УДК 004.67

АНАЛІЗ СТРУКТУР ВХІДНИХ ДАНИХ В ЗАДАЧАХ ІНДУКТИВНОГО МОДЕЛЮВАННЯ

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

Одним із важливих етапів на шляху розробки інтегрованого середовища обробки та зберігання інформації в задачах індуктивного моделювання є проектування засобів інтеграції зовнішніх даних у сховище. У статті розглядаються ключові проблеми, що виникають на цьому етапі, в тому числі проаналізовано структури вхідних даних за різними параметрами.
Ключові слова: індуктивне моделювання, метод групового урахування аргументів, структури даних, обробка та зберігання інформації, інтегроване середовище.

One of the most important steps towards developing an integrated environment of storing and handling information in the tasks of inductive modeling is the design of the integration of external data in the repository. The article discusses key issues arising at this stage, including the structure of input data analyzed by various parameters.
Keywords: inductive modeling, group method of data handling, structures of data, handling and storing of information, integrated environment.

Одним из важнейших этапов на пути разработки интегрированной среды обработки и хранения информации в задачах индуктивного моделирования является проектирование средств интеграции внешних данных в хранилище. В статье рассматриваются ключевые проблемы, возникающие на этом этапе, в том числе проанализированы структуры входных данных по различным параметрам.
Ключевые слова: индуктивное моделирование, метод группового учета аргументов, структуры данных, обработка и хранение информации, интегрированная среда.

Вступ. Для початку розглянемо що є вхідними даними в задачах індуктивного моделювання. Взагалі дані – це подання фактів та ідей у формалізованому вигляді, придатнім для передачі та обробки в деякому інформаційному процесі [1]. Для вирішення задач з використанням алгоритмів індуктивного моделювання вхідні статистичні данні мають бути строго формалізовані та зведені до табличного вигляду.

Дані є всюди навколо нас, в різних форматах представлення, з різних джерел, мають різну структуру та характеристики. Забезпечення сумісного, єдиного формату зберігання даних зі всіх джерел, хоча б у межах однієї організації на сьогоднішній день є чи не найважливішою проблемою з якою стикаються інформаційні технології. Розробка інструментів для збору, обробки та представлення даних надає можливості швидше та якісніше приймати рішення, спираючись на існуючі дані. В останні роки все більшу популярність набувають системи бізнес аналітики, які розділяють на три компоненти: сховище даних, статистичний аналіз та представлення даних, інтелектуальний аналіз.

Сучасні системи бізнес-аналітики зазвичай в своєму складі мають інструменти для інтеграції даних, а саме ETL-інструменти (Extraction,

Transformation and Loading) для вилучення даних із зовнішніх джерел, їх трансформації та загрузки у сховище. Вилучення даних – це копіювання із оперативних систем, документів та інших джерел, забезпечуючи цілісність та унікальність даних. Трансформація передбачає перетворення даних до загального вигляду, очистку від помилок, прив'язку до розмірностей. Перенос трансформованих даних у сховище виконується на етапі загрузки. Інтегровані в систему дані можна використати як для побудови прямого звіту, так і для подальшого аналізу з використанням алгоритмів інтелектуального аналізу даних.

Далі проаналізуємо характеристики вхідних даних в задачах індуктивного моделювання за різними параметрами.

1. Формалізація подання вхідних даних

В [2] використано теорію множин для формалізації подання даних на кожному з етапів побудови моделей з використанням алгоритмів МГУА [3]. Маємо такі компоненти (побудовані з використанням матеріалу аналізу методу структурної ідентифікації [4]):

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

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

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

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

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

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

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

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

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

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

$$i = \overline{1, I},\ I = J \cdot K \cdot L \cdot R \cdot Q \cdot T$$

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

У цій статті розглянемо детальніше множини статистичних даних та нормалізованих даних. Зупинімося на структурах вхідних даних, джерелах, засобах зберігання та інших характеристиках. Окремо розглянемо питання перетворювання множини статистичних даних до множини нормалізованих даних.

2. Аналіз типових структур вхідних даних

За ступенем структуризації виділяють наступні форми представлення даних: структуровані, неструктуровані та слабоструктуровані [5].

До неструктурованих відносяться дані, довільні за формою, включаючи тексти і графіку, мультимедіа (відео, мова, аудіо). Ця форма представлення даних широко використовується, наприклад, в Інтернеті, а самі дані представляються користувачеві у вигляді відгуку пошуковими системами.

Структуровані дані відображають окремі факти предметної області. Структурованими називаються дані, певним чином впорядковані і організовані з метою забезпечення можливості застосування до них деяких дій (наприклад, візуального або машинного аналізу). Це основна форма подання відомостей у базах даних. Всі структуровані дані поділяються на п’ять основних типів: цілий, дійсний, строковий (впорядкований, категоріальний), логічний, дата/час. Джерелом структурованих даних є база даних.

Слабоструктуровані дані – це дані, для яких визначено деякі правила і формати, але в самому загальному вигляді. Наприклад, рядок з адресою, рядок у прайс-листі, ПІБ і т. п. На відміну від неструктурованих, такі дані з меншими зусиллями перетворяться до структурованої формі, однак без цієї процедури вони теж непридатні для аналізу. До джерел неструктурованих даних можна віднести типові електроні формати наприклад PDF, Excel, HTML, CSV та інші.

Деякі джерела [6,7] виділяють також квазіструктуровані дані, маючи на увазі дані наприклад в форматі XML.

Організація того чи іншого виду зберігання даних (структурованих або неструктурованих) пов'язана із забезпеченням доступу до них. Під доступом мається на увазі можливість виділення елемента даних (або безлічі елементів) серед інших елементів за будь-якими ознаками з метою виконання деяких дій над елементом. Однією з найпоширеніших моделей зберігання структурованих даних є таблиця. У ній всі дані упорядковуються в двовимірну структуру, що складається із стовпців і рядків. В комірках такої таблиці містяться елементи даних: символи, числа, логічні значення.

Неструктуровані дані непридатні для обробки безпосередньо методами аналізу даних, тому такі дані піддаються спеціальним прийомам структуризації, причому сам характер даних у процесі структурування може істотно змінитися. Наприклад, в аналізі текстів (Text Mining) при структуруванні з вихідного тексту може з'явитися таблиця з частотами зустрічі слів, і вже такий набор даних буде оброблятися методами, придатними для структурованих даних. Більшість методів аналізу даних, в тому числі і алгоритми індуктивного моделювання, працює тільки з добре структурованими даними, що представлено в табличному вигляді.

За критерієм постійності своїх значень в процесі вирішення задачі дані можуть бути змінні, сталі, умовно-сталі [8]. Змінні дані – це такі дані, котрі змінюють свої значення в процесі рішення задачі. Сталі дані – це такі дані, котрі зберігають свої значення в процесі рішення (математичні константи,

координати нерухомих об'єктів) та не залежать від зовнішніх факторів. Умовно-сталі дані – це такі дані, котрі можуть інколи змінювати свої значення, але ці зміни не залежать від процесу рішення задачі, а визначаються зовнішніми факторами.

В залежності від тих функції які виконують дані, вони можуть бути довідникові, оперативні, архівні.

Слід відрізняти дані за період та точкові дані. Ці відмінності важливі при проектуванні системи збору інформації, а також в процесі змін. Дані за період характеризують деякий період часу. Зразком даних за період можуть бути: прибуток підприємства за місяць, середня температура за тиждень. Точкові дані представляють значення деякої змінної в конкретний момент часу. Зразок точених даних: залишок на рахунку на перше число місяця, температура о сьомій годині ранку.

Дані бувають первинні та вторинні. Вторинні дані – це дані, які являються результатом визначених обчислень, застосованих до первинних даних. Вторинні дані, як правило, призводять до прискореного отримання відповіді на запит користувача за рахунок збільшення об'єму інформації.

Структури вхідної інформації за типами вхідних даних: опис об'єктів за ознаками, опис взаємовідношень між об'єктами, часовий ряд або сигнал, зображення або відеоряд.

Окремо слід виділити поняття метадані – це дані про дані. В склад метаданих можуть входити: каталоги, довідники, реєстри. Метадані містять відомості про состав даних, зміст, статус, походження, якість, формат та форму представлення, умови доступу, придбання та використання авторських, майнових та сумісних з ними правах на дані та інше. Бізнес метадані містять бізнес-терміни та означення, належність даних. Оперативні метадані – це інформація, зібрана під час роботи сховища даних: походження перенесених та перетворених даних, статус використання даних, дані моніторингу, такі як статистика використання, повідомлення про помилки та інше.

Аналізуючи структури вхідної інформації окремо слід зупинитися на питанні вимірювань, а саме в якому форматі спеціаліст веде протокол своїх досліджень або у організації прийнято зберігати оперативні дані. Іншими словами як представити дані отримані з певного джерела. Зазвичай в таблиці заносяться цифри, але зустрічаються букви малюнки, тексти, аудіо, відео та інше [9]. Якщо отримані дані призначені для використання в системах аналітики, то потрібно забезпечити однозначне розуміння змісту даних. Питаннями пов'язаними із вивченням засобів відображення властивостей об'єктів спостереження займається теорія вимірювань [10]. Коротко розглянемо відомості цієї теорії що потрібні для опису методів аналізу даних.

Теорія вимірювань оперує поняттям «емпірична система відношень» (E), яка включає в себе множину вимірюваних об'єктів (A) та набір відношень між об'єктами (R). Для запису результатів досліджень використовується символьна система (W). Домовленість використання саме такого відображення системи E

на систему W означає вибір деякого визначеного правила відображення g. Трійка елементів (E, W, g) називається шкалою. На практиці розповсюдження отримали наступні типи шкал:

- абсолютна шкала (допустиме перетворення для шкал такого типу представляє собою тотожність);
- шкала відношень (між різними протоколами, фіксуючими один і той же емпіричний факт на різних мовах, при цьому типі шкали повинно виконуватись співвідношення: $y = ax$, де a – будь-яке додатне число);
- шкала інтервалів (між протоколами припустимі перетворення $y = ax + b$, де a – будь-яке додатне число, b – будь-яке додатне або від'ємне число);
- шкала порядку (допустимі перетворення для даного типу шкали є всі монотонні перетворення, тобто такі, які не порушують порядок слідування значень вимірюваних величин);
- шкала найменувань (фіксує два відношення: «дорівнює», «не дорівнює»).

3. Первинна обробка даних в системах інтеграції даних

Сучасні інструменти інтеграції даних зазвичай надають ряд стандартних можливостей обробки даних [11], а саме:

- експорт даних із інших баз даних, текстових файлів, Excel-таблиць, та інших джерел;
- імпорт даних із сховища в інші бази даних, текстові файли, Excel-таблиці та інші ресурси;
- міграція між різними СКБД (в тому числі різними версіями однієї СКБД);
- дослідження даних в існуючих базах даних;
- додавання даних в різні сховища;
- інтеграцію комбінованих даних з різних джерел;
- попередню обробку даних застосовуючи комплексні підходи;
- та інше.

Слід зазначити, що попередня обробка даних зазвичай передбачає наступні операції:

- графічний аналіз (візуалізація) даних;
- пошук аналіз, оцінювання, заповнення пропусків;
- пошук, аналіз, оцінювання або дослідження аномалій (викидів);
- фільтрація (згладжування шумів даних);
- дисперсійний аналіз середніх значень;
- перевірка однорідних даних.

4. Приклад структур вхідних даних в задачах індуктивного моделювання. Розглянемо типову структуру вхідних даних в задачах індуктивного моделювання на прикладі задачі прогнозу стану хворого діабетом за даними моніторингу, який веде сам хворий [12]. Вхідна вибірка даних наведена в Таблиці 1 (для прикладу із всіх даних моніторингу хворого діабетом за 6 місяців вибрані 28 днів хвороби).

Характеристичний вектор пацієнта включає в себе всі показники, що реєструються в карті хворого (стать, вік, зріст, вага та інше). В характеристичний вектор хворого входять також чотири значення рівня глюкози, що виміряються за день: X_{6k} – о 6 годині, перед сніданком; X_{12k} – о 12 годині, перед обідом; X_{17k} – о 17 годині; X_{22k} – о 22 годині, перед сном; а також дози інсуліну, що було призначено хворому протягом доби: X_{6k} – доза інсуліну зранку о 6 годині; X_{12k} – доза інсуліну о 12 годині; X_{17k} – доза інсуліну о 17 годині; X_{22k} – доза інсуліну о 22 годині. У задачі потрібно отримати прогноз рівня глюкози в крові хворого діабетом, котрий буде у нього на задану дату.

Таким чином на вході ми маємо структуровані, змінні, оперативні, первинні дані за заданий період часу. Дані містять пропуски, що слід урахувати при первинній обробці.

Таблиця 1

Вхідні дані спостереження одного пацієнта за 28 днів

Дата	X_{6k}	X_{12k}	X_{17k}	X_{22k}	X_{6k}	X_{12k}	X_{17k}	X_{22k}
1993/03/22	12	8	12	14	13,4	12,6	11,6	12,9
1993/03/23	12	6	12	14	8	13,9	13,2	9
1993/03/24	12	8	11	14	9,5	5,8	8,5	6,3
1993/03/25	12	8	10	14	5,9	12,6	5,4	10,5
1993/03/26	11	6	10	14	13,5	4	11,6	
1993/03/27	11	8	10	14	14,2	14,6	5,6	13,7
1993/03/28	11	8	9	14	15,3	16,5	13,8	3,9
993/03/29	11	8	8	14	12,8	9,3	9,4	12
1993/03/30	11	8	8	14	8,3	6,9	13,1	12,5
1993/03/31	11	8	9	14	13,9	11,5	13,6	10
1993/04/01	11	8	10	14	12,4	10,3	11,2	5,6
1993/04/02	12	8	10	14	5,6	10,9	11,5	7,7
1993/04/03	12	8	11	14	12	6,8	11,2	9,3
1993/04/04	12	8	11	14	12,5	10,1	12,9	10,8
1993/04/05	12	8	11	14	13,5	15,4	11	4,3
1993/04/06	12	8	12	14	10,9	12,9	9	4,8
1993/04/07	13	8	12	14	12,8	16	12,8	14,5
1993/04/08	13	8	12	14	8,7	14,4	17,1	8,2
1993/04/09	13	8	12	16	8,2	16,8	10,9	6
1993/04/10	13	8	12	16	6,7	15	16,9	14,6
1993/04/11	13	8	12	16	5,1	5,9	9,7	5,8
1993/04/12	13	8	12	16	4	13,1	11	12,8

Продовження таблиці 1

1993/04/13	13	8	12	16	10,3	7,9	1,7	4,6
1993/04/14	13	8	12	14	9	8,5	9,2	10,6
1993/04/15	12	8	12	16	10,3	11,4	11,3	3,2
1993/04/16	13	8	12	16	11,3	4	9,8	8,5
1993/04/17	12	8	12	16	5,4	13,2	2,4	5,9
1993/04/18	13	8	11	16	12,2	6	2,9	10,6

Висновки. В статті проаналізовано структури вхідних даних в задачах індуктивного моделювання. Наведено опис даних за ступенем структуризації, класифікацію видів даних, опис шкал вимірювання, опис засобів інтеграції даних у сучасних системах бізнес-аналітики та приклад вхідних статистичних даних, що використовуються при розв'язанні задачі за допомогою алгоритмів МГУА. Проведений аналіз використовується при проектуванні модуля інтеграції даних розроблюваного середовища обробки та зберігання інформації в задача індуктивного моделювання.

Література

1. http://wikipedia.org/
2. Щербакова Н.В Формалізація структур зберігання інформації в задачах індуктивного моделювання // Моделювання та керування станом еколого-економічних систем регіону. Збірник праць. К.: МННЦІТС, 2009. – С. 229- 234.
3. Айстраханов Д.Д., Пугачова М.В., Степашко В.С. та ін. Концептуальні основи статистичного моніторингу. – К.: ІВЦ Держкомстату України, 2003. – 343с.
4. Ефименко С.Н., Степашко В.С. Имитационный эксперимент как средство для исследования эффективности методов моделирования по данным наблюдений // УСІМ. –2009. – №1. –С. 69-78.
5. Орешков В.И., Паклин Н.Б. Бизнес-аналитика: от данных к знаниям. – М.: Питер, 2009, – 624 с.
6. Артемьев В. Что такое Business Intelligence? // Открытые системы, 2003. – № 1.
7. Грейвс М. Проектирование баз данных на основе XML. Пер. с англ. – М.: Вильямс, 2002. – 640 с.
8. Дейт К. Жд. Введение в системы баз данных, 8-е издание. – М.: Вильямс, 2005. – 1328 с.
9. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики, 1999. – 270 с.
10. Супес П., Зинес Дж. Основы теории измерений. – М.: Мир, 1967. – С. 117–132.
11. Ian H. Witten, Eibe Frank. Data mining. Practical Machine Learning Tools and Techniques. – NY.: Morgen Kaufmann Publishers, 2005. – 525 p.
12. Савченко Е.А. Экспресс-прогноз уровня глюкозы в крови по комбинаторному алгоритму МГУА // УСИМ. – 2003. – №3. –С. 107–111.