УДК 004.942:517.958

Ю. Нікольський
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

МОДЕЛЬ ПРОЦЕСУ АНАЛІЗУ ДАНИХ

Описано процес аналізу даних у вигляді формальної математичної моделі. Особливістю такого підходу є те, що усі дії із перетворення даних описано у вигляді спеціальних функцій. Такий підхід дає змогу розглядати різні процеси аналізу даних на основі однакових поглядів на склад, структуру, послідовність дій та зміст.

The data analysis process as a formal mathematical model is offered. The features of such approach describe all operations of data converting as the special functions. Such approach allows to examine the different processes of data analysis as identical in composition, structure, actions' sequence and maintenance.

Вступ

Сучасний етап розвитку інформаційних технологій характеризується переходом від екстенсивного до інтенсивного їх використання. Екстенсивне використання полягало в накопиченні та структуризації даних, необхідних для експлуатації інформаційних систем. Це призвело до появи великих і надвеликих баз даних, а також спеціальних об'єктів комплексного зберігання – сховищ даних. Інтенсивне використання ґрунтується на складнішому, тоншому та глибшому опрацюванні даних без істотного збільшення їх обсягів. Це дає змогу вирішувати завдання абсолютно іншого рівня складності, які характерні для експертних систем і систем прийняття рішень. Інтенсивне використання ґрунтується на використанні нових спеціальних підходів, моделей і алгоритмів.

Метою аналізу даних є виявлення прихованих правил та залежностей у великих масивах даних. Людський розум не пристосований до сприйняття великих масивів різнорідної інформації. Людина спроможна виявляти незначну кількість взаємозв’язків лише у невеликих обсягах даних. За допомогою традиційної математичної статистики, яка довший час претендувала на роль основного інструмента аналізу даних, також часто не можна отримати прийнятні розв’язки складних реальних задач, оскільки необхідно оперувати середніми характеристиками вибірок, які часто є фіктивними величинами. Мета процесу аналізу даних полягає у поясненні великої кількості розрізнених даних, які отримано з деякої предметної області [1].

Постановка задачі дослідження

Найбільше потребує досліджень, пов’язаних із аналізом даних, великий бізнес. Саме тут накопичені великі масиви даних та формулюються завдання знаходження змісту цих даних. Метою відповідних досліджень є підвищення конкурентоспроможності та прибутковості компаній. Замовники намагаються не лише краще зрозуміти свій бізнес на основі зібраних даних, але й отримати нові знання про предметну область та на основі залежностей, які приховані в даних, вирішувати свої проблеми новими, кращими та нестандартними способами.

Для отримання корисних знань з даних потрібно усвідомити, що існує певний загальний підхід, за яким можна досягти успіху. З іншого боку, велика кількість алгоритмів аналізу даних ще не є достатньою умовою такого успіху. Процес аналізу визначений послідовністю кроків для знаходження шаблонів даних; він є ітераційним та вимагає багатократного, циклічного повторення певних кроків. Тому для формалізації процесу аналізу даних введено поняття моделі процесу [1].

Процес аналізу ще називають процесом відкриття знань у базах даних та пов’язаним із ним пошуком нових знань у певній прикладній галузі. Також його називають процесом ідентифікації

реальних, нових, потенційно корисних зрозумілих шаблонів у даних. Цей процес узагальнюється до даних, які не обов'язково зберігаються у базах даних, хоча часто підкреслюється, що саме бази даних є первинним джерелом даних.

Він складається з багатьох кроків, кожний з яких виконує часткову задачу формування залежностей та полягає у застосуванні певного математичного методу.

Знаходження залежностей у даних є цілісним процесом, який передбачає кроки підготовки даних, їх збереження та доступ до них, із застосовуванням алгоритмів аналізу великих обсягів даних, інтерпретування та візуалізації результатів та взаємодії людини та машини. Сучасні вимоги до процесу аналізу вимагають формалізації та узгодження всіх цих кроків.

Аналіз останніх досліджень

На початку 90-х років XX ст. було запропоновано кілька різних моделей процесів аналізу даних. Процес розглядають як послідовність кількох кроків. Кожний наступний крок, який ініціалізують після успішного завершення попереднього, є складовою ланцюжка дій від розуміння предметної області та даних через підготовку даних та аналіз до оцінювання, розуміння та застосування отриманих результатів.

Весь процес аналізу складається з кількох кроків, або фаз, які можуть повторюватись. З джерела даних отримують "сирі", неопрацьовані дані, після чого їх повністю або їхню частину відбирають для подальшого опрацювання. Ці дані опрацьовують та перетворюють, інтерпретують та перевіряють, аж поки оброблять алгоритмами виявлення залежностей та отримають нові залежності, які приховані у даних.

Вхідні дані можуть бути подані у різних форматах, зокрема, числових та номінальних, які зберігаються у базах даних, образах, відео, напівструктурованих даних, таких як XML, HTML тощо. Результатом є нові знання, подані у формі правил, шаблонів, класифікаційних моделей, асоціацій, трендів, результатів статистичного аналізу тощо.

В останні роки набули великої популярності дослідження з метою отримання якомога загальнішої модель процесу [2]. Першу базову модель процесу запропонували U. Fayyad та ін. [3]. Цю модель надалі було вдосконалено та модифіковано іншими дослідниками. Основні відмінності між моделями полягають у кількості та специфіці кроків.

Розглядають три основні типи моделей процесів аналізу: дослідницькі, промислові та гібридні. Прикладом дослідницької моделі є модель процесу аналізу U. Fayyad та ін. [3], яку було покладено в основу багатьох інших моделей.

Типовою промисловою моделлю є п'ятикрокова модель P. Cabena та ін. [4], розроблена спільно із фірмою IBM, та шестикрокова модель CRISP-DM [5], яку поширює консорціум європейських компаній. Останню вважають провідною промисловою моделлю. Модель процесу аналізу CRISP-DM (CRoss-Industry Standard Process for Data Mining) [5] вперше задекларовано в кінці 90-х років XX ст. чотирма компаніями, кожна з яких мала свою спеціалізацію у проекті.

Знаходження залежностей є пошуком нового знання про предметну область [6]. Процес аналізу складається з багатьох кроків, які виконуються послідовно. Наступний крок, яким ініціюють у разі успішного завершення попереднього кроку, вимагає результатів попереднього кроку як вхідних даних.

Процес аналізу з усіма його складовими – від розуміння предметної області, підготовки даних, їх аналізу до оцінювання, розуміння та застосування знайдених залежностей – істотно ітеративним та містить багато зворотних зв'язків. Створення моделей процесу вимагає формалізації кроків для скорочення вартості та часу, а також покращання розуміння даних, оцінювання результатів та розширення сфер застосування таких процесів. Формальні моделі дають змогу зменшити залежність процесу аналізу від специфіки предметної області та розширити сферу застосувань, інструментів та потенційних користувачів.

До найпоширеніших моделей аналізу належать п'ять представників, які визначають певний стандарт у створенні систем такого типу. Це дев'ятикрокова модель U. Fayyad та ін. [3],

восьмикрокова модель S. Anand та A. Buchner [7,8], шестикрокова модель K. Cios та ін. [9,10,11], п'ятикрокова модель P.Cabena [4] та модель CRISP-DM [5].

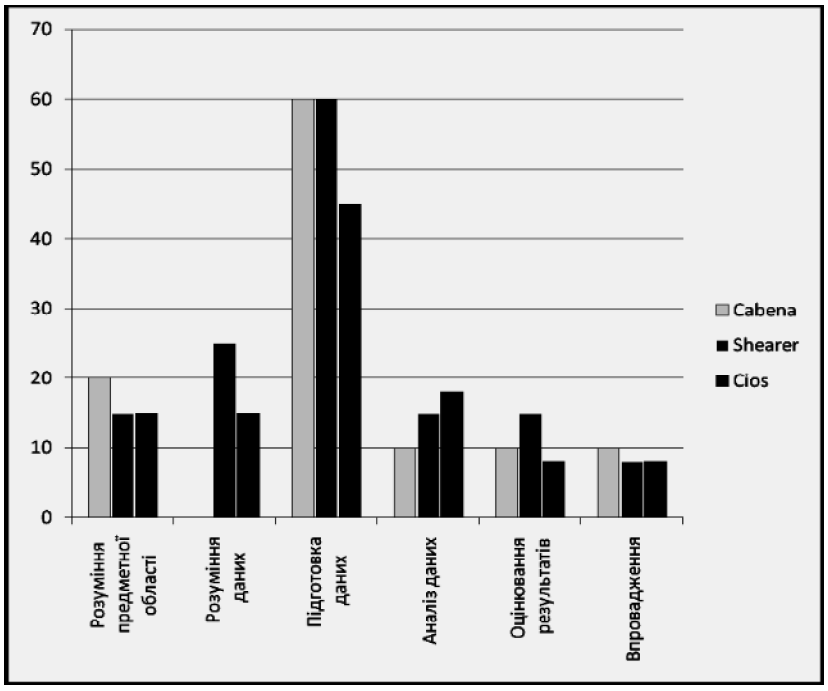
Дані, наведені у таблиці, допомагають здійснити порівняльну характеристику цих моделей. Кожна з моделей має сильні та слабкі сторони, визначені сферою застосування та специфікою розв'язуваних задач. Дуже важливою особливістю процесів аналізу є відносний час, якого вимагає реалізація кожного кроку.

Порівняння моделей аналізу даних

	Приклади моделей аналізу				
	U. Fayyad [3]	**Amand & Buchner [7]**	**K. Cios [9]**	**Cabena [4]**	**CRISP-DM [5]**
Область застосування	Дослідницька	Дослідницька	Гібридна	Промислова	Промислова
Кількість кроків	9	8	6	5	6
Кроки	Дослідження та розуміння проблемної області	Опис людських ресурсів	Розуміння проблемної області	Визначення цілей бізнесу	Розуміння бізнес-процесу
	Створення цільової множини даних	Специфікація проблеми	Розуміння даних	Підготовка даних	Розуміння даних
	Очищення та попереднє опрацювання даних	Попереднє дослідження даних	Підготовка даних	Аналіз даних	Підготовка даних
	Скорочення та проектування даних	Виявлення знань предметної області	Аналіз даних	Аналіз результатів	Моделювання
	Вибір цілей аналізу даних	Ідентифікація методології	Оцінювання отриманих знань	Засвоєння знань	Оцінювання
	Вибір алгоритму аналізу даних	Попереднє опрацювання даних	Використання отриманих знань		Застосуваня
	Аналіз даних	Відкриття шаблонів			
	Інтерпретація виявлених шаблонів	Опрацювання отриманих знань			
	Консолідація отриманих знань				
Спеціалізоване програмне забезпечення	Комерційна система MineSet	Невідоме	Невідоме	Невідоме	Комерційна система Clementine®
Галузі застосування	Медицина, проектування, промисловість, електронний бізнес, розроблення програмного забезпечення	Маркетинг, торгівля	Медицина, розроблення програмного забезпечення	Маркетинг, торгівля	Медицина, проектування, промисловість, розроблення програмного забезпечення, маркетинг, торгівля

Усвідомлення цього факту та оцінювання зусиль на виконання кожного кроку можуть стати вирішальними для планування усього комплексу робіт, які треба виконати для аналізу даних. Існують різні оцінки відносної кількості часу, необхідного для цього.

На рисунку показано діаграму, яка дає змогу порівняти такі оцінки. Наведені тут оцінки є експертними та мають відносний характер. Вони показують відносні середні витрати часу, у зв'язку з чим сума наведених значень може не дорівнювати 100 %.



Порівняння часу (у %) на виконання кроків процесу аналізу даних [12, 13]

Специфіка оцінок залежить від багатьох чинників, зокрема, знання предметної області, кваліфікації фахівців, складності проблеми тощо. Загальна ідея наведених оцінок є підтвердженням того, що підготовка даних є найістотнішим етапом процесу аналізу з погляду витрат часу.

Цілі статті

Як показав аналіз останніх досліджень, існування різних процесів аналізу даних створює проблеми вибору моделі процесу, адекватної предметній області та задачам аналізу даних, для розв'язання яких цей процес створюється.

Аналіз даних є складною, багатокроковою процедурою, на кожному кроці якої побудовано спеціальне відображення та застосовано математичні методи опрацювання даних. Формування процесу аналізу пов'язано із необхідністю узгодження цих методів за множинами аргументів та значень відповідних операторів.

Таке узгодження можна виконати лише у тому випадку, якщо кожний крок процесу буде подано оператором, а узгодження буде встановлене на основі відповідності результатів одного кроку та аргументів іншого. Це дасть змогу описати різні процеси аналізу на основі загальних міркувань. Запропоновано підхід до виконання формального опису процесу аналізу даних, який враховує таку особливість даних, як часткова невизначеність та надлишковість. Усунення таких особливостей даних є важливою складовою розв'язування задач аналізу, оскільки воно дає змогу зменшити розмірність задачі або усунути неістотні дані, які не впливають на якість розв'язку.

Основний матеріал

У зв'язку із побудовою формальної моделі процесу аналізу будемо розуміти його як певну послідовності відображень та відповідних обчислювальних процедур. Ці процедури мають на меті знаходження прихованої залежності в таблиці даних як функції визначеного вигляду – явного або параметричного – та обчислення її у всіх точках області визначення. Процес аналізу даних пропонується розглядати як сукупність таких підпроцесів:

1. Підпроцес формування опису предметної області.
2. Підпроцес попереднього опрацювання даних.
3. Підпроцес виявлення залежностей у даних.
4. Підпроцес оцінювання та інтерпретації.

В основу побудови та дослідження математичної моделі процесу аналізу покладено розуміння її як задачі виявлення залежностей в сенсі задачі наближення функції за таблицею її значень. Опис предметної області Π дає змогу виділити в ній певну множину X, $X \subset \Pi$ об'єктів, кожний з яких x, $x \in X$ заданий множиною значень A його властивостей.

Таблицю даних з описами множини об'єктів предметної області надалі буде використано як основу для аналізу. Її вважатимемо табличним поданням невідомої функції $F(x)$ багатьох змінних, значення d якої відомі у точці x, $x \in X$ m-вимірного простору. Розмірність області визначення функції задана кількістю ознак m, за якими описано об'єкти предметної області Π. Кожний з об'єктів предметної області Π може мати різну кількість значень властивостей. Задача аналізу полягає у відновленні функції та обчисленні значення цієї функції у кожній точці Π.

Модель процесу аналізу даних розглядатимемо його як послідовність його складових підпроцесів такого загального вигляду

$$M = (M_1, M_2, M_3, M_4). \tag{1}$$

Під математичною моделлю процесу аналізу даних розуміємо функцію $F: X \to D$ множини об'єктів $X = \{x_1, x_2, ..., x_n\}$ певної предметної області із значеннями з множини ознак $d \in D$. Тут d – числовий інтервал, множина булевих сталих, множина цілих чисел або множина лінгвістичних змінних.

У моделі (1) позначено M_1 – модель підпроцесу формування опису предметної області, M_2 – модель підпроцесу опрацювання даних, M_3 – модель підпроцесу виявлення прихованих залежностей, M_4 – модель підпроцесу оцінювання та інтерпретації результатів аналізу.

Оскільки процес аналізу відбувається у предметній області Π, то почнемо її побудову з моделі підпроцесу формування її опису. Для цього введемо низку позначень.

У загальному випадку модель M описує процес знаходження розв'язку задачі класифікації, тобто побудови відображення $F: X \to D$ множини об'єктів $X \subset \Pi$ у множину класів D. Знаходження такого відображення дає змогу кожному об'єкту x, $x \in X$ надати його значення $d = F(x)$. Кожний об'єкт x, $x \in X$ описаний множиною властивостей a, $a \in A$.

Нехай у предметній області Π розглядається множина об'єктів $X = \{x_1, x_2, \mathbf{K}, x_n\}$. Кожний об'єкт x_i, $x_i \in X$, $i = \overline{1, n}$ розглядаємо як кортеж його властивостей $a(x_i) = (a_1(x_i), a_2(x_i), \mathbf{K}, a_{m_i}(x_i))$.

Множину властивостей усіх об'єктів X предметної області можна сформувати за правилом $A = \bigcup_{i=1}^{n} a_i$. Для побудови математичної моделі предметної області множину властивостей можна сформувати ще у такий спосіб $A = \bigcap_{i=1}^{n} a_i$. Кожний такий спосіб формування множини атрибутів породжуватиме невизначеність.

Вважатимемо, що $A = \{a_1, a_2, \mathbf{K}, a_m\}$. Кожне значення v цих властивостей $v \in V_{a_i}$, $i = \overline{1, m}$ є дійсним числом $v \in \mathsf{R}$, цілим числом $v \in \mathsf{Z}$, булевою сталою $v \in \mathsf{B}$, $\mathsf{B} = \{T, F\}$ або лінгвістичною змінною, визначеною на скінченній множині слів $v \in \mathsf{L}$. Для кожної властивості a визначимо множину її значень V_a потужності $|V_a|$. Якщо $v \in \mathsf{Z}$, $v \in \mathsf{B}$ чи $v \in \mathsf{L}$, то множина V_a є скінченною. Для властивостей, значення яких є дійсними числами, тобто $v \subset \mathsf{R}$, множина V_a є, взагалі кажучи, нескінченною.

Кожний об'єкт x_i, $x_i \in X$, $i = \overline{1,n}$ має ще одну властивість d, $d \in V_d$ Ця властивість є узагальненою характеристикою об'єкта, класифікаційною ознакою, класом або рішенням, яке прийняте щодо цього об'єкта. Приклади вважаємо класифікованими, якщо для кожного з них визначено клас, до якого він належить. Значення $d_i, d_i \in V_d$ породжує клас еквівалентності $[d]_i$, а множина усіх класів, яка здійснює розбиття $X = \{X_1, X_2, \mathbf{K}, X_{|V_d|}\}$ множини X. Важливим є той факт, що $|V_d| < n$. Щодо типу значень елементів кожного класу можуть бути висловлені такі самі припущення, що і для значень інших властивостей. Отже, множину властивостей позначимо $A \mathbf{U} \{d\}$, а інформаційний опис предметної області зображатимемо двоелементним кортежем $(X, A \mathbf{U} \{d\})$.

Тепер задачу побудови моделі процесу аналізу даних можна сформулювати так. Нехай існує невідоме відображення F множини об'єктів $X = \{x_1, x_2, \mathbf{K}, x_n\}$, заданих множинами своїх властивостей, у множину класів d. Задачу аналізу розглядатимемо як таку, що полягає у визначенні функції F та класу еквівалентності, до якого можна віднести об'єкт $y \in \Pi \setminus X$ на основі припущення, що $F(y) = d'$ та $d' \in d$.

Модель підпроцесу формування опису предметної області повинна надати у формалізованому вигляді вхідні дані та результати застосування функцій для попереднього опрацювання даних. Формування опису предметної області повинне враховувати особливості даних, пов'язані із їх можливою невизначеністю та надлишковістю. Виявленням цих особливостей можна зменшити розмірність задачі аналізу в результаті їх виявлення та усунення.

Дані здебільшого зібрано в процесі натурного експерименту або обстеження предметної області. Задача аналізу даних вимагає спеціального підготовлення даних. Спеціального аналізу предметної області з метою виникнення надлишковості та невизначеності здебільшого не здійснюють.

Модель M_1 підпроцесу формування опису предметної області встановлює відповідність множині описів об'єктів X досліджуваної предметної області таблицю прийняття рішень, яку запишемо кортежем $(X, A \mathbf{U} \{d\})$. Подамо математичне формулювання моделі як кортеж такого вигляду:

$$M_1 = (X, O, A, d, \rho(x,a), \eta(x,d), \Psi(X), \Phi(A)). \quad (2)$$

У формулюванні (2) моделі прийнято позначення: $X = \{x_1, x_2, \mathbf{K}, x_n\}$ – множина об'єктів предметної області; $A = \{a_1, a_2, \mathbf{K}, a_m\}$ – множина атрибутів об'єктів; d – клас еквівалентності. У рівності (2) позначено O – модель онтології предметної області

$$O = (X, R, I),$$

де R – множина відношень на X, I – множина інтерпретацій відношень R.

Функції, які реалізує модель підпроцесу, визначимо так: $\rho(x,a)$: $X \times A \to V_a$ – функція, яка надає значення атрибуту a $(a \in A)$ об'єкта x $(x \in X)$; V_a – множина значень функції $\rho(x,a)$; $\eta(x,a)$: $X \times A \to V_d$ – функція, яка надає значення атрибуту прийняття рішень d об'єкта x $(x \in X)$; V_d – множина значень функції $\eta(x,d)$.

Постановка задачі побудови моделі процесу аналізу даних виконана із врахуванням часткової невизначеності та надлишковості даних. Проблема надлишковості у такій постановці еквівалентна одній із таких ситуацій: або значення функції $F(x)$ невідоме для $x \in \Pi \setminus X$, або існують два об'єкти x_i та x_j, $i \neq j$, для яких $a(x_i) = a(x_j)$ та $h(x_i, d) = h(x_j, d)$. Надлишковість може

існувати у такій формі, що існує дві властивості x_i та $x_j, i \neq j$, для яких $a(x_i) = a(x_j)$ та $h(x_i, d) \neq h(x_j, d)$.

Надлишковість полягає в наявності властивостей, які не впливають на отримання результату або створюють інформаційний шум. За проявом невизначеності та надлишковості даних моделі поділятимемо на явні та приховані. У моделі формування підпроцесу опису предметної області невизначеність та надлишковість є явною.

Оператори проектування, які зменшують розмірність задачі вилученням надлишкових об'єктів та властивостей, є такими: $\Psi(X)$ – здійснює відображення $\Psi : X \to X'$, де X' – множина об'єктів, яку отримано скороченням множини об'єктів X для усунення у ній надлишкових об'єктів у кількості $|X \setminus X'|$; $\Phi(A)$ – оператор, який здійснює відображення $\Phi : A \to A'$, де A' – множина властивостей, отриманих усуненням надлишкових властивостей з множини A у кількості $|A \setminus A'|$.

У введених позначеннях можна розглядати різні випадки, а саме: $V_a, V_d \subset \mathsf{R}$, де R – множина дійсних чисел; $V_a, V_d = \{0;1\}$ – множина булевих значень; $V_a, V_d \subset \mathsf{L}$ – множина лінгвістичних значень, причому $|\mathsf{L}| \geqslant 2$; зокрема, надзвичайно поширеним є випадок $|\mathsf{L}| = 2$, коли $\mathsf{L} = \{\text{Так, Ні}\}$, або $\mathsf{L} = \{\text{Yes, No}\}$.

Результатом виконання підпроцесу формування опису предметної області є таблиця прийняття рішень вигляду

$$T_1 = (X_1, A_1 \cup \{d\}). \tag{3}$$

У таблиці (3) позначено через X_1 – множину об'єктів, A_1 – множину властивостей об'єктів, d – ознака або клас об'єкта. Модель підпроцесу попереднього опрацювання інформації визначимо так:

$$M_2 = (T_1, Comp(a), Disc(x), EcsC(a'), EcsR(x)). \tag{4}$$

У формуванні моделі (4) враховано таку особливість таблиці T_1, як відсутні значення властивостей. Відсутнім значенням називаємо таку ситуацію, що існує певний об'єкт x у таблиці T_1 та його властивість a, що $\rho(x, a) = *$, де знаком "*" позначено невідоме значення властивості. Функція $Comp(a)$ визначає це невідоме значення. У результаті отримано властивість a', для якої $\rho(x, a') \neq *$. У процесі попереднього опрацювання даних невизначеність та надлишковість є прихованими та вимагають ідентифікації або усунення.

Якщо властивість a є такою, що $V_a \subset \mathsf{R}$, то виникає ситуація, в якій застосування методів машинного навчання для виявлення залежностей у даних може призвести до ситуації, яку називають перенавчанням, або overfitting. У цій ситуації функцією $Disc(a)$ дискретизації неперервних значень властивостей формується властивість a', значення якої $V_{a'}$ утворюють з V_a розбиттям її на скінченну кількість інтервалів $[a_i, b_i]$, $i = \overline{1,k}$, $k < n$, кожному з яких надають значення i – номер цього проміжку. У припущенні, що існує принаймні одна пара індексів i, j, $i \neq j$ та $a(x_i) = a(x_j)$, але $d_i \neq d_j$, тобто можна стверджувати, що існує невизначеність. Усуває таку невизначеність функція $EcsR(X')$, значення якої – множина об'єктів X_2'' – не містить вказаної невизначеності.

Нехай є така властивість a, що всім об'єктам множини X відповідає множина об'єктів $\overline{X}$, яку отримано з об'єктів X описом властивостями множини $A \setminus \{a\}$. При цьому $F(\overline{X}) = F(X)$, тобто усунення властивості a не змінило значень функції на множині об'єктів. У результаті отримано множину властивостей A_2 за допомогою функції $EcsC(a)$ для вилучення властивості a.

У результаті обчислення функцій моделі підпроцесу попереднього опрацювання даних отримаємо таку таблицю прийняття рішень

$$T_2 = \left(X_2, A_2 \cup \{d\}\right).$$

Модель підпроцесу виявлення залежностей запишемо у вигляді

$$M_3 = \left(T_2, S, Edu\left(X_2\right), Test\left(X_2\right), Pat\left(X_2^н, S\right)\right). \quad (5)$$

Для подання залежностей у даних використовують функції різного типу для класифікації об'єктів віднесенням їх до певних класів еквівалентності. Такі функції мають відомий вигляд і є багатопараметричними. Під шаблоном S у формулі (5) розумітимемо загальний вигляд такої функції.

Залежність, приховану у даних, будують методами машинного навчання. Такі залежності отримують у результаті знаходження параметрів заданого шаблону S. Для застосування методів машинного навчання треба в таблиці прийняття рішень T_2 виділити певну частину об'єктів (навчальні об'єкти), на яких побудувати цю залежність. Множину навчальних об'єктів $X_2^н$ отримують з об'єктів X_2 функцією $Edu\left(X_2\right)$. Для знаходження множини значень параметрів z застосовано метод машинного навчання, за яким обчислюють функцію $Pat\left(X_2^н, S\right)$ на множині навчальних об'єктів $X_2^н$ для залежності, структуру якої задано шаблоном S.

Модель підпроцесу оцінювання та інтерпретації сформулюємо у такому вигляді

$$M_4 = \left(T_2, \tau, F\left(z, X_2^m\right), \theta\left(f, \tau\right), Evl\left(d_3\right)\right) \quad (6)$$

Перевіряють отриману залежність на множині тестових об'єктів X_2^m, які є частиною об'єктів множини X_2 та сформовані функцією $Test\left(X_2\right)$ у співвідношенні (6). У результаті обчислення значень параметрів z для функції, визначеної шаблоном S, можна обчислити значення f як значення функції $F\left(z, X_2^m\right)$, якого вона набуває на множині тестових об'єктів X_2^m. Класифікаційну ознаку d_3 обчислюємо як значення функції $\theta\left(f, \tau\right)$ на множині тестових об'єктів X_2^m на основі оцінок f та заданого порогового значення τ, $\tau \in \lfloor 0;1 \rfloor$. Оцінка якості побудованого класифікатора p є значенням функції $Evl\left(d_3\right)$.

Висновки

Наведено результати досліджень, присвячених побудові формальної математичної моделі процесу аналізу даних. Формулювання моделі у запропонованому вигляді дає змогу розглядати процеси аналізу даних на основі однакових поглядів на склад, структуру, послідовність та зміст кожної складової такого процесу. В основу формулювання моделі покладено поняття процесу як послідовного обчислення спеціальних функцій, за допомогою яких опрацьовують дані, отримані описом предметної області. Істотна особливість наведеної моделі полягає в наявності спеціальних функцій, за допомогою яких усувається часткова невизначеність та надлишковість у даних для зменшення їх розмірності та скорочення часу виконання процедур із виявлення залежностей. Запропоновану модель можна адаптувати до конкретної задачі шляхом уточнення змісту відповідних змінних, параметрів та функцій.

1. Cios K. J. Data Mining: A Knowledge Discovery Approach. /Cios K. J., Pedrycz W., Swiniarski R. W., Kurgan L. A. – New York : Springer Science + Business Media,LLC, 2007. 2. Chapman P. The CRISP-DM process model /Chapman P., Clinton J., Khabaza T., Reinartz T., Wirth R. – Режим доступу: http://www.crisp-dm.org. – 03. 1999. 3. Fayyad, U. The KDD Process for Extracting Useful Knowledge from Volumes of Data / Fayyad U., Piatesky-Shapiro G., Smyth P. // Communications of the ACM. – 1996. – №39(11). – P. 27–34. 4. Cabena P. Discovering Data Mining: From Concepts to Implementation / Cabena P., Hadjinian P., Stadler R., Verhees J., Zanasi A. – Newark:Prentice Hall Saddle River, 1998. 5.

Режим доступу: http://www.crisp-dm.org. 6. Kurgan L. A survey of knowledge discovery and data mining process models / Kurgan L., Musilek P. // Knowledge Engineering Review. – 2006. – №21(1). – P. 1–24. 7. Anand S. Decision Support Using Data Mining. Financial Times Pitman Publishers / Anand S., Buchner A. – London, 1998.8. Anand S. A data mining methodology for cross-sales / Anand S., Hughes P., Bell D. // Knowledge Based Systems Journal.– 1998. – №10.– P. 449–461. 9. Cios K. Diagnosing myocardial perfusion from SPECT bull's-eye maps – a knowledge discovery approach / Cios K., Teresinska A., Konieczna S., Potocka J., Sharma S. // IEEE Engineering in Medicine and Biology Magazine: Special issue on Medical Data Mining and Knowledge Discovery.- 2000.- №19(4). – P. 17–25.10. Cios K. Trends in data mining and knowledge discovery / Cios K., Kurgan L. // Pal N.R. Advanced Techniques in Knowledge Discovery and Data Mining / Pal N.R., Jain L.C.- London: Springer Verlag, 2005. – P. 1–26.11. Kurgan L. Mining the Cystic Fibrosis Data / Kurgan L., Cios K., Sontag M., Accurso F. // Zurada J. Next Generation of Data-Mining Applications / Zurada J., Kantardzic M. – : IEEE Press Piscataway, 2005. – P. 415–444. 12. Pal N.R. Advanced Techniques in Knowledge Discovery and Data Mining / Pal N.R., Jain L.C. – London: Springer Verlag, 2005. 13. Shearer C. The CRISP-DM model: the new blueprint for data mining // Journal of Data Warehousing. – 2000. –№5(4). – P. 13–19.

УДК 004.4'232

О. Овсяк
Львівська філія Київського національного університету
культури і мистецтв

МОДЕЛІ РЕКУРСІЇ ТА РЕКУРЕНЦІЇ

Засобами розширеної алгебри алгоритмів описано моделі рекурсії та рекуренції, наведено приклади їхнього використання.

Recursive and reversing models are described by means of expanded algorithms algebra, the using examples are given.

Вступ

Рекурсія і рекуренція відіграють надзвичайно важливу роль у математиці, алгоритмах і програмуванні. Результатом їх використання у математиці та алгоритмах є компактні вирази функцій та алгоритмів, а у програмуванні, крім компактності коду, досягають істотного зменшення кількості виконуваних операцій і, тим самим, зменшення затрат обсягів пам'яті та часу виконання програм. Однак сьогодні ще немає моделі опису рекурсії та рекуренції у формулах алгоритмів, отриманих застосуванням розширеної алгебри алгоритмів. Власне ці питання і розглядаються у статті. Для створення коректної моделі насамперед необхідно проаналізувати моделі рекурсії та рекуренції у математиці і програмуванні.

Терміни "рекурсії" і "рекуренції" у математиці

В енциклопедії математики [1] так означено **рекурсію**: "РЕКУРСИЯ – способ определения функций, являющийся объектом изучения в теории алгоритмов и других разделах математической логики. Это способ давно применяется в арифметике для определения числовых последовательностей (*прогрессии, чисел Фибоначчи* и пр.). Существенную роль играет рекурсия в вычислительной математике (рекурсивные методы). Наконец, в теории множеств часто используется трансфинитная рекурсия. Долгое время термин рекурсия употреблялся математиками, не будучи