Н.Шаховська, Р.Кравець
Національний університет "Львівська політехніка"

УДК 681.3

ЗАСТОСУВАННЯ БАГАТОЗНАЧНОЇ ЛОГІКИ ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Шаховська Н., Кравець Р., 2002

Описано контекстну схему системи інтелектуального аналізу результатів соціологічних опитувань. Запропоновано алгоритми для вирішення основних задач.

There are described the conceptual schema of intelligence analyses the social questions system.

У великій кількості предметних областей потрібно опрацьовувати нечітку інформацію, причому результат аналізу даних повністю залежить від ступеня їх повноти у системі. Типовими предметними областями появи нечіткості є соціологічна сфера (біржа праці, громадські фонди, маркетингові дослідження ринку тощо), історичні дослідження, планування господарської діяльності тощо.

У статті пропонуються алгоритми класифікації та класифікування об'єктів, інформація про які зберігається у реляційній базі даних. Також подається опис системи, призначеної для автоматизації структурування та, частково, аналізу даних.

Основними проблемами, які виникають в задачах аналізу та структурування даних, є проблеми створення класів та віднесення до них об'єктів, інформація про які щойно надійшла у базу даних. Задача створення класу розбивається на дві підзадачі:

1) побудова класифікаційних функцій, згідно з якими об'єкт класифікується як представник певного класу;
2) розбиття на класи та ідентифікація отриманих класів.

У цій статті розглянемо першу з цих підзадач.

1. Опис алгоритмів класифікації

1.1.Віднесення до класу

Вхідними даними для класифікування (віднесення до класу) є множина значень цільових атрибутів. Цільовими атрибутами назвемо атрибути, які використовуються для аналізу даних, і згідно із значеннями яких здійснюється розбиття на класи. До цільових атрибутів, у першу чергу, належать усі атрибути, які входять до множини ключів. До цільових атрибутів віднесемо усі атрибути, що входять у множину лівих частин функціональних залежностей (крім первинних ключів), а також ті атрибути, які будуть впливати на ступінь довіри до отриманого результату аналізу. Крім того, для конкретної предметної області за допомогою експертного опитування визначається додаткова підмножина атрибутів, які вважатимуться цільовими для аналізу. Наприклад,

для задачі соціологічного опитування такими атрибутами є вік, освіта, матеріальний стан тощо.

Атрибути, над якими виконуються операції агрегації та порівняння, назвемо критичними (подають результати аналізу). Критичними атрибутами є атрибути, які містять числові дані, невизначеності, подані у довільному вигляді, та праві частини функціональних залежностей. До них також належать атрибути, що містять назви класів [4].

Для спрощення задачі вважатимемо, що класи є визначимі, і їх характеристики (тобто назви та правила, за якими об'єкт вважається представником цього класу) зберігаються у базі даних.

Віднесення до класу здійснюється на основі визначення підмножини значень цільових атрибутів. Наприклад, для класу "Студент" значення цільових атрибутів мають задовольняти умови: вік (16, 23), освіта - (середня, середня професійна, незакінчена вища), матеріальний стан - (50 грн., 150 грн.).

У зв'язку з тим, що важко отримати повну інформацію про об'єкти предметної області, то можуть бути визначені не всі цільові атрибути. Тому для кожного класу визначається значення межі – величини у межах одиничного інтервалу, яка позначає мінімальний ступінь довіри до об'єкта, за яким об'єкт може бути класифікований як представник цього класу. Ступінь довіри s до об'єкта визначається як кількість цільових атрибутів із визначеними значеннями до усіх визначених цільових атрибутів цього класу (чим більше відомо про об'єкт, тим вищим буде ступінь довіри).

$$s = \sum \begin{cases} 0, & cr_i \text{ Is Null;} \\ 1, & cr_i \text{ Not Null,} \end{cases}$$

де cr_i – значення цільового атрибута cr_i.

Вважатимемо, що якщо значення атрибута визначене, то воно достовірне, тобто проблема подання неправдивої інформації у цій статті не розглядатиметься. Наприклад, для розглянутого вище класу "Студент" межа становитиме 2/3 (для того, щоб об'єкт інтерпретувався як представник класу "Студент", то не менше ніж два атрибути повинні мати визначені значення, і ці значення повинні входити у підмножину значень цільових атрибутів, визначених для цього класу). Така інтерпретація ступеня довіри до об'єкта вигідна тим, що не потребує використання коефіцієнтів важливості (значення яких може бути отримане шляхом опитування експертів), оскільки вважається, що для даного класу усі цільові атрибути є рівноважливими. Але у цій простоті є і недолік: ігноруються можливі зв'язки між значеннями цільових атрибутів.

Розглянемо питання усунення невизначеності.

Віднесення до класу можна розглядати як один із способів усунення невизначеності, адже у процесі класифікування інтелектуальним чином здійснюється заповнення порожнього значення атрибута, який містить значення назви класу. Крім того, класифікаційні правила можна вважати нечіткими (наближеними) функціональними залежностями. У базі даних підтримується нечітка функціональна залежність

$$e(X > A),$$

якщо співвідношення кортежів, на яких виконується ця функціональна залежність, до кортежів, на яких вона не виконується, не менше, ніж e, де e – значення межі пропускання, визначене на основі експертного опитування [3]. Зрозуміло, значення e – не менше значення межі класу.

Значення межі пропускання позначатимемо ступенем багатозначної логіки Лукасевича (змінюється у межах [0, 1]). Звідси випливає, що алгоритми усунення невизначеностей за допомогою функціональних залежностей можна застосувати для класифікування об'єктів. Наприклад, у базі даних існує класифікаційне правило Вік, Освіта, Матеріальний стан > Соціальна група.

Розглянемо один із способів усунення невизначених значень. Виходячи із того, що класифікаційне правило вважатимемо наближеною функціональною залежністю із визначеним ступенем довіри A, використаємо для цього метод, аналогічний до відомого методу прогонки [2]: рівність значень атрибутів у лівій частині правила зі ступенем довіри А означає і рівність значень атрибутів у правій частині.

Опишемо алгоритм застосування модифікованого методу прогонки.

Нехай у відношенні r підтримується наближена функціональна залежність

$$e(X_1, ..., X_n > A).$$

Символ ↓ позначає визначене значення, а ⊥ – його відсутність; t_i – кортеж відношення r (послідовність кортежів значення не має)

1. Якщо $\{t_1(X_1)\downarrow, ..., t_1(X_n)\downarrow\}$ і $\{t_2(X_1)\downarrow, ..., t_2(X_n)\downarrow\}$ і $\{t_1(X_1)\downarrow, ..., t_1(X_n)\downarrow = t_2(X_1)\downarrow, ..., t_2(X_n)\downarrow\}$ і $\{t_1(\mathrm{A})\downarrow\}$ і $\{t_2(\mathrm{A}) = \perp\}$, то заміняємо кожне входження ⊥ у r на $t_1(\mathrm{A})$.
2. Якщо $\{t_1(X_1)\downarrow, ..., t_1(X_n)\downarrow\}$ і {в t_2 m з n значень атрибутів - ↓, $n - m$ значень атрибутів – ⊥, $n \le m$ } і $\{e \le m/n\}$ і {по визначених значеннях $t_1(X^m)\downarrow = t_2(X^m)\downarrow\}$ і $\{t_1(A)\downarrow\}$ і $\{t_2(A) = \perp\}$ і, то заміняємо кожне входження ⊥ у r на $t_1(A)$.
3. Якщо {в t_i m_i з n значень атрибутів - ↓, $m_i \le n$} і {в t_j m_j з n значень атрибутів - ↓, $m_j \le n$} і {по визначених значеннях $t_i(X^m)\downarrow = t_2(X^m)\downarrow\}$ і {по визначених значеннях $t_j(X^m)\downarrow = t_2(X^m)\downarrow\}$ і $\{m_i/n \le m_j/n\}$ і $\{t_i(A)\downarrow\}$ і $\{t_j(A)\downarrow\}$ і $\{t_2(A) = \perp\}$, то заміняємо кожне входження ⊥ у r на $t_j(A)$.

1.2 Побудова класифікаційних функцій

Для того, щоб мати можливість класифікувати об'єкти, необхідно побудувати функції класифікації. Взагалі, у базі даних може зберігатися інформація про декілька типів класів, і для кожного типу класу є своя підмножина функцій. Одна й та ж функція може застосовуватись для визначення кількох типів класів.

Розглянемо алгоритм породження класифікаційних функцій (правил).

Правила можуть генеруватись двома способами: на основі аналізу характеристик класів; на основі існуючих правил.

1.2.1.Породження класифікаційних правил на основі аналізу характеристик класу

У разі застосування першого способу класифікаційні правила, перш за все, будуть будуватися на основі функціональних залежностей, що підтримуються у відношенні. Ступінь довіри до такого правила буде максимальним ($A = 1$).

Решту атрибутів, які входитимуть до правил, визначають на основі аналізу характеристик класів.

Послідовність кроків:

1. Кортежі відношення групуються за назвами класів.
2. Всередині групи проходить почергово групування за кожним цільовим атрибутом.
3. Якщо кількість елементів підгрупи, зокрема з порожніми значеннями, не дорівнює кількості кортежів у групі класу, то обираємо інший атрибут для перевірки та переходимо на крок 2.
4. Визначаємо значення *e* як відношення кількості кортежів з непорожнім значенням аналізованого атрибуту до кількості усіх кортежів у групі (тобто визначаємо частотну характеристику).
5. До отриманих частотних характеристик застосовуємо багатозначне "або": $u \,\&\, v = \max\{0, u + v - 1\}$
6. До класифікаційних правил як ліва частина будуть входити усі атрибути, частотні характеристики яких більші або дорівнюють значенню, отриманому на кроці 6, а сама частотна характеристика вважатиметься ступенем довіри до правила.

1.2.2.Породження класифікаційних правил на основі існуючих

У обгрунтовано використання ступенів багатозначної логіки для подання довіри до правила. За такого подання праву та ліву частини правила можна вважати дискретними і працювати з їх частинаками як з окремими елементами. Оскільки у попередньому розділі показано, що класифікаційне правило вважається наближеною функціональною залежністю, то до них можна застосувати основні аксіоми виведення [3]. Використовуючи логічні операції багатозначної логіки [1] "і" для нащадків та "або" для предків, ми отримуємо можливість генерувати нові правила на основі існуючих та автоматично визначати до них ступені довіри (які можуть бути перевірені експериментально). Звідси випливає, що у базі даних потрібно зберігати лише мінімальне покриття наближених функціональних залежносией (тобто класифікаційних правил), а усі решта можна вивести на основі їх комбінацій з використанням операцій багатозначної логіки [1] та аксіом виведення.

Приклад породження правил наведено в таблиці.

Алгоритми класифікації із застосуванням апарата багатозначної логіки були застосовані для вирішення задачі аналізу результатів соціологічних опитувань.

2. Опис задачі аналізу результатів соціологічних опитувань

Система, що розробляється для цієї задачі, призначена для аналізу вподобань осіб (респондентів). Залежно від області застосування такими вподобаннями можуть бути: історичні цінності (музейна сфера), професійні зацікавлення (біржа праці, працевлаштування), групи продуктів та їх виробники (соціологічні опитування) тощо. Вхідними даними для аналізу є: відомості про особу, системи її переваг та вподобань. Система переваг може бути подана у вигляді бальної шкали, матриць порівнянь, ієрархії цілей тощо.

Приклад генерації правил на основі існуючих

Існуючі правила	Породжені правила
Вік, Освіта $\xrightarrow{0,8}$ Соціальна група Матеріальний стан $\xrightarrow{0,4}$ Соціальна група	Вік, Освіта, Матеріальний стан $\xrightarrow{0,4}$ Соціальна група
Вік, Освіта $\xrightarrow{0,8}$ Соціальна група Освіта $\xrightarrow{0,4}$ Рівень матеріального забезпечення	Вік, Освіта $\xrightarrow{0,2}$ Рівень матеріального забезпечення, Соціальна група

Аналізуючи дані про особу, з'являється можливість віднести її до певного класу осіб, а для кожного класу встановити систему вподобань. Тоді для особи, відомості про яку ми отримали, можна визначити клас, а, отже, прогнозувати її вподобання. Крім того, на основі характеристик класу можна доповнити відсутню інформацію про особу.

Слід передбачити можливість ручного визначення класу та його властивостей (задається на початку роботи системи), а також утворення нових класів на основі аналізу даних.

Концептуальна модель

Розглянемо контекстну схему системи інтелектуального аналізу результатів соціологічних опитувань.

На рис. 1 подані чотири зовнішні сутності, умовно названі Особа, Виробник, Продукт, Класифікатори.

Сутність Особа володіє персональними даними та системою переваг про певні характеристики сутності Продукт. Остання може позначати конкретні види продукції, перелік професій, каталог історичних об'єктів тощо, які цікавлять Особу. Продукти об'єднуються у групи.

Сутність Виробник позначає суб'єкти (об'єкти), які створюють Продукт або володіють ним, та надають свої послуги для Особи. Залежно від предметної області Виробником можуть бути:

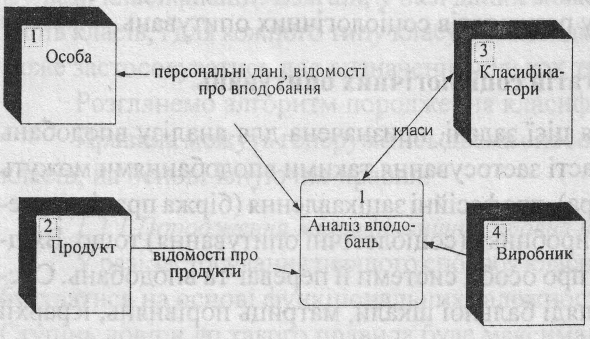

Рис. 1. Контекстна схема

- музеї, приватні колекції, віртуальні музеї (для музейної справи);
- роботодавці, посередники (для бірж праці);
- виробники продукції та послуг (для аналізу ринку споживання).

У свою чергу, Виробника цікавить рейтинг Продуктів, визначений на основі аналізу ситеми переваг класів Осіб, та самі систе-

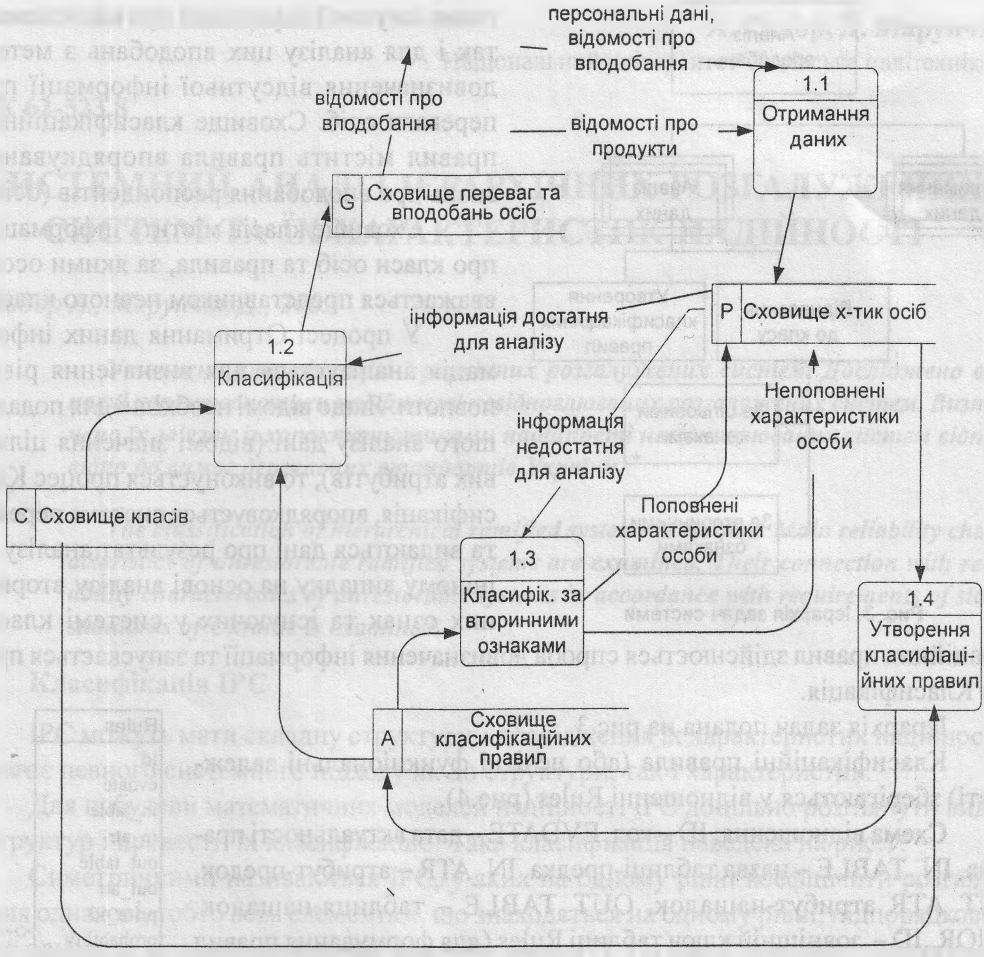

Рис. 2. Діаграма потоків даних аналізу вподобань

ми переваг, встановлені для класів.

Кожен об'єкт з сутності Особа класифікується за цільовими характеристиками (наприклад, соціальна група, вікова категорія тощо), причому для цього об'єкта визначається ступінь приналежності до класу, а для класу вказується мінімальна межа "проходження", тобто мінімальний ступінь приналежності об'єкта, за яким об'єкт ще вважається представником цього класу. Якщо для Особи відсутні дані про цільові ознаки або вподобання, то на основі правил, які містяться у сутності Класифікатори, здійснюється спроба їх довизначення шляхом аналізу вторинних характеристик (професія, рівень освіти, стать тощо). Якщо у результаті аналізу було визначено, що аналізований об'єкт "не проходить" у жодний клас, то здійснюється спроба породити нові правила.

Зовнішня сутність Класифікатори містить систему класифікаторів (правил) об'єктів Особа, на основі яких і формується система переваг особи.

Оцінки, які визначені для особи або отримані безпосередньо від неї, передаються у Сховище переваг та вподобань. Це сховище використовується як для збері-

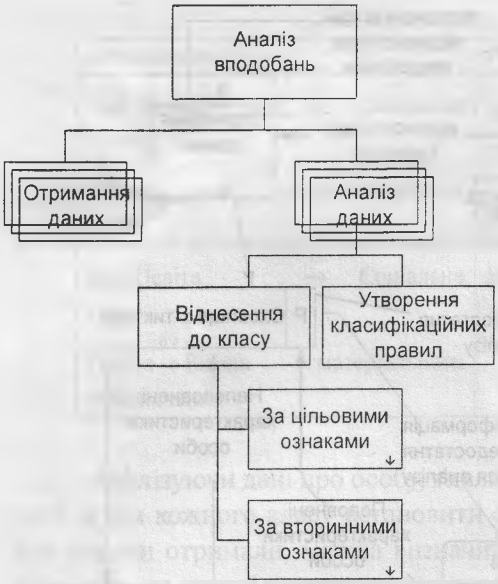

Рис. 3. Ієрархія задач системи

гання існуючої інформації про вподобання, так і для аналізу цих вподобань з метою довизначення відсутньої інформації про переваги осіб. Сховище класифікаційних правил містить правила впорядкування знань про вподобання респондентів (осіб).

Сховище класів містить інформацію про класи осіб та правила, за якими особа вважається представником певного класу.

У процесі Отримання даних інформація аналізується для визначення рівня повноти. Якщо відомі необхідні для подальшого аналізу дані (відомі значення цільових атрибутів), то виконується процес Класифікація, впорядковується система переваг та видаються дані про результат аналізу. В іншому випадку на основі аналізу вторинних ознак та існуючих у системі класифікаційних правил здійснюється спроба довизначення інформації та запускається процес Класифікація.

Ієрархія задач подана на рис.3.

Класифікаційні правила (або нечіткі функціональні залежності) зберігаються у відношенні Rules (рис.4).

Схема відношення: ID – код, EVDATE – дата актуальності правила, IN_TABLE – назва таблиці-предка, IN_ATR – атрибут-предок, OUT_ATR атрибут-нащадок, OUT_TABLE – таблиця-нащадок, PRIOR_ID – зовнішній ключ таблиці Rules (для формування правил зі складеними частинами предків чи нащадків), PROBABILITY – довіра до правила.

Модифікований метод прогонки почергово перебирає усі правила з відношення Rules та застосовує його до кортежів відношень, вказаних у відповідному кортежі по правилу, що застосовується.

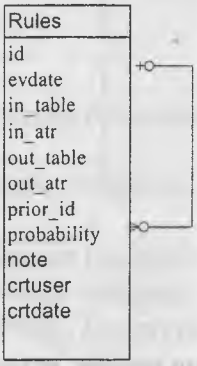

Рис. 4. Схема відношення Rules

Підсумки

1. Застосування чисел багатозначної логіки для подання ступеня довіри до правил дозволило зберігати у базі даних лише мінімальне покриття класифікаційних правил.

2. Усунення невизначеностей у базі даних можна вважати класифікуванням за певними ознаками.

1. Panti, G. Multi-valued logics, in: D. Gabbay, P. Smets (eds.) Handbook of Defeasible Reasoning and Uncertainty Management Systems. vol. 1: P. Smets (ed.) Quantified Representation of Uncertainty and Imprecision. Kluwer Acad. Publ., Dordrecht. - 1998. - P. 25-74 .

2. Д.Мейер Теория реляционных баз данных: Пер. с англ.- М.: Мир, 1987. - 608 с., ил.

3. Huhtala Y., Karkainen J. Tane: An Efficient Algoritm for discovering Functional and Approximate Dependencies// The Computer Journal. 1999. - Vol. 42. - № 2.

4. Шаховська Застосування багатозначної логіки у базах даних