УДК 681.5

# МЕТОД АВТОМАТИЧНОГО УТВОРЕННЯ ГІПОТЕЗ З КВАНТОРОМ БАГАТОВИМІРНОЇ АСОЦІАЦІЇ В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

к.т.н. В.Л. Петров, к.т.н. А.О. Феклістов, О.О. Феклістов
(представив д.т.н. Г.В. Пєвцов)

*В статті пропонується метод автоматичного утворення гіпотез з квантором багатовимірної асоціації, який може використовуватись у складі математичного забезпечення інтелектуальних систем підтримки прийняття рішень.*

**Постановка проблеми.** Одним з видів інформаційно-аналітичних матеріалів, що використовуються в процесі підтримки прийняття рішень, є гіпотези про наявність або відсутність асоціативних зв'язків між ознаками об'єктів (подій, явищ). Найбільш цікавим представляється дослідження методів автоматичного утворення гіпотез на основі індуктивних логічних числень [1], тому що вони, на відміну від дедуктивних методів, працюючих із заздалегідь заданою множиною правил, орієнтовані на роботу з фактично наявним емпіричним матеріалом.

**Аналіз літератури**. Проведений аналіз робіт в області автоматичного утворення гіпотез з кванторами асоціації показав, що сьогодні існує ряд робіт, присвячених методам автоматичного утворення гіпотез з квантором бінарної асоціації [1], визначення для квантора бінарної асоціації чисельної міри асоціативної схожості [2], визначення чисельної міри асоціативної схожості з урахуванням логічної невизначеності вхідних даних [3]. Крім того, існують й інші підходи до побудови асоціативних правил [4, 5]. Аналіз ряду практичних задач показав, що в деяких випадках актуальним є вирішення задачі розробки методу автоматичного утворення гіпотез з квантором багатовимірної асоціації.

**Мета статті.** Метою цієї роботи є представлення методу автоматичного утворення гіпотез з квантором багатовимірної асоціації без урахування логічної невизначеності вхідних даних на основі атрибутованого бінарного дерева.

Згідно [1] під бінарною асоціацією розуміється таке відношення між двома ознаками $x_i$ і $x_j$ ($i \neq j$) деякого об'єкта, при якому «збіг си-

© к.т.н. В.Л. Петров, к.т.н. А.О. Феклістов, О.О. Феклістов, 2004

льніший за відмінність», тобто є кількість записів (в термінології реляційних баз даних (БД)), в яких ознаки $x_i$ і $x_j$ водночас мають або водночас не мають певні значення, більша за кількість інших записів, в яких одна з ознак не має даного значення. Ознаки об'єкта представляються множиною $X = \{x_1, \ldots, x_n\}$. На основі функцій відповідності [1] (правил логічної інтерпретації [3]) їм ставляться у відповідність унарні предикати $\vartheta(\tau) = \{\varphi_1(\tau), \ldots, \varphi_n(\tau)\}$, що приймають три логічні значення («істина» (1), «неправда» (0) і «не визначене» (x)). Логічне значення записується перед предикатом в круглих дужках. Наприклад, запис (1) $\varphi(\tau)$ означає, що унарний предикат $\varphi(\tau)$ має значення «істина». Змінна «τ» відповідає покажчику запису в файлі реляційної бази даних. Введемо функцію $|\{M\}|$, що визначає потужність (кількість елементів) множини $\{M\}$. Коефіцієнти кількості записів в БД, для яких предикати мають задані значення, розраховуються наступним чином:

$$a_{11} = |\{\tau|(1)\varphi_i(\tau) \wedge (1)\varphi_j(\tau)\}|;\ a_{10} = |\{\tau|(1)\varphi_i(\tau) \wedge (0)\varphi_j(\tau)\}|;$$

$$a_{01} = |\{\tau|(0)\varphi_i(\tau) \wedge (1)\varphi_j(\tau)\}|;\ a_{00} = |\{\tau|(0)\varphi_i(\tau) \wedge (0)\varphi_j(\tau)\}|.$$

В роботі [2] за чисельну міру асоціації пропонується використовувати величину «зворотна спряженість» (Δ), що розраховується наступним чином: $\Delta = \frac{p_{11} * p_{00}}{p_{10} * p_{01}}$, де $p_{ij} = \frac{a_{ij}}{m}$ – імовірність виконання формули $(i)\varphi(\tau) \wedge (j)\psi(\tau)$, $i, j \in \{0, 1\}$, m – загальна кількість записів БД. Далі в роботі вказується, що кожна розумна міра залежності є строго монотонною функцією Δ. При цьому, якщо взяти величину $\delta = \log(\Delta)$ (логарифмічна спряженість), то при $\delta > 0$ залежність позитивна, при $\delta < 0$ залежність негативна, а при $\delta = 0$ – залежність буде відсутня [1, с. 115]. В роботі [3] чисельна міра квантора бінарної асоціації $\gamma_{\approx_2}$ вводиться як величина, зворотна «зворотної спряженості» Едвардса: $\gamma_{\approx_2} = 1/\Delta = \frac{p_{10} \cdot p_{01}}{p_{11} \cdot p_{00}} = \frac{a_{10} \cdot a_{01}}{a_{11} \cdot a_{00}}$. Величина $\gamma_{\approx_2} \in [0 \ldots +\infty)$. Вона не визначена в двох випадках: коли чисельник і знаменник в формулі дорівнюють ($\gamma_{\approx_2} = 1$), і коли обчислення чисельника або знаменника неможливе внаслідок рівності 0 одного з її складових коефіцієнтів. Гіпотеза з квантором бінарної асоціації має наступний вигляд: $\approx_2 (\varphi_i(\tau), \varphi_j(\tau))$.

Під багатовимірною асоціацією пропонується розуміти гіпотезу $\approx_n (\varphi_1(\tau), \varphi_2(\tau), \ldots, \varphi_n(\tau))$, що описує таке відношення асоціації між n ознаками $x_1, x_2, \ldots, x_n$ об'єкта, при якому кількість записів в БД, в яких ці ознаки водночас мають або не мають певні значення, більше кількості записів, в яких одна або деякі з ознак не мають даних значень.

Для обчислення чисельної міри гіпотез з квантором багатовимірної асоціації ($\gamma_{\approx_n}$) пропонується метод, оснований на представленні коефіцієнта a у вигляді вузлів атрибутованого бінарного дерева (рис. 1). Далі при вживанні поняття «вузол» буде підрозумітись «коефіцієнт» формули для обчислення чисельної міри квантора багатовимірної асоціації.

Кожний вузол має обов'язковий атрибут «код», що є послідовністю нулів та одиниць (логічні значення унарних предикатів). Наприклад, код вузла $a_{1101}$ є «1101». Даний атрибут є основою алгоритму класифікації вузлів на дві рівних множини: $\{a^+\}$ – множина

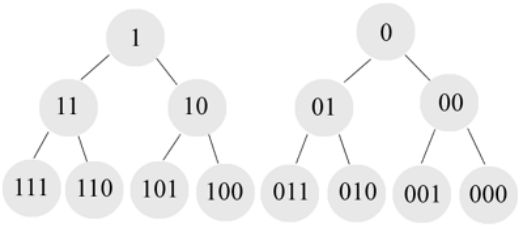

Рис. 1. Фрагмент атрибутованого бінарного дерева для представлення коефіцієнта a у вигляді вузлів

вузлів, які характеризують асоціативну схожість і $\{a^-\}$ – множина вузлів, які характеризують асоціативну відмінність. Для обчислення чисельної міри квантора багатовимірної асоціації пропонується наступна формула:

$$\gamma_{\approx_n} = \prod_{i=1}^{n^-} a_i^- \Big/ \prod_{i=1}^{n^+} a_i^+ ,$$

де П – операція добутку; $n = |\{a\}|$ – загальна кількість вузлів; $n^+ = |\{a^+\}|$ – кількість вузлів асоціативної схожості; $n^- = |\{a^-\}|$ – кількість вузлів асоціативної відмінності: $\{a\} = \{a^+\} \cup \{a^-\}$; $n = n^+ + n^-$; $n^+ = n^- = 1/2\, n$.

Аналіз наведених формул показує, що для класифікації достатньо визначити одну з множин $\{a^+\}$ або $\{a^-\}$. Далі буде розглядатися метод визначення $\{a^+\}$. При цьому $\{a^-\}$ є різницею $\{a\}$ і $\{a^+\}$: $\{a^-\}=\{a\}\setminus\{a^+\}$. Алгоритм визначення множини $\{a^+\}$ оснований на введенні ряду атрибутів для вузла і зв'язаних з ними коефіцієнтів. В табл. 1 представлені всі атрибути і коефіцієнти, що використовуються в даному методі.

Таблиця 1

Класифікаційні атрибути і коефіцієнти атрибутованого бінарного дерева

| Атрибут вузла | Семантика позначення атрибута вузла | Коефіцієнт кількості записів |
|---|---|---|
| $a$ | вузол бінарного дерева | $n = \|\{a\}\|$ |
| $a^+$ | вузол «асоціативної схожості» | $n^+ = \|\{a^+\}\|$ |
| $a^-$ | вузол «асоціативної відмінності» | $n^- = \|\{a^-\}\|$ |
| $a_{=1}$ | код має всі 1 | $n_{=1} = \|\{a_{=1}\}\|$ |
| $a_{=0}$ | код має всі 0 | $n_{=0} = \|\{a_{=0}\}\|$ |
| $a_{>1}$ | кількість 1 більша за кількість 0 | $n_{>1} = \|\{a_{>1}\}\|$ |
| $a_{>0}$ | кількість 0 більша за кількість 1 | $n_{>0} = \|\{a_{>0}\}\|$ |
| $a_{0=1}$ | кількість 1 дорівнює кількості 0 | $n_{0=1} = \|\{a_{0=1}\}\|$ |
| $a_?$ | вузол множини $\{a_?\} = \{a_{>1}\} \cup \{a_{>0}\}$ | $n_? = \|\{a_?\}\|$ |
| $a_?^+$ | вузол «можливої асоціативної схожості» | $n_?^+ = \|\{a_?^+\}\|$ |
| $a_?^-$ | вузол «можливої асоціативної відмінності» | $n_?^- = \|\{a_?^-\}\|$ |
| $a_{>1}^\Delta$ | вузол типу $a_{>1}$, в якому кількість 1 більша за кількість 0 на величину $\Delta$ | $n_{>1}^\Delta = \|\{a_{>1}^\Delta\}\|$ |
| $a_{>0}^\Delta$ | вузол типу $a_{>0}$, в якому кількість 0 більша за кількість 1 на величину $\Delta$ | $n_{>0}^\Delta = \|\{a_{>0}^\Delta\}\|$ |

Об'єднання всіх вузлів є: $\{a\} = \{a_{=1}\} \cup \{a_{=0}\} \cup \{a_{>1}\} \cup \{a_{>0}\} \cup \{a_{0=1}\}$. Коефіцієнт n дорівнює: $n = n_{=1} + n_{=0} + n_{>1} + n_{>0} + n_{0=1}$. Позначимо множину вузлів $\{a_{=1}\}$, $\{a_{=0}\}$, $\{a_{>1}\}$ і $\{a_{>0}\}$ як $\{a_\Sigma\}$: $\{a_\Sigma\} = \{a_{=1}\} \cup \{a_{=0}\} \cup \{a_{>1}\} \cup \{a_{>0}\}$. Потужність $\{a_\Sigma\}$ є: $n_\Sigma = n_{=1} + n_{=0} + n_{>1} + n_{>0}$.

Процес класифікації вузлів оснований на аналізі виконання однієї з двох умов: коефіцієнт $n_\Sigma = 1/2\ n$ або коефіцієнт $n_\Sigma > 1/2\ n$.

1. Якщо коефіцієнт $n_\Sigma = 1/2\ n$, то вузли множин $\{a_{=1}\}$, $\{a_{=0}\}$, $\{a_{>1}\}$, $\{a_{>0}\}$ є вузлами асоціативної схожості, а вузли $\{a_{0=1}\}$ – вузлами асоціативної відмінності:

$$\{a^+\} = \{a_{=1}, a_{=0}, a_{>1}, a_{>0}\};\ \{a^-\} = \{a_{0=1}\}.$$

2. Якщо коефіцієнт $n_\Sigma > 1/2\ n$, то пропонується виконати наступні дії. Об'єднаємо множини $\{a_{>1}\}$ і $\{a_{>0}\}$ в множину $\{a_?\}$: $\{a_?\} = \{a_{>1}\} \cup \{a_{>0}\}$). З $\{a_?\}$ необхідно вибрати вузли, які б характеризували асоціативну схожість, а всі інші вузли класифікувати як вузли асоціативної відмінності, тобто необхідно «зменшити асоціативну схожість». Для

визначення вузлів, що будуть характеризувати схожість, пропонується використовувати загальне припущення про те, що чим частіше ознаки зустрічаються спільно, тим вище асоціативна схожість між ними, і вибрати в якості $a^+$ вузли з максимальними значеннями коефіцієнтів. Позначимо ці вузли як вузли «можливої асоціативної схожості» ($a^+_?$). Кількість даних вузлів є: $n^+_? = |\{a^+_?\}| = 1/2\ n - n_{=1} - n_{=0}$. Інші вузли визначимо як вузли «можливої асоціативної відмінності» ($a^-_?$). Кількість даних вузлів є: $n^-_? = |\{a^-_?\}| = 1/2\ n$. До множини $\{a^-\}$ будуть відноситись також всі вузли $\{a_{0=1}\}$:

$$\{a^+\} = \{a^+_{=1},\ a^+_{=0},\ a^+_?\};\ \{a^-\} = \{a_{0=1},\ a^-_?\}.$$

При розгляді вузлів арності 5 і більше було відзначено, що серед вузлів $a_{>1}$ і $a_{>0}$ є такі, в яких за кількістю одиниць (нулів) одні вузли більш підходять на відбір в множину $\{a^+\}$, ніж інші. Наприклад, вузол $a_{11101}$ має більше 1, ніж $a_{11100}$ і, отже, в більшій мірі характеризує «асоціативну схожість». Введемо додаткове обмеження на обчислення чисельної міри для асоціацій для коефіцієнтів арності 5 і більше при відборі коефіцієнтів в множину $a^+_?$. Пропонується наступним чином визначити коефіцієнти $\Delta_1$ і $\Delta_0$. Коефіцієнт $\Delta_1$ характеризує різницю між кількістю одиниць і середньою кількістю знаків (1 і 0) в коді вузла і дорівнює: $\Delta_1 = \text{abs}\ ((1/2\,c(a)) - c_1(a)$, де abs (x) – функція визначення абсолютного значення числа x, $c(a)$ – кількість знаків в коді вузла, $c_1(a)$ – кількість одиниць в коді вузла. Коефіцієнт $\Delta_0$ характеризує різницю між кількістю нулів і середньою кількістю знаків в коді вузла і дорівнює: $\Delta_0 = \text{abs}\ ((1/2\,c(a)) - c_0(a)$, де $c_0(a)$ – кількість нулів в коді вузла. Через те, що коефіцієнти дорівнюють один одному ($\Delta_1 = \Delta_0$), пропонується використовувати для них одне загальне позначення $\Delta$.

Розглянемо коефіцієнт $a_{11101}$. Значення $\Delta$ коефіцієнтів для нього дорівнюють: $c(a) = c(a_{11101})$; $c_1(a) = 4$; $c_0(a) = 1$; $\Delta_1 = \text{abs}\ (5/2 - 4) = 1.5$; $\Delta_0 = \text{abs}\ (5/2 - 1) = 1.5$. Введення параметра $\Delta$ дозволяє уточнити вузли $a_{>1}$ і $a_{>0}$ наступним чином: $a^{\Delta}_{>1}$ і $a^{\Delta}_{>0}$. Пропонується вважати, що чим більше значення $\Delta$, тобто більша різниця між кількістю одиниць і нулів у коді, тим вище асоціативна схожість між ознаками. Таким чином, в

множину $a_?^+$ в першу чергу повинні відбиратися вузли з більшими значеннями Δ. Позначимо мінімальне значення Δ як $\Delta^{min}$. Для чотирьохвимірної асоціації $\Delta \in \{0, 1\}$, для п'ятивимірної асоціації $\Delta \in \{0.5, 1.5\}$ і так далі. Для врахування різниці в кількості одиниць і нулів в коді пропонується в першу чергу вибирати коефіцієнти з $\Delta^* > \Delta^{min}$: $\{a_?^+\} = \{a_{>1}^{\Delta^*}\} \cup \{a_{>0}^{\Delta^*}\}$. Розглянемо два приклади застосування методу, що пропонується.

**Приклад 1**. Визначення чисельної міри для квантора бінарної асоціації з коефіцієнтами: $\{a\} = \{a_{11}, a_{00}, a_{10}, a_{01}\} = \{5, 4, 2, 4\}$ (n = 4). Множина $\{a_\Sigma\} = \{a_{=1}\} \cup \{a_{=0}\} \cup \{a_{>1}\} \cup \{a_{>0}\} = \{a_{11}\} \cup \{a_{00}\} \cup \{\varnothing\} \cup \{\varnothing\}$, де ∅ – символ порожньої множини. Через те, що коефіцієнт $n_\Sigma = 2 = 1/2\ n$, то множина вузлів $\{a_{=1}\}$, $\{a_{=0}\}$, $\{a_{>1}\}$, $\{a_{>0}\}$ є вузлами асоціативної схожості ($\{a^+\} = \{a_{11}, a_{00}\}$), а множина вузлів $\{a_{0=1}\}$ – вузлами асоціативної відмінності ($\{a^-\} = \{a_{10}, a_{01}\}$).

Чисельна міра для квантора бінарної асоціації визначається наступним чином:

$$\gamma_{\approx_2} = \prod_{i=1}^{2} a_i^- \Big/ \prod_{i=1}^{2} a_i^+ = \frac{a_{10} \cdot a_{01}}{a_{11} \cdot a_{00}} = \frac{2 \cdot 4}{5 \cdot 4} = 0.4.$$

**Приклад 2**. Визначення чисельної міри для квантора тривимірної асоціації з коефіцієнтами: $\{a\} = \{a_{111}, a_{110}, a_{101}, a_{011}, a_{100}, a_{010}, a_{001}, a_{000}\} = \{2, 5, 8, 2, 4, 1, 9, 3\}$ (n = 8). Множина $\{a_\Sigma\} = \{a_{=1}\} \cup \{a_{=0}\} \cup \{a_{>1}\} \cup \{a_{>0}\} = \{a_{111}\} \cup \{a_{000}\} \cup \{a_{110}, a_{101}, a_{011}\} \cup \{a_{100}, a_{010}, a_{001}\}$. Коефіцієнт $n_\Sigma = 8 > 1/2\ n$. Для «зменшення асоціативної схожості» об'єднаємо $\{a_{>1}\}$ і $\{a_{>0}\}$ в множину $\{a_?\}$: $\{a_?\} = \{a_{>1}\} \cup \{a_{>0}\} = \{a_{110}, a_{101}, a_{011}, a_{100}, a_{010}, a_{001}\} = \{5, 8, 2, 4, 1, 9\}$. Виберемо з $\{a_?\}$ вузли «можливої асоціативної схожості»: $\{a_?^+\} = \{9, 8\} = \{a_{001}, a_{101}\}$ як вузли з максимальними значеннями коефіцієнтів. Кількість таких вузлів дорівнює $n_?^+ = |\{a_?^+\}| = 1/2\ n - n_{=1} - n_{=0} = 4 - 1 - 1 = 2$. Решту елементів $\{a_?\}$ визначимо як вузли «можливої асоціативної відмінності»: $\{a_?^-\} = \{a_{110}, a_{011}, a_{100}, a_{010}\}$. Отже маємо: $\{a^+\} = \{a_{=1}, a_{=0}, a_?^+\} = \{a_{111}, a_{000}, a_{001}, a_{101}\}$; $\{a^-\} = \{a_?^-\} = \{a_{110}, a_{011}, a_{100}, a_{010}\}$.

Чисельна міра для квантора тривимірної асоціації визначається наступним чином:

$$\gamma_{\approx_3} = \prod_{i=1}^{3} a_i^- \Big/ \prod_{i=1}^{3} a_i^+ = \frac{a_{110} \cdot a_{011} \cdot a_{100} \cdot a_{010}}{a_{111} \cdot a_{000} \cdot a_{001} \cdot a_{101}} = \frac{5 \cdot 2 \cdot 4 \cdot 1}{2 \cdot 3 \cdot 9 \cdot 8} = 0.009 .$$

**Висновки**. В роботі пропонується метод автоматичного утворення гіпотез з квантором багатовимірної асоціації на основі атрибутованого бінарного дерева без урахування логічної невизначеності вхідних даних. Наведений підхід використання бінарного дерева для класифікації коефіцієнтів вузлів асоціативної схожості і відмінності, множина класифікаційних атрибутів і зв'язаних з ними коефіцієнтів, запропоновано вперше. Показано, що відомий метод визначення чисельної міри для квантора бінарної асоціації [1, 3] може бути реалізований як окремий випадок чисельної міри квантора багатовимірної асоціації. Запропонований метод дозволяє одержувати чисельну оцінку міри багатовимірної асоціації і є основою для проведення подальших досліджень з розробки математичного забезпечення інтелектуальних систем підтримки прийняття рішень.

## ЛІТЕРАТУРА


1. *Гаек П., Гавранек Т. Автоматическое образование гипотез: математические основы общей теории. – М.: Наука, 1984. – 280 с.*
2. *Edwards A.W.F. The measure of association in 2x2 table // Journal of the Royal Statistical Society. – Ser. A29. – P. 109 – 114.*
3. *Феклістов А.О., Лазарєва О.Я., Феклістов О.О. Метод формалізації операторів логічної невизначеності в інтелектуальних системах підтримки прийняття рішень // Системи обробки інформації. – Х.: ХВУ. – 2003. – Вип. 6. – С. 22 – 26.*
4. *Amir A., Feldman R., Kashi R. A new and versatile method for association generation // Information systems. – 1997. – Vol. 22, No. 6/7. – P. 333 – 347.*
5. *Ситников Д.Э., Титова Е.В. Метод поиска обобщенных ассоциативных зависимостей между дискретными признаками // Системи обробки інформації. – Х.: НАНУ, ПАНМ, ХВУ. – 2002. – Вип. 6(22). – С. 194 – 202.*
6. *Мартин Дж. Организация баз данных в вычислительных системах. – М.: Мир, 1980. – 662 с.*


*Надійшла 7.10.2003*

***ПЕТРОВ Вадим Лук'янович**, канд. техн. наук, доцент, професор кафедри ХВУ. В 1978 закінчив ВІРТА ім. Говорова. Область наукових інтересів – інформаційна боротьба.*

***ФЕКЛІСТОВ Андрій Олександрович**, канд. техн. наук, старший науковий співробітник, начальник лабораторії НЦ Військ ППО. В 1993 році закінчив ХВВКІУ РВ. Область наукових інтересів – штучний інтелект.*

***ФЕКЛІСТОВ Олексій Олександрович**, ад'юнкт кафедри ХВУ. В 1998 році закінчив ХВУ. Область наукових інтересів – системи обробки інформації.*