

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

В.В. Литвин, В.В. Пасічник, Ю.В. Нікольський

АНАЛІЗ ДАНИХ ТА ЗНАНЬ

Навчальний посібник

Видавництво ПП "Магнолія 2006"

Львів 2024

УДК 32.970
Л64

*Відтворення цієї книги або будь-якої її частини
заборонено без письмової згоди видавництва.
Будь-які спроби порушення авторських прав
переслідуватимуться у судовому порядку.*

*Затверджено Міністерством освіти і науки України як посібник для
студентів вищих навчальних закладів*

Аналіз даних та знань [навчальний посібник] – Львів, «Магнолія 2006»,
– 276 с.

Автори:

В.В. Литвин, В.В. Пасічник, Ю.В. Нікольський

Викладено основні методи аналізу даних та знань, особливу увагу звернуто на багатовимірний та інтелектуальний аналіз даних. Висвітлено особливості технологій Data Mining, їх теоретичні та прикладні аспекти. Розглянуто моделі онтологічних систем та методику розроблення онтологій. Детально описано побудову онтології за допомогою програмного засобу Protégé. Розглянуто методи машинного навчання, які використовуються під час аналізу даних та знань.

Навчальний посібник призначений для студентів, що навчаються за напрямами підготовки «Комп'ютерні науки», «Системний аналіз», для магістрів спеціальностей, які базуються на цих напрямах підготовки, а також для магістрів спеціальностей «Управління проектами» та «Консолідована інформація».

«Магнолія 2006»

ISBN 978-617-574-181-8

УДК 32.970
Л64

© В.В. Литвин, В.В. Пасічник,
Ю.В. Нікольський
© «Магнолія 2006»

ЗМІСТ

Передмова наукового редактора.....	7
Вступ.....	9
РОЗДІЛ 1. Інформація. Дані. Знання.....	10
1.1. Поняття про інформацію.....	10
1.2. Інформаційні сигнали матеріального світу.....	13
1.3. Знакове середовище існування інформації.....	13
1.4. Поняття про знакові системи.....	15
1.5. Форми знакового подання інформації.....	16
1.6. Інформація як результат взаємодії даних і методів.....	17
1.7. Властивості взаємозв'язку інформації, даних і методів.....	19
1.8. Синтаксичні міри інформації.....	20
1.9. Семантичні міри інформації.....	21
1.10 Прагматичні міри інформації.....	23
Контрольні запитання.....	24
РОЗДІЛ 2. Факторний аналіз.....	25
2.1. Концепція факторного аналізу.....	25
2.2. Основи факторного аналізу.....	26
2.3. Основні алгоритми та методи.....	27
2.4. Методи виділення первинних факторів.....	28
2.5. Головні компоненти, власні значення та вектори.....	29
2.6. Методи факторного аналізу.....	33
2.6.1. Метод найменших квадратів.....	34
2.6.2. Метод максимальної правдоподібності.....	35
2.6.3. Альфа-факторний аналіз.....	37
2.6.4. Аналіз образів.....	38
2.7. Методи обертання.....	40
2.7.1. Геометричний метод обертання.....	40
2.7.2. Методи ортогонального обертання.....	43
2.7.3. Методи косокутного обертання.....	45
2.7.4. Обертання з використанням цільової матриці.....	47
Контрольні запитання.....	48
РОЗДІЛ 3. Дискримінантний аналіз.....	49
3.1. Суть дискримінантного аналізу.....	49
3.2. Канонічні дискримінантні функції.....	52
3.2.1. Геометрична інтерпретація.....	53
3.2.2. Кількість канонічних дискримінантних функцій.....	53
3.2.3. Одержання коефіцієнтів канонічної дискримінантної функції.....	54
3.2.4. Коефіцієнти v_i	56

3.2.5. Нестандартизовані коефіцієнти.....	57
3.3. Процедури класифікації.....	57
3.3.1. Класифікаційні функції.....	58
3.3.2. Врахування апріорних ймовірностей.....	60
3.3.3. Класифікація за допомогою канонічних дискримінантних функцій.....	61
3.3.4. Графічне зображення областей.....	62
3.3.5. Класифікаційна матриця.....	62
3.3.6. Обґрунтування за допомогою розбиття вибірки.....	63
Контрольні запитання.....	64
РОЗДІЛ 4. Кластерний аналіз.....	65
4.1. Суть кластерного аналізу.....	65
4.1.1. Застереження стосовно використання кластерного аналізу....	67
4.1.2. Поняття подібності.....	67
4.1.3. Вибір змінних.....	69
4.2. Міри подібності.....	70
4.2.1. Коефіцієнти кореляції.....	70
4.2.2. Міри відстані.....	71
4.2.3. Коефіцієнти асоціативності.....	72
4.2.4. Імовірнісні коефіцієнти подібності.....	73
4.3. Методи кластерного аналізу.....	74
4.3.1. Ієрархічні агломеративні методи.....	74
4.3.2. Ітеративні методи групування.....	78
4.3.3. Інші методи.....	81
Контрольні запитання.....	83
РОЗДІЛ 5. Інтелектуальний аналіз даних.....	84
5.1. Що таке інтелектуальний аналіз даних (Data Mining)?.....	84
5.2. Типи закономірностей.....	88
5.2.1. Класи систем Data Mining.....	89
5.2.2. Предметно-орієнтовані аналітичні системи.....	89
Контрольні запитання.....	95
РОЗДІЛ 6. Машинне навчання.....	96
6.1. Інтуїтивне розуміння навчання.....	96
6.2. Означення навчання.....	97
6.3. Програми, що навчаються.....	98
6.4. Мотивація до навчання.....	101
6.5. Таксономія машинного навчання.....	102
6.6. Споріднені галузі.....	104
6.7. Навчання як розділ штучного інтелекту.....	106
6.8. Загальне формулювання задачі навчання за прецедентами.....	107
6.9. Основні поняття та означення.....	107

6.10	Типологія задач навчання за прецедентами.....	108
6.11.	Задачі з описом об'єктів на основі ознак.....	111
6.12.	Приклади задач машинного навчання.....	111
6.12.1.	Задачі класифікації.....	111
6.12.2.	Задача відновлення регресії.....	116
6.12.3.	Задачі прогнозування та прийняття рішень.....	117
6.12.4.	Задачі кластеризації.....	119
6.12.5.	Задачі аналізу клієнтських середовищ.....	120
6.13.	Навчання понять в штучному інтелекті.....	121
6.13.1.	Задача навчання понять – пошук у просторі гіпотез.....	122
6.13.2.	Упорядкування гіпотез „від загальної до конкретної”.....	124
6.13.3.	Алгоритм Find-S пошуку максимально конкретної гіпотези.....	126
6.13.4.	Алгоритм „вилучення кандидата”.....	129
	Контрольні запитання.....	144
	Задачі для самостійного розв'язування.....	145
	РОЗДІЛ 7. Деревя рішень.....	149
8.1.	Означення дерева рішень.....	149
8.2.	Алгоритми побудови дерева рішень.....	152
	Контрольні запитання.....	166
	Задачі для самостійного розв'язування.....	166
	РОЗДІЛ 8. Нейронні мережі.....	169
8.1.	Навчання на основі зв'язків.....	169
8.1.1.	Біологічні нейронні мережі.....	170
8.1.2.	Модель штучного нейрона.....	171
8.1.3.	Подання нейромереж та їхньої архітектури.....	175
8.1.4.	Сучасні архітектури нейромереж.....	177
8.1.5.	Навчання одношарових нейромереж прямого поширення.....	182
8.1.6.	Навчання багатшарових нейромереж прямого поширення.....	189
8.2.	Метод опорних векторів.....	194
8.3.	Мережі, що самоорганізуються.....	210
8.3.1.	Опис мереж, що самоорганізуються.....	210
8.3.2.	Міри відстані між векторами.....	212
8.3.3.	Проблема нормалізації векторів.....	212
8.3.4.	Міра організації мережі.....	213
8.3.5.	Механізм стомлення нейронів.....	214
8.3.6.	Методи навчання мереж, що самоорганізуються.....	215
	Контрольні запитання.....	218
	Задачі для самостійного розв'язування.....	219
	РОЗДІЛ 9. Онтології й онтологічні системи.....	222
9.1.	Поняття онтології.....	222
9.2.	Моделі онтології й онтологічної системи.....	229

9.3. Методології створення і “життєвий цикл” онтологій.....	233
9.4. Мови опису онтологій.....	234
Контрольні запитання.....	238
РОЗДІЛ 10. Програмні засоби побудови онтологій.....	239
10.1. Онтологія як засіб формалізації та алгоритмізації знань.....	239
10.1.1. Аналіз підходів до навчання онтологій.....	240
10.1.2. Загальні принципи проектування онтологій.....	242
10.1.3. Формати та стандарти подання інформації.....	243
10.1.4. Засоби для створення онтологій.....	246
10.2. Технологія розроблення онтологій в редакторі Protégé.....	247
10.2.1. Еволюція Protégé.....	247
10.2.2. Protégé-OWL. Мова Web онтологій OWL.....	249
10.2.3. Основні терміни та поняття у Protégé-OWL.....	251
10.2.4. Методика розроблення онтологій засобами Protégé.....	253
10.2.5. Створення онтологій.....	255
Контрольні запитання.....	268
Використана література.....	269

Передмова наукового редактора серії підручників «КОМП'ЮТИНГ»

Шановний читачу!

Започатковуючи масштабний освітньо-науковий проект підготовки і видання серії сучасних підручників під загальним гаслом «КОМП'ЮТИНГ» та загальним методичним патронуванням його Інститутом інноваційних технологій та змісту освіти МОН України, мені, як ініціатору та науковому керівнику, неодноразово доводилось прискіпливо аналізувати загальну ситуацію в царині сучасного україномовного підручника комп'ютерно-інформатичного профілю. Загалом, позитивна тенденція останніх років ще не співмірна з надзвичайно динамічним розвитком як освітньо-наукової та виробничої сфери комп'ютингу, так і стрімким розширенням потенційної цільової читачької аудиторії цього профілю. Іншими словами, попередній аналіз засвідчує наявність значного соціального замовлення під реалізацію пропонованого Вашій увазі проекту.

Ще одним фактором формування освітньо-наукової ініціативи, пропонованої групою відомих вітчизняних науковців-педагогів та практиків, які організують наукові дослідження, готують фахівців та провадять бізнес в галузі комп'ютингу, постало завдання широкомасштабного включення Української Вищої Школи до загальноєвропейських і всесвітніх об'єднань, структур і асоціацій. Виконуючи функцію науково-технічного локомотиву суспільства, галузь комп'ютингу невідворотно зобов'язана зіграти роль активного творця загальної освітньо-наукової платформи, яка має бути методологічно об'єднаною та професійно-інтеграційною основою для багатьох сфер людської діяльності.

Третім суттєвим фактором, який спонукав започаткувати проповану серію підручників є те, що об'єктивно визріла ситуація, коли фахівцям та науковцям треба подати чіткий сигнал щодо науково-методологічного осмислення та викладення базових знань галузі комп'ютингу як освітньо-наукової, виробничо-економічної та сервісно-обслуговувальної сфери.

Читач, безсумнівно, зверне увагу, на нашу послідовну промоцію нового терміну КОМП'ЮТИНГ (computing, англ.), який є вдалим та комплексно узагальнювальним для означення галузі знань, науки, виробництва, надання відповідних послуг та сервісів. Видається доречним подати ретроспективу як самого терміну комп'ютинг, так і широкої освітньої, наукової, бізнесової та виробничої сфери діяльності, що іменується комп'ютигом.

Гносеологічний аналіз подальшого формування інженерного рівня сфери КОМП'ЮТИНГУ невідворотно веде до структурного подання базових типів інженерій, які трактуються у класичному розумінні. ІНЖЕНЕРІЯ (майстерний – від лат. *ingeniosus*) – це наука про проектування та побудову (чит. створення) об'єктів певної природи. У цьому контексті природними для сфери КОМП'ЮТИНГУ є декілька видів інженерії. Мова йтиме про:

КОМП'ЮТЕРНУ ІНЖЕНЕРІЮ (computer engineering, англ.), яка охоплює проблематику проектування та створення об'єктів комп'ютерної техніки;

ПРОГРАМНУ ІНЖЕНЕРІЮ (software engineering, англ.), яка опікується проблематикою проектування та створення об'єктів, що іменуються програмними продуктами;

ІНЖЕНЕРІЮ ДАНИХ ТА ЗНАНЬ (data & knowledge engineering, англ.), інженерія, яка опікується проектуванням та створенням інформаційних продуктів; інженерію, яка опікується проектуванням та створенням міжкомпонентних (інтерфейсних) взаємозв'язків та формуванням цілісних системних об'єктів, усе частіше іменують СИСТЕМНОЮ ІНЖЕНЕРІЄЮ (systems engineering, англ.).

У разі такого структурно–класифікаційного подання видів інженерій сфери комп'ютерингу зазначимо, що кожен з них у цьому трактуванні є „відповідальним” за певний тип забезпечення, а саме апаратного (hardware, англ.), програмного (software, англ.), інформаційного (dataware, англ.) та між компонентного (middleware, англ.). Інформаційну технологію (ІТ) можна трактувати як певну точку в чотирирівимірному просторі зазначених інженерій. При цьому слід обов'язково зважити на певну частку наближення та інтерпретації цього простору як дискретного та неметричного. Інформаційні технології (ІТ) використовують комп'ютерні та програмні засоби для реалізації процесів відбору, реєстрації, подання, збереження, опрацювання, захисту та передавання інформації – інформаційного ресурсу у формі даних та знань – з метою створення інформаційних продуктів.

Автори підручників серії «КОМП'ЮТИНГ» пропонують значний перелік навчальних дисциплін, котрі, з одного боку, включаються до сфери комп'ютерингу за означенням, а, з іншого боку, їх предмет ще не знайшов якісного висвітлення у вітчизняній навчальній літературі для вищої школи. Структурно серія подається узагальненими профілями як то:

- фундаментальні проблеми комп'ютерингу;
- комп'ютерні науки;
- комп'ютерна інженерія;
- програмна інженерія;
- інженерія даних та знань;
- системна інженерія;
- інформаційні технології та системи.

Ми розуміємо, що чітка завершена будівля комп'ютерингу з'явиться лише в перспективі, а наша праця буде подаватись як активний труд будівничих з якнайшвидшого втілення в життя проекту цієї, без перебільшення, грандіозної будівлі сучасного інформаційного суспільства. Я запрошую потенційних авторів долучитись до цього освітньо–наукового проекту, а шановних читачів виступити в ролі творчих критиків та опонентів. Буду вдячний за Ваші побажання, зауваження та пропозиції.

З глибокою повагою науковий редактор серії підручників «КОМП'ЮТИНГ», д.т.н., професор Володимир Пасічник.

Вступ

Аналіз даних – це розділ математики, що займається розробкою методів обробки даних незалежно від їх природи.

Можна виділити такі етапи аналізу даних: отримання даних, обробка, аналіз та інтерпретація результатів обробки. Тобто процес отримання самих даних не на стільки важливий, як зробити на їх основі правильні висновки.

Аналіз даних можна вважати прикладним розділом математичної статистики, проте потрібно підкреслити, що аналіз даних охоплює обробку як кількісних, так і якісних даних. При чому, не обов'язково використовуються імовірнісні моделі при описі об'єктів, явищ та процесів що досліджуються.

Для обробки кількісних даних використовуються такі статистичні аналізи: розвідувальний аналіз; кореляційний аналіз; дисперсійний аналіз; регресійний аналіз; коваріаційний аналіз; факторний аналіз; дискримінаційний аналіз; кластерний аналіз; аналіз часових рядів.

Ці аналізи детально розглянуті у навчальному посібнику В.Є.Бахрушина Методи аналізу даних (Бахрушин В.Є. Методи аналізу даних: навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя: КПУ, 2011. – 268 с.) та інших книгах. Тому у нашому посібнику ми зупинимось лише на багатовимірних аналізах (факторному, дискримінаційному та кластерному – розділи 2-4).

Для якісного аналізу даних використовується так званий „інтелектуальний аналіз даних” (Data Mining). Методами такого аналізу є так звані методи машинного навчання, зокрема дерева рішень, нейронні мережі. Ці методи детально розглянуті у розділах 5-8 навчального посібника.

Що стосується аналізу знань, то сучасний підхід проведення такого аналізу ґрунтується на онтологічному інжинірингу. Онтологічний підхід до аналізу знань розглядається в розділах 9-10 цього навчальноо посібника.

Однак, насамперед ми зупинимось на основних поняттях, означеннях, які будуть використовуватись в цьому навчальному посібнику (див. розділ 1).

В.В. Литвин, В.В. Пасічник, Ю.В. Нікольський

АНАЛІЗ ДАНИХ ТА ЗНАНЬ

Навчальний посібник

Формат 60x84/16. Папір друк. цифровий.
Гарнітура Times New Roman

Умовн. друк. арк. 22,43.

ПП «Магнолія 2006»

м. Львів-53, 79053, Україна, тел.+380503701957

e-mail: magnol06@ukr.net

Свідоцтво про внесення суб'єкта видавничої справи
до Державного реєстру видавців, виготівників і розповсюджувачів видавничої
продукції: серія ДК № 2534 від 21.06.2006 року,
видане Державним комітетом інформаційної політики,
телебачення та радіомовлення України

Надруковано у друкарні видавництва «Магнолія 2006»