задач дисциплін, що базуються на емпіричній основі, зокрема уфології та інших.

Як правило, на практиці дослідник розглядає кожне з не ототожнених явищ окремо, поступово приймаючи ті або інші гіпотези та аналізуючи їх придатність у ототожненні. З огляду на те, що гіпотези мають вигляд множин проявів та число, як груп розподілу, так і самих гіпотез, може бути достатньо великим, що ускладнює зіставлення, актуальною уявляється задача виведення *єдиного чисельного еквіваленту* для кожної гіпотези, який би спростив їх порівняння.

Виклад основного матеріалу дослідження. Рішення даної задачі викладемо за допомогою теоретико-множинного підходу. У загальному вигляді прояви класифікуються за довільно обраними групами a_i:

$$a_1\{a_{11};a_{12};a_{13}\ldots a_{1m_1}\},$$
$$a_2\{a_{21};a_{22};a_{23}\ldots a_{2m_1}\},$$
$$a_3\{a_{31};a_{32};a_{33}\ldots a_{3m_1}\},$$
$$\ldots$$
$$a_X\{a_{X1};a_{X2};a_{X3}\ldots a_{Xm_X}\}.$$

Елементи виду a_{ij} – прояви у групах. $i= 1, 2, 3 \cdots X$ – номер групи, j – номери проявів у відповідних групах. Індекси $j = m_1 , m_2 , m_3 \cdots m_X$ показують довільність і скінченність числа проявів у кожній з груп. Сукупність груп розподілу проявів $\{a_i\}$ будемо називати *класифікатором* проявів. X - загальна кількість груп розподілу. Очевидно, що число проявів у кожній групі не обов'язково однакове і залежить від вибору та деталізації класифікатору.

З огляду на те, що групи розподілу проявів мають вигляд таблиці зі стовпцями та рядками, їх простіше структурувати у матричній формі із відповідними проявам компонентами.

Явищем-гіпотезою (далі - гіпотезою), що пояснює причину виникнення не ототожненого явища *a'* будемо називати множину проявів, що характеризують будь-яке ототожнене явище *a*, що приймається у розгляд при ототожненні. Кожна з гіпотез представлена у вигляді множини проявів $G_N\{a_{ij}\}$:

$$G_N\begin{Bmatrix} a_1 \\ a_2 \\ a_3 \\ \ldots \\ a_X \end{Bmatrix},$$

де $a_1, a_2, a_3 \cdots a_X$ – підмножини, що показують загальні групи проявів, а N – номер гіпотези, що приймається у розгляд при ототожненні. Кожна з гіпотез може як містити більше одного прояву з кожної групи, так і не містити їх взагалі. Об'єднання

$$G_1 \cup G_2 \cup G_3 \cup \cdots \cup G_N \cdots \cup G_L = G\begin{Bmatrix} a_{11};a_{12};a_{13}\ldots a_{1m_1} \\ a_{21};a_{22};a_{23}\ldots a_{2m_1} \\ a_{31};a_{32};a_{33}\ldots a_{3m_1} \\ \ldots \\ a_{X1};a_{X2};a_{X3}\ldots a_{Xm_X} \end{Bmatrix}$$

де L – кількість гіпотез, утворює загальну множину проявів, що містяться у всіх існуючих гіпотезах, котру далі будемо називати *основним масивом проявів.*

Кожне з не ототожнених явищ позначимо як множини $A_K\{a_{ij}\}$, що матимуть вигляд варіацій проявів у відповідних групах:

$$A_K\begin{Bmatrix} a_1 \\ a_2 \\ a_3 \\ \ldots \\ a_X \end{Bmatrix},$$

де $a_1, a_2, a_3 \cdots a_X$ – підмножини, що показують загальні групи проявів, аналогічні групам основного масиву, але різні за місткістю проявів. K – порядковий номер не ототожненого явища, що розглядається. Кожне з явищ може як містити більше одного прояву з кожної відповідної групи, так і не містити їх взагалі. Для збереження відповідності

у такому випадку на місці відсутніх проявів ставиться нуль. Якщо прояви у деяких групах відсутні через недостатність наявних даних щодо явища, у таких групах проставляється символ, що означає невизначеність, тобто що припускає можливу присутність проявів. Фактично, чим більше співвідношення груп проявів, у яких проставлено невизначеність, до загальної кількості груп, що містять прояви, тим менша кількість показників, за якими провадиться порівняння, і, відповідно, знижується достовірність оцінки розбіжностей. Також можливий варіант, коли в A_K входять прояви, що не відносяться до жодної з груп, що містяться у основному масиві. Такі прояви логічно розподіляються за групами і отримують індекс j, відмінний від записаних у основному масиві.

Для подальших викладок потрібне сумарне число компонентів у кожній з груп основного масиву для гіпотез, та не ототожнених явищ, що розглядаються:

$$G'_N \begin{Bmatrix} g_1 \\ g_2 \\ g_3 \\ \cdots \\ g_X \end{Bmatrix}, A'_K \begin{Bmatrix} t_1 \\ t_2 \\ t_3 \\ \cdots \\ t_X \end{Bmatrix}$$

де компоненти g_1 , g_2 , g_3 , ... g_X (для гіпотез) та t_1, t_2, $t_3 \cdots t_X$ (для явищ, що розглядаються) являють собою числа компонентів у відповідних групах розподілу, і знаходяться, як $g_i(i = 1,2,3,\cdots X) = \sum_{j=1}^{m_i} g_{ij}$, $t_i(i = 1,2,3,\cdots X) = \sum_{j=1}^{m_i} t_{ij}$, g_{ij} та t_{ij} приймають значення за умовами імплікації

$$\begin{cases} (a^{G}_{ij} = \overline{\varnothing}_j) \to (g_{ij} = 1), \\ (a^{G}_{ij} = \varnothing_j) \to (g_{ij} = 0); \end{cases} \quad \begin{cases} (a^{A}_{ij} = \overline{\varnothing}_j) \to (t_{ij} = 1), \\ (a^{A}_{ij} = \varnothing_j) \to (t_{ij} = 0); \end{cases}$$

де $\varnothing_j$ – одномісні компонент-підмножини, а індекси A і G показують множину - джерело, з якої було обрано компонент. Групи, у яких проставлена відсутність проявів, у сумі будуть давати нуль, а ті, у яких проставлена недостатність даних – 1, тобто номінально припускається, що такі підмножини можуть містити хоча б один прояв, відмінний, проте від проявів основного масиву.

Для оцінки розбіжностей між не ототожненими явищами A_K та явищами-гіпотезами G_N, слід зіставити обрані згідно класифікатору прояви з відповідними у кожній з прийнятих гіпотез. Для цього для кожної G_N гіпотези знаходиться множина спільних проявів S_N – перетин множин проявів, які містяться у гіпотезі, і у не ототожненому явищі, що розглядається, відповідно групам розподілу:

$$S_N \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_X \end{Bmatrix} = G_N \cap A_K = G_N \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_X \end{Bmatrix} \cap A_K \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_X \end{Bmatrix},$$

Для кожної групи множини S_N обчислюється число проявів, що збіглись :

$$S'_N\begin{Bmatrix} s_1 \\ s_2 \\ s_3 \\ \cdots \\ s_X \end{Bmatrix} = \begin{Bmatrix} s_{11} + s_{12} + s_{13} + ... + s_{1m_1} \\ s_{21} + s_{22} + s_{23} + ... + s_{2m_2} \\ s_{31} + s_{32} + s_{33} + ... + s_{3m_3} \\ \cdots \\ s_{X1} + s_{X2} + s_{X3} + ... + s_{Xm_X} \end{Bmatrix},$$

s_1 , s_2 , s_3 , ... s_X – число спільних проявів для окремо узятого явища і гіпотези, що розглядається у кожній з груп. Числа s_{11} , s_{12} , s_{13} , ... s_{Xm_X} характеризують збіг кожного з компонентів, і приймають значення за умовами: $\begin{cases} (\frac{a^{A}_{ij} + a^{G}_{ij}}{2} = a_{ij}) \rightarrow (s_{ij} = 1), \\ (\frac{a^{A}_{ij} + a^{G}_{ij}}{2} = \bar{a}_{ij}) \rightarrow (s_{ij} = 0); \end{cases}$

Якщо в A_K та G_N у відповідних групах збігаються нульові значення, то проявів, що збіглися не буде, і такі групи взаємно викреслюються і надалі не враховуються. Групи A_K, у яких проставлена невизначеність при зіставленні дають нульове значення.

Сума проявів по групам для множин A'_K , G'_N , та S'_N визначається як:

$$G'_N = g_1 + g_2 + g_3 + \cdots + g_X = \sum_{n=1}^{X} g_n \; ; S'_N = s_1 + s_2 + s_3 + \cdots + s_X = \sum_{n=1}^{X} s_n \; ;$$

$$A'_K = t_1 + t_2 + t_3 + \cdots + t_X = \sum_{n=1}^{X} t_n \; .$$

Число, що характеризує пріоритетність даної гіпотези G_N при її оцінці для ототожнення окремого, на початковому етапі досліджень не ототожненого, явища A_K, будемо називати *застосовністю* даної гіпотези. Математично, застосовність виражає відношення суми усіх проявів, що збіглись по групам розподілу у даному не ототожненому явищі, та прийнятої у розгляд гіпотези, до більшого з чисел сумарних компонентів в гіпотезі, або у самому явищі.

Застосовність позначається літерою P , та знаходиться як

$$\begin{cases} (A'_K < G'_N) \rightarrow (P_N = S'_N / G'_N), \\ (A'_K > G'_N) \rightarrow (P_N = S'_N / A'_K), \\ (A'_K = G'_N) \rightarrow \{(P_N = S'_N / G'_N) \vee (P_N = S'_N / A'_K)\}; \end{cases}$$

причому очевидно, що $P_N \in [0;1]$.

Наведені вище записи та підрахунки проводяться послідовно для кожної гіпотези. Якщо для якої-небудь гіпотези виконується умова $(P_N = 1) \wedge (\sum_{N=1}^{L} (P_N = 1) = 1)$, тобто якщо існує одна і тільки одна гіпотеза, яка містить усі прояви явища A_K , застосовність даної гіпотези до не ототожненого явища, що розглядається, можна вважати умовно[1] високою, а гіпотезу – прийнятою.

Однак такий варіант зустрічається достатньо рідко, оскільки потребує повної відповідності всіх проявів, що містяться у явищі, що розглядається, з усіма проявами в одній гіпотезі, та часткової або повної не відповідності проявів у всіх інших гіпотезах.

Теоретично можливий варіант $(P_N = 1) \wedge (\sum_{N=1}^{L} (P_N = 1) > 1)$, тобто умовно високу застосовність має більш, ніж одна гіпотеза: $(P_1 = S'_1 / A'_1 = 1) \wedge (P_2 = S'_2 / A'_1 = 1)$,але тоді

[1] Умовність пов'язана із достовірністю проявів, що містяться в A_K , перевірка якої не є предметом даної роботи.

$S_1' = A_1'$, $S_2' = A_1'$, $S_1' = S_2'$, отже $G_1 = G_2$, що свідчить пре невірне складання основного масиву.

З практичної точки зору, найбільш розповсюдженим є варіант, коли $(0 < P_N < 1) \wedge (\sum_{N=1}^{L} (P_N = 1) = 0)$, тобто жодна з гіпотез не містить у повному обсязі проявів, записаних в A_K:

$A_K \neq G_1 \neq G_2 \neq G_3 \neq ... \neq G_N \neq ... \neq G_L$, $A_K \subseteq G$, $C_{A_K} G = A_K \setminus G = \varnothing$;

або A_K містить окрім проявів, належних основному масиву, додаткові прояви, що не входять у основний масив проявів (рис. 1):

$A_K \neq G_1 \neq G_2 \neq G_3 \neq ... \neq G_N \neq ... \neq G_L$, $C_{A_K} G = A_K \setminus G \neq \varnothing$.

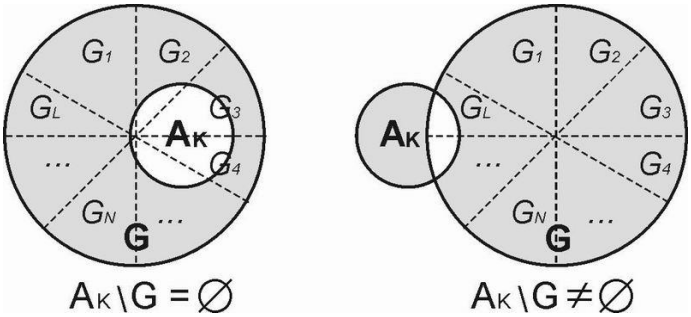

Рис.1 Випадки перетину A_K із основним масивом у діаграмах Вена.

В останньому вказаному випадку потрібен окремий аналіз кожного додаткового прояву з метою встановлення його сполучуваності із загальною картиною реєстрації явища. У залежності від джерела даних щодо проявів слід враховувати людський фактор (суб'єктивне сприйняття людей-регістраторів явищ) а також похибки та особливості застосування використовуваної апаратури.

Якщо додаткові прояви дійсно мали місце, і не являються результатом помилки (невірного тлумачення людиною-регістратором проявів, що містяться у основному масиві, або, відповідно, неточності даних приладових вимірювань), то значення s_i для кожної i-групи, у яку записуються додаткові прояви, знижується на їх число. А отже, знижується і загальна застосовність P_N кожної з гіпотез. Якщо для будь-яких двох і більше гіпотез значення P_N рівні, то такі гіпотези є рівно застосованими.

З будь-якої застосовності гіпотези G_N може бути також знайдена її *незастосовність*, інакше кажучи, не ототожненість явища за даною гіпотезою:

$$U_N = 1 - P_N .$$

Для систематизації результатів складається зведена таблиця вигляду табл.1.

Таблиця 1

A_K	G_1	G_2	G_3	...	G_N	...	G_L
P	P_1	P_2	P_3	...	P_N	...	P_L
r^p	r^p_1	r^p_2	r^p_3	…	r^p_N	…	r^p_L

r^p – ранги гіпотез, за якими у відповідності до мети оцінки обираються приоритетні застосовності. Графічна візуалізація результатів можлива за допомогою побудови гістограми значень застосовності (або незастосовності) гіпотез, зразок якої представлено на рис.2 .

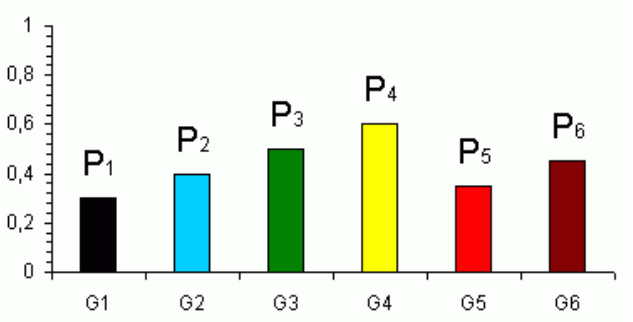

Рис.2 Зразок гістограми застосовності гіпотез

Процес зіставлення множин явищ можливо автоматизувати за допомогою ЕОМ, створивши спеціалізовані експертні системи, що використовують тематичні бази даних.

Слід також зазначити, що представлена методика порівняння гіпотез, що пояснюють природу не ототожнених явищ, не тільки не залежить від жорстко обраного класифікатору, але й від предмета дослідження взагалі, тобто застосовна при зіставленні будь-яких множин якісних даних.

Список літератури:

1. *Бронштейн И.Н., Семендяев К.А.,* Справочник по математике для инженеров и учащихся ВТУЗОВ. – М.: Наука, 1981. – С. 496–503.

2. *Корн Г., Корн Т.*, Справочник по математике для научных работников и инженеров. – М.: Наука, 1984. – С. 99–102.

3. *Бериков В.Б.,* Построение решающей функции в задаче анализа структурированных объектов в условиях неопределенности/ Сборник докладов конф. по мягким вычислениям и измерениям. SCM'99, -СПб.: СПГЭТУ, 1999.

4. *Микони С.В., Козченко, Р.В., Созоновский П.Г.,* Выбор наилучших вариантов из баз данных/ Сборник докладов конф. по мягким вычислениям и измерениям. SCM'99, -СПб.: СПГЭТУ, 1999.