УДК 624.074.046

Білик А.С.

Аналіз ієрархічно структурованих інформаційних масивів в умовах невизначеності

В статті розглядається критичний вплив невизначеності, пов'язаної з неповністю вихідних даних на функцію належності структурованого об'єкта масиву порівняння. Показана залежність функції належності від кількості невизначених параметрів.

Постановка проблеми. Сучасний розвиток науки пов'язаний з необхідністю прийняття рішень з урахуванням все більшої кількості чинників, що на них впливають.

Одним із шляхів прийняття ефективних рішень є вибір з масиву інформації або бази знань про існуючі плани рішень чи моделей об'єктів тих, що найбільш відповідають наявному об'єкту або ситуації, що склалася. Проте досить часто, особливо при аналізі динамічних систем та даних, отриманих емпіричним шляхом, вихідна інформація для прийняття рішень може бути неповною, фрагментованою або опосередкованою, причому строки прийняття рішення не дозволяють встановити невизначені параметри більш точно або це встановлення взагалі не уявляється можливим.

Разом з тим, прийняття зважених рішень або вибір операцій підтримки цілеспрямованих дій на основі отриманих результатів обробки даних повинно враховувати усі особливості та типи даних, що містяться в умовах поставлених завдань. Інакше кажучи, при прийнятті рішень в умовах недостатньої вихідної інформації потрібно вміти оцінити критичний вплив, який може спричинити невизначеність на область найбільш прийнятних рішень, визначених на основі обробки висхідної інформації.

Аналіз останніх досліджень. Прийняття рішень в умовах невизначеності в широкому розумінні цього поняття має розгалужене прикладне підґрунтя [4,7]. Останнім часом широке застосування метод обробки інформаційних масивів шляхом створення інтелектуальних систем, що використовують спеціалізовані алгоритми на основі м'яких обчислень і дозволяють формалізувати інформацію або знання про об'єкт дослідження та невизначеність, що міститься у його описанні. Як математичний інструмент інтелектуальні системи використовують теоретико-множинний підхід, теорію нечітких множин, математичну логіку та інші розділи дискретної математики [6].

Задачі, пов'язані з розпізнаванням та класифікацією структурованих об'єктів при невизначеності, пов'язаної з розпливчастими умовами, описанні об'єкта вектором належностей або полем імовірностей широко представлені і розглянуті: [2,3,5] та ін. Наближений підхід до визначення впливу невизначеності, пов'язаної з відсутністю даних, окреслений у роботі [1].

Постановка завдання. Для наочності розглядатимемо множини даних, що містять певні характерні ознаки деяких явищ або об'єктів, тобто їх *прояви*. Прояви у множинах даних розподілені по визначеним групам та підгрупам, які підпорядковані ієрархічній структурі, і утворюють *класифікатор* даних.

Нехай ми маємо деяку множину даних, в якій містяться дані що описують певне явище або об'єкт дослідження A_K, і розподілені згідно обраного класифікатору по групам розподілу: $A_K{}^T\{a_1; a_2; a_3; ...; a_X\}$. X - кількість груп розподілу. Число проявів у кожній з груп розподілу довільне і скінчене:

$$a_1{}^T\{a_{11}; a_{12}; a_{13}...a_{1\alpha_1}\};$$
$$a_2{}^T\{a_{21}; a_{22}; a_{23}...a_{2\alpha_2}\};$$
$$a_3{}^T\{a_{31}; a_{32}; a_{33}...a_{3\alpha_3}\};$$
$$...;$$
$$a_X{}^T\{a_{X1}; a_{X2}; a_{X3}...a_{X\alpha_X}\}.$$

При матричному представленні, кожен з елементів виду a_{ij} являє собою одномісну компонент підмножину, що може містити, або не містити прояв явища A_K: $A_K\{a_{ij}\}$.

Розглянемо тепер іншу множину даних, що також описує певне явище або об'єкт G_N, причому всі його прояви є чітко визначені на основі багаторазових прямих спостережень, вимірювань, експериментів та інших достовірних даних, тобто є детермінованими. Прояви в G_N розподілені довільним чином, але відносно тих самих груп розподілу: ${G_N}^T\{a_1;a_2;a_3;...;a_X\}$, $G_N\{a_{ij}\}$. При дослідженні множина явища або об'єкта $A_K\{a_{ij}\}$ підлягає класифікації відносно множин виду $G_N\{a_{ij}\}$. Інакше кажучи, кожна з множин $G_N\{a_{ij}\}$ є гіпотезою-множиною при розпізнаванні або *ототожненні* явища-множини $A_K\{a_{ij}\}$. Об'єднання сукупності множин $G=\bigcup_{N=1}^{L}G_N$ де L – кількість гіпотез, утворює *основний масив* порівняння.

Уведемо невизначеність. Припустимо, що деякі з груп розподілу $\{a_1;a_2;a_3;...;a_X\}$ множини A_K містять невідоме число проявів, причому число одномісних компонент-множин у кожній з таких груп та саме число груп є відомим. Відповідно, найменшою структурною одиницею невизначеності для будь-якої множини буде група невизначеності a_{Ui}.

$A_{KU}=\bigcup_{i=1}^{\theta_K}a_{Ui}$ – підмножина, що містить усі групи невизначених проявів для даного явища, θ_K - кількість невизначених груп у A_K.

Задача полягає у тому, щоб порівняти послідовно множину проявів явища або об'єкта A_K з кожною із множин основного масиву порівняння $G=\{G_1;G_2;G_3;...;G_L\}$ і обрати з цього масиву множини, що найбільше відповідають A_K, урахувавши при цьому вплив невизначеності.

Виклад основного матеріалу дослідження. У даній роботі розглядатимемо випадок присутності невизначеності, коли не відомі ані імовірнісні характеристики невідомих проявів, ані вектори належностей до кожної з гіпотез основного масиву. Вагові характеристики для кожного типу проявів у групах вважаються однаковими.

Функцію належності [5] множини окремого, на початковому етапі досліджень не ототожненого, явища або об'єкту $A_K\{a_{ij}\}$ множині гіпотези $G_N\{a_{ij}\}$, що характеризує її пріоритетність при застосуванні для ототожнення $A_K\{a_{ij}\}$ названо *застосовністю* даної гіпотези [1]:

$$\mu_{G_N}(A_K)=P_N=\begin{cases}(A'_K>G'_N)\to(P_N=S'_N/G'_N),\\(A'_K\le G'_N)\to(P_N=S'_N/A'_K);\end{cases}\qquad P_N=\begin{cases}(0,1],A_K\cap G_N=\overline{\varnothing},\\0,A_K\cap G_N=\varnothing.\end{cases}\qquad(1)$$

Повне урахування невизначеності будемо реалізовувати через пошук максимальних та мінімальних значень, які може приймати застосовність при всіх можливих випадках розкриття невідомих проявів для фіксованої кількості груп невизначеності.

1. Знайдемо максимальну застосовність, що може виникати при урахуванні невизначеності. Для цього розкладемо: $A_K=A_{K\overline{U}}\cup A_{KU}$; $G_N=G_{N\overline{U}}\cup G_{NU}$;

G_{NU} – підмножина гіпотези G_N, відповідна за групами класифікаційного розподілу підмножині явища A_{KU}. Сума проявів по групам для множин A'_K та G'_N, визначається як: $G'_N=\sum_{n=1}^{X}g_n$; $A'_K=\sum_{n=1}^{X}t_n$ де g_n і t_n є числа компонентів по групах розподілу у відповідних множинах. Чисельно можна записати: $A'_K=A'_{K\overline{U}}+A'_{KU}$; $G'_N=G'_{N\overline{U}}+G'_{NU}$.

Для визначення впливу невизначеності на максимальне значення застосовності уведемо підмножину m_U – варіант заповнення підмножини A_{KU} проявами, що містить найбільшу їхню кількість з усіх реально можливих комбінацій проявів в усіх класифікаційних групах, що містять невизначеність.

m_U визначається через відповідні класифікаційні групи розподілу з максимальним заповненням виходячи з властивостей класифікаційних груп розподілу проявів та реальних чинників, що впливають на A_K та G_N:

$m_{Ui} = \max\left(G_{1U}\{a_{Ui}\}; G_{2U}\{a_{Ui}\}; \cdots; G_{LU}\{a_{Ui}\}\right); m_U = \bigcup_{i=1}^{\theta_K} m_{Ui}$. Чисельно можна записати:

$m'_{Ui} = \max\left(G_{1U}\{g_{Ui}\}; G_{2U}\{g_{Ui}\}; \cdots; G_{LU}\{g_{Ui}\}\right); m'_U = \sum_{i=1}^{\theta_K} m'_{Ui}$. Очевидно, що $m'_U \geq G'_{NU}$.

Множина спільних проявів для A_K та G_N визначається як перетин цих множин:

$S_N{}^T\{a_1; a_2; a_3; ...; a_X\} = G_N \cap A_K = G_N{}^T\{a_1; a_2; a_3; ...; a_X\} \cap A_K{}^T\{a_1; a_2; a_3; ...; a_X\}; S'_N = \sum_{n=1}^{X} s_n$;

Відповідно, множину спільних проявів можна представити як $S_N = (A_{KU} \cup A_{K\bar{U}}) \cap (G_{NU} \cup G_{N\bar{U}})$. Максимальна застосовність досягатиметься за умови максимуму перетину, тобто коли $m_U = G_{NU}$:

$S_{N\max} = (m_U \cup A_{K\bar{U}}) \cap (G_{NU} \cup G_{N\bar{U}}) = (G_{NU} \cup A_{K\bar{U}}) \cap (G_{NU} \cup G_{N\bar{U}})$;

застосувавши принцип дистрибутивності [6], отримаємо:

$S_{N\max} = (A_{K\bar{U}} \cap G_{N\bar{U}}) \cup G_{NU} = (A_{K\bar{U}} \cap G_{N\bar{U}}) \cup G_{NU} = S_{N\bar{U}} \cup G_{NU}$.

Тоді з системи (1), максимальна застосовність визначатиметься як

$$\begin{cases} (A'_{K\bar{U}} > G'_{N\bar{U}}) \rightarrow P_{N\max} = (S'_{N\bar{U}} + G'_{NU})/(A'_{K\bar{U}} + G'_{NU}), \\ (A'_{K\bar{U}} \leq G'_{N\bar{U}}) \rightarrow P_{N\max} = (S'_{N\bar{U}} + G'_{NU})/(G'_{N\bar{U}} + G'_{NU}); \end{cases} \quad (2)$$

$P_{N\max}$ - максимальне значення застосовності P_N при даній невизначеності m_U.

З системи (2) видно, що максимальне значення застосовності з урахуванням невизначеності не залежить власне ні від кількості груп невизначеності, ні від кількості одномісних компонент множин, що містяться у цих групах, а лише від кількості проявів у відповідних їм групах множини гіпотези G_{NU}. Можна також зробити висновок, що $P_{N\max} = 1$ тоді і тільки тоді, коли $A_{K\bar{U}} = G_{N\bar{U}}$.

2. Визначимо тепер мінімальну застосовність з урахуванням невизначеності.

Мінімальному випадку буде відповідати заповнення підмножини невизначеності усіма проявами з максимального варіанту заповнення, за винятком проявів, що містяться у відповідній підмножині гіпотези G_{NU}. Сукупність таких проявів можна записати через різницю множин $A_{KU} = m_U \setminus G_{NU}$. Тоді мінімальний перетин множин A_K та G_N буде:

$S_{N\min} = (A_{KU} \cup A_{K\bar{U}}) \cap (G_{NU} \cup G_{N\bar{U}}) = \left\{(m_U \setminus G_{NU}) \cup A_{K\bar{U}}\right\} \cap (G_{NU} \cup G_{N\bar{U}})$;

$S_{N\min} = \left\{(m_U \setminus G_{NU}) \cup A_{K\bar{U}}\right\} \cap G_N$.

Підставивши в (1) чисельні значення проявів у відповідних підмножинах, отримаємо:

$$\begin{cases} (A'_{K\bar{U}} + (m'_U - G'_{NU}) \leq (G'_{N\bar{U}} + G'_{NU}) \rightarrow (P_{N\min} = \dfrac{S'_{N\bar{U}}}{G'_{N\bar{U}} + G'_{NU}}), \\ (A'_{K\bar{U}} + (m'_U - G'_{NU}) > (G'_{N\bar{U}} + G'_{NU}) \rightarrow (P_{N\min} = \dfrac{S'_{N\bar{U}}}{A'_{K\bar{U}} + (m'_U - G'_{NU})}); \end{cases} \quad (3)$$

Приклад. Нехай ми маємо множину даних, що описує досліджуване явище A_1 за чотирма групами розподілу, кожна з яких містить по п’ять компонент підмножин для охарактеризування проявів. Одна з груп містить невизначеність, пов’язану із відсутністю інформації щодо проявів:

$$a_1^T\{a_{11};a_{12};a_{13};a_{14};a_{15}\};$$
$$a_2^T\{a_{21};a_{22};a_{23};a_{24};a_{25}\};$$
$$a_{U3}^T\{a_{U31};a_{U32};a_{U33};a_{U34};a_{U35}\};$$
$$a_4^T\{a_{41};a_{42};a_{43};a_{44};a_{45}\}.$$

Представимо множину A_1 у матричному вигляді, надавши групам розподілу конкретного змісту щодо проявів:

$$A_1\begin{pmatrix}0&0&1&1&0\\1&0&0&1&1\\U&U&U&U&U\\0&1&1&0&1\end{pmatrix}.$$

Аналогічно представимо множину гіпотези G_1 та відповідну множину спільних проявів S_1:

$$G_1\begin{pmatrix}0\,1\,0\,0\,0\\1\,1\,0\,1\,1\\0\,1\,1\,0\,0\\0\,1\,0\,1\,0\end{pmatrix};\quad S_1\begin{pmatrix}0&0&0&0&0\\1&0&0&1&1\\U&U&U&U&U\\0&1&0&0&0\end{pmatrix}.$$

У даному прикладі вважається, що усі прояви є сумісними між собою, тобто кожна група може містити їх будь-яку варіацію, містити всі або ж не містити жодного. Аналізуючи функцію належності $P_1=\mu_{G_1}(A_1)$ за формулою (1), ми мусили б розглянути всі можливі варіанти заповнення підмножини невизначеності.

Загалом, кількість можливих комбінацій при заповненні підмножини невизначеності залежить від числа значень, які можуть приймати змінні проявів та від числа одномісних компонент-підмножин у групах розподілу підмножини невизначеності [8]:

$$c_{KU}=n_{ij}^{\,i\alpha_i}. \tag{4}$$

У даному випадку, коли прояви описуються бінарними змінними (тобто може бути зареєстрована тільки наявність або відсутність прояву у окремій одномісній компонент-підмножині), формулу (4) можна записати як:

$$c_{1U}=2^{i\alpha_i}. \tag{5}$$

Таким чином, за формулою (5), потрібно проаналізувати кількість варіантів заповнення підмножини невизначеності, що дорівнює 32. Безперечно це є досить працемісткий процес навіть на етапі складання варіантів. До того ж слід зазначити, що реальні інформаційні масиви є набагато складнішими, ніж представлений у прикладі.

Для аналізу максимальної застосовності застосуємо формулу (2):

$$A'_{1\bar{U}}=8>G'_{1\bar{U}}=7;\ \rightarrow P_{1\max}=(S'_{1\bar{U}}+G'_{1U})/(A'_{1\bar{U}}+G'_{1U})=(4+2)/(8+2)=0,6.$$

Мінімальну застосовність визначимо за формулою (3):

$$A'_{1\bar{U}}+(m'_U-G'_{1U})=8+(5-2)=11>(G'_{1\bar{U}}+G'_{1U})=7+2=9\rightarrow$$
$$\rightarrow P_{1\min}=\frac{S'_{1\bar{U}}}{A'_{1\bar{U}}+(m'_U-G'_{1U})}=\frac{4}{8+(5-2)}=0,364.$$

Експериментально можна пересвідчитися, що множина допустимих дискретних значень, що їх прийматиме функція належності в залежності від кількості проявів у підмножині невизначеності при її розкритті, буде утворювати умовну область (рис.1), нижньою та верхньою точками якої по осі абсцис будуть $P_{1\min}$ та $P_{1\max}$ відповідно. Внаслідок дискретності та малої кількості значень проявів із проаналізованих 32 варіантів функція належності набирає лише 12 різних значень.

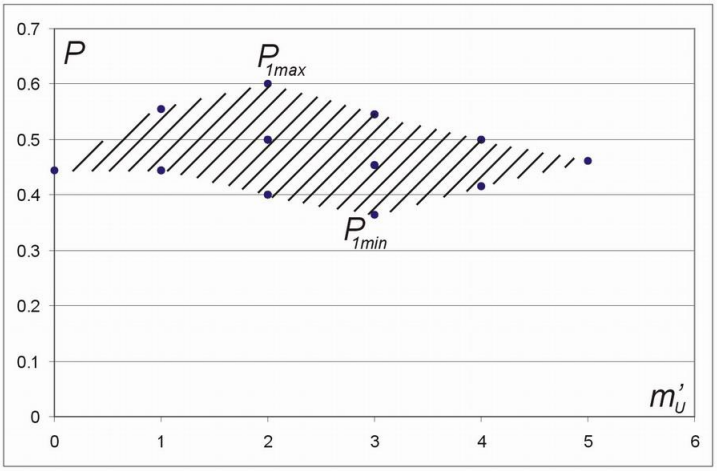

Рис.1 Область допустимих дискретних значень функції належності при усіх можливих комбінаціях проявів у підмножині невизначеності

Висновки. Всі інші можливі значення застосовності будуть знаходитися тільки у межах визначеної області, а значить, справляти передбачуваний вплив на результати прийняття рішень на основі неповних вихідних даних. Слід також зазначити, що встановлені залежності (1), (2), та (3) дійсні для даних як якісного, так і для кількісного характеру.

На основі розробленої методики створено пробну модель експертної системи, що дозволяє класифікувати невідомі літальні апарати та атмосферні явища за їх характерними проявами. Перші результати обробки тематичних інформаційних повідомлень за допомогою створеної системи показали високу чутливість до вхідних даних та значне скорочення часових та трудових витрат в порівнянні з існуючими схемами обробки повідомлень.

Перспективним є вирішення задачі урахування невизначеності, пов'язаної з неповністю інформації із залученням систем уведення вагових коефіцієнтів ознак для визначення сукупності моделей рішень або об'єктів, більш адекватних критеріям пріоритетності.

Список літератури:

1. *Білик А.С.,* Порівняння масивів якісних даних на прикладі не ототожнених явищ // Зб. наук. праць IV Міжн. наук. конф. „Політ", – К.: НАУ, 2004.
2. *Вятченин Д.А.,* Об индексации нечеткой иерархии и методах нечеткой классификации на основе понятия распределения // Мат. междунар. науч. сем. „Интеллектуальный анализ информации", – К.: НТУУ «КПИ», Просвіта, 2004. – С.69-77.
3. Обработка информации и принятие решений в условиях неопределенности: Сб. науч. ст. // АНКиргССР, Ин-т автоматики, – Фрунзе: Илим, 1980.
4. *Ходаков В.Е., Граб М.В.,* Классификация ситуаций в задаче планирования управления лесным пожаром // Зб. наук. статей конф. „Проблеми моделювання та прогнозування надзвичайних ситуацій", – К.: КНУБіА, Чорнобильінтерінформ, 2002. – С.3-5.
5. *Шлепаков Л.Н.,* Системы с базами данных по решению задач распознавания и калссификации информационных сообщений // Интеллектуализация систем обработки информ. сообщений: Сб. науч. тр., – К.: НАНУ, Ин-т матем., 1995. – С.11-38.
6. *Шарапов О.Д.,* Дискретний аналіз: навч.-метод. посібник, – К.:КНЕУ, 2002.
7. Принятие решений в условиях неопределенности: Межвуз. науч. сб., – Уфа: Уфимский гос. авиац. техн. ун-т, 2000.
8. Виленкин Н.Я., Комбинаторика, - М.: «Наука», 1969г. – 328 с.