УДК 004.67

ФОРМАЛІЗАЦІЯ СТРУКТУР ЗБЕРІГАННЯ ІНФОРМАЦІЇ В ЗАДАЧАХ ІНДУКТИВНОГО МОДЕЛЮВАННЯ

Н.В. Щербакова
Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України,
nataliya.shcherbakova@gmail.com

При розв'язанні реальних задач побудови моделей за статистичними даними постає питання зберігання інформації та забезпечення ефективного доступу до неї. Для вирішення існуючої проблеми розробляється інтегроване середовище зберігання інформації. Головне питання, яке виникає при розробці середовища, є формалізація структур зберігання інформації в середовищ, а саме формат зберігання даних та структури представлення даних в сховищі.
Ключові слова: інтегроване середовище, обробка та зберігання інформації, індуктивне моделювання, метод групового урахування аргументів

When solving real tasks of model construction from statistical data, the question arises regarding storage of and providing effective access to the information. To solve such a problem, an integrated environment of information storage is developed. The main question arising when developing the environment is formalization of information storing structures within the environment, namely, a format of data storing and presentation of data structures in the storage.
Keywords: integrated environment, handling and storing of information, inductive modeling, group method of data handling.

При решении реальных задач построения моделей по статистическим данным возникает вопрос сохранения информации и обеспечения эффективного доступ к ней. Для решения существующей проблемы разрабатывается интегрированная среда хранения информации. Главный вопрос, который возникает при разработке среды, это формализация структур хранения информации в среде, а именно формат сохранения данных и структуры представления данных в хранилище.
Ключевые слова: интегрированная среда, обработка и хранение информации, индуктивное моделирование, групповой метод учета аргументов.

Вступ

Використання алгоритмів індуктивного моделювання є ефективним для застосування при розв’язанні практичних задач моделювання економічних, екологічних, технологічних та інших складних процесів і систем [1, 2]. Проблема наявності засобів доступного зберігання та оброблення даних наукових досліджень і використання їх результатів залишається актуальною.

У попередніх дослідженнях було запропоновано архітектуру середовища, яка надає можливості вільно маніпулювати наявною інформацією завдяки використанню реляційної бази даних, що містить лише мета-дані та сховища, в якому зберігаються вхідні статистичні дані та результати обчислень.

Подальші дослідження спрямовані на вирішення можливих труднощів, що виникають при розробці інтегрованого середовища обробки та зберігання інформації в задачах індуктивного моделювання. По-перше, на етапі обробки вхідних даних незалежно від вибору метода моделювання поступають дані з різних типів джерел, що містять пропуски і нетипово малі значення, які

необхідно звести до єдиного вигляду. По-друге, постає питання зберігання вихідної інформації, а саме структури та параметри моделей, оцінки достовірності та точності, графіки та інше. Увага зосереджується на узагальненні основних форматів вхідної інформації, формалізації структур зберігання інформації на всіх стадіях моделювання та на можливості зберігання допоміжної інформації.

1. Архітектура системи

В [3] запропонована архітектура інтегрованого середовища зберігання та обробки інформації в задачах індуктивного моделювання, яка надає можливості вільно маніпулювати наявною інформацією завдяки побудові макету середовища, який складається з реляційної бази даних [4,5], що містить лише мета-дані та сховища, в якому зберігаються вхідні статистичні дані та результати обчислень[6].

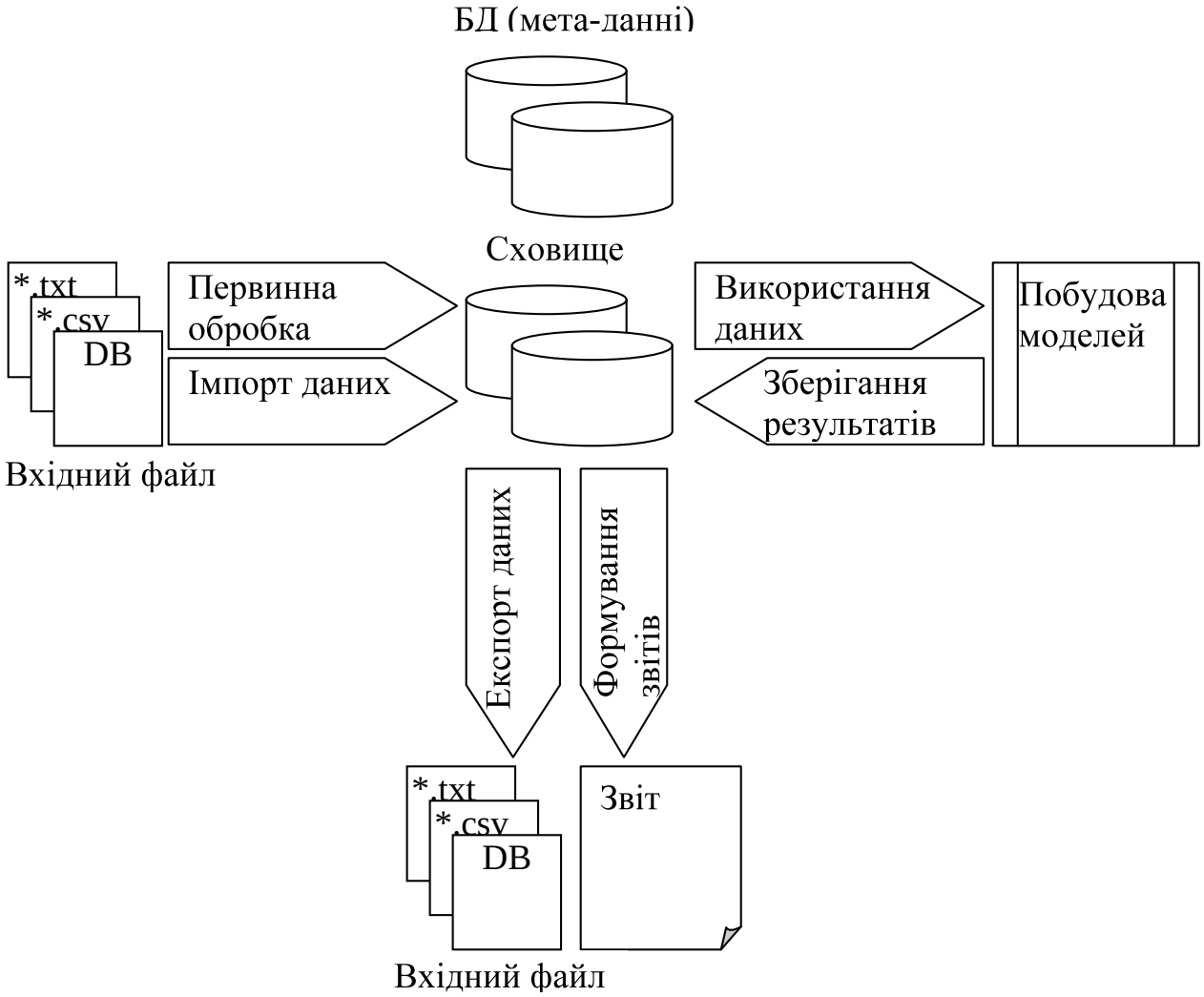

Рис.1 Архітектура інтегрованого середовища оброблення та зберігання інформації

Пропонований макет середовища спрямований на вирішення проблем зберігання вхідних статистичних даних та результатів. Розроблена архітектура середовища зберігання та обробки інформації в задачах індуктивного моделювання (Рис.1) надає можливість розробити на базі неї програмну систему. Модульна архітектура системи надає можливості розширення її функціональності.

Основними вимогами до системи є можливість імпортувати (включаючи первинну обробку) та експортувати вхідні дані, зберігати та обробляти існуючу інформацію, зберігати вихідні дані з повною інформацією про розрахунки, формувати звіти по результатах обчислень. Слід зауважити, що результати обчислень мають зберігатись в системі у стандартизованому вигляді, що дозволить побудувати строго формалізовані звіти по результатах обчислень. Відкритим залишалось питання узагальнення та структуризації основних форматів вхідної інформації та розробка стандартизованих форматів зберігання результатів обчислень.

Розглянемо більш детально яку саме інформацію потрібно зберігати у системі. По-перше це як вже обговорювалось вище вхідні статистичні данні, що приведено до єдиного формату, та оброблені данні, в яких ліквідовано пропуски і нетипово малі значення, тощо. По-друге опорну функцію, згенеровані моделі, оцінки параметрів структур, критерії якості моделей та найкращу модель. Всю цю інформацію доречно зберігати у сховищі. Допоміжну інформацію, а саме данні про виконавця, дату та час, які файли було використано тощо доречно зберігати в реляційній базі даних.

2. Первинна формалізація структур зберігання інформації в задачах індуктивного моделювання

Розв’язання реальних задач побудови моделей з використанням алгоритмів МГУА за допомогою розроблюваного середовища можна розділити на такі етапи:

- імпорт вхідних статистичних даних з різних типів джерел;
- первинна обробка вхідних статистичних даних;
- вибір класу моделей;
- вибір алгоритму генерації (перебору) структур;
- вибір методу оцінювання параметрів структур;
- вибір критерію селекції кращих моделей;
- оцінка адекватності моделі;
- застосування результуючої моделі.

Використаємо теорію множин для формалізації представлення даних на кожному з цих етапів. Маємо такі компоненти (побудовані з використанням матеріалу аналізу методу структурної ідентифікації [7]):

$W = (X, Y)$ – множина статистичних даних (послідовність N значень випадкової величини Y, що характеризується M ознаками X)

$\mathrm{W} = \left\{ w_j \right\},\ j = \overline{1,J},\ J = n \cdot m,\ n = \overline{1,N},\ m = \overline{1,M}$;

NW – множина нормалізованих статистичних даних $NW = \left\{ \overline{w}_j \right\},\ j = \overline{1,J}$;

F – множина класів моделей $F = \left\{ f_k \right\}, k = \overline{1,K}$;

G – множина генераторів структур моделей $G = \left\{ g_l \right\}, l = \overline{1,L}$;

P – множина методів оцінювання параметрів структур $P = \left\{ p_r \right\}, r = \overline{1,R}$;

CR – множина критеріїв якості моделей $\mathrm{CR} = \left\{ cr_q \right\}, q = \overline{1,Q}$;

V – множина прогнозуючих моделей $\mathrm{V} = \left\{ v_t \right\}, t = \overline{1,T}$.

Тоді формально процес побудови множини всіх можливих моделей можна представити у вигляді прямого добутку складових множин $Z = W \times NW \times F \times G \times P \times CR \times V$. Деякий елемент множини Z, описаний як

$$z_i = \left\{ w_j, \overline{w} j, f_k, g_l, p_r, cr_q, v_t \right\},\ j = \overline{1,J}, k = \overline{1,K}, l = \overline{1,L}, r = \overline{1,R}, q = \overline{1,Q}, t = \overline{1,T},$$
$$i = \overline{1,I},\ I = J \cdot K \cdot L \cdot R \cdot Q \cdot T$$

будемо розглядати як конкретні дані, які було збережено у середовищі при проходженні конкретного повного циклу моделювання за статичними даними.

Таким чином, на основі апарату теорії множин виконано первинну формалізацію даних, які зберігаються в сховищі.

Розглянемо детальніше яка саме інформація зберігається після відпрацювання кожного з модулів системи. По-перше модуль імпорту розміщує в сховищі вхідні статистичні данні та нормалізовані данні. На другому етапі в процесі побудови моделі, використовуючи існуючи в сховищі статистичні данні, класи опорних моделей, генератори структур моделей, методи оцінювання параметрів структур та критерії якості моделі, будується модель, інформація про структуру якої повертається в сховище. Нижче розглянемо існуючи формати для зберігання прогнозуючих моделей та можливість їх використання в нашому випадку.

3. Використання PMML для зберігання інформації

Найбільш поширеним засобом зберігання інформації, що використовується при моделюванні, враховуючи системи інтелектуального аналізу даних є мова розмітки PMML.

PMML (Predictive Model Markup Language) - це XML-діалект, який використовується для опису статистичних моделей. Його головна перевага

полягає в тому, що PMML-сумісні програми дозволяють легко обмінюватися моделями з іншими PMML-інструментами. Класи моделей, які можна зберігати засобами даної мови розмітки такі: асоціативні правила, дерева рішень, кластери, регресія, загальна регресія, нейронні мережі, Баєсівська мережа, послідовності, текстові моделі, часові ряди, множини правил, дерева, опорні вектори.

В нашому випадку для зберігання прогнозуючих моделей можна використати схему, що описує регресійні функції. Регресійні функції, що мають змогу зберігатися засобами PMML використовуються для визначення взаємозв'язку між залежною змінною (цільової області) і одинією або більше незалежних змінних. Залежною змінною є одна, значення якої передбачається, у той час як незалежні змінні є змінними, на яких базується прогноз. Термін регресія, як правило, відноситься до прогнозування числових значень, тому PMML елемент RegressionModel також може бути використаний для класифікації. Це пояснюється тим фактом, що кілька рівнянь регресії можуть бути об'єднані, щоб передбачити, категорію цінностей [8].

Використання PMML дає можливість зберігати просту регресійну модель в наступному вигляді:

$$Dependent\ variable = intercept + Sum_i\ (coefficient_i * independent\ variable_i) + error.$$

А модель класифікації може мати кілька регресійних рівнянь виду:

$$y_j = intercept_j + Sum_i\ (coefficient_{ji} * independent\ variable_i).$$

Слід зауважити що автори використовують найбільш поширену версію PMML – 3.2. Однак використання PMML не дає можливості зберігання повного набору даних що мають міститись у сховищі. Тому в даному випадку в цілях стандартизації форматів збереження повної інформації має місце розробка власної XML-схеми для можливості зберігання всієї необхідної інформації по розрахунках в єдиному форматі в сховищі.

4. Використання XML для зберігання інформації

Використання XML для зберігання даних у сховищі, збільшує можливості щодо розширення системи, в тому числі можливість розробки власних схем XML-документів.

В сховищі повинні зберігатись вхідні статистичні дані, а саме множина значень випадкової величини та множина залежних параметрів. Крім того слід зберігати у сховищі нормалізовані дані.

Слід звернути увагу на можливість зберігання не лише прогнозуючих моделей, а також і допоміжної інформації. А саме: можливі класи моделей, генератори структур, інформацію щодо методів оцінки параметрів структур,

інформацію щодо критеріїв якості моделей. Розробка власних XML-схем для зберігання інформації в сховищі також дає можливості побудови звітів не лише для прогнозуючих моделей, а також для всієї допоміжною інформації, що міститься в сховищі.

Висновки

В статті наведено архітектуру інтегрованого сховища обробки та зберігання інформації в задачах індуктивного моделювання, яку було розроблено у попередніх дослідженнях.

Наведено опис первинної формалізації структур зберігання інформації в сховищі (вхідних даних та результатів обчислень). Використано теорію множин для формалізації процесу моделювання на всіх стадіях, в тому числі вибір класу моделей, алгоритму генерації структур, методу оцінювання параметрів структур, критерію селекції кращих моделей, оцінки адекватності моделі та безпосередньо моделі.

Виконано огляд можливостей застосування PMML для зберігання результатів обчислень в задачах індуктивного моделювання на основі МГУА. А також проаналізовано напрямки розробки власної XML-схеми, що дозволить зберігати всі необхідні дані, що використовуються при прогнозуванні за допомогою алгоритмів МГУА.

Література

1. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. – К.: Наук. думка, 1985. – 216с.

2. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем. – К.: Наук. думка, 1982 – 296с.

3. Щербакова Н.В., Степашко В.С. Про концепцію інтегрованого середовища обробки та зберігання інформації в задачах індуктивного моделювання. // Моделювання та керування станом еколого-економічних систем регіону. Збірник праць. К.:МННЦІТС, 2008. – 260с.

4. К. Жд. Дейт Введение в системы баз данных, 8-е издание. – М.: Вильямс, 2005. – 1328с.

5. Томас К., Карелин Б. Базы данных. Проектирование, реализация и сопровождение. Теория и практика. – М.: Вильямс, 2003. – 1440с.

6. Грейвс М. Проектирование баз данных на основе XML. Пер. с англ. – М.: Вильямс, 2002. – 640с.

7. Ефименко С.Н., Степашко В.С. Имитационный эксперимент как средство для исследования эффективности методов моделирования по данным наблюдений // УСІМ. – К:МНУЦІТС, 2008. – 160с.

8. http://www.dmg.org/ - Data Mining Group