МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД
УЖГОРОДСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
Факультет математики та цифрових технологій
Кафедра теорії ймовірностей і математичного аналізу

Тегза А.М.

КОНСПЕКТ ЛЕКЦІЙ
З КУРСУ "Інтелектуальні технології Data Mining"

для студентів-магістрів математичного факультету

Ужгород – 2020

Зміст

Вступ

Інтелектуальний аналіз даних (ІАД, Data Mining), або розвідка даних - термін, що застосовується для опису здобуття знань у базах даних, дослідження даних, обробки зразків даних, очищення і збору даних. Це процес виявлення кореляції, тенденцій, шаблонів, зв'язків і категорій.

Термін Data Mining дістав назву від двох понять: дані - data і переробка сирого матеріалу (гірської руди) - mining.

Data Mining - предметна область"що виникла і розвивається на базі таких наук, як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних тощо.

Виникнення і розвиток Data Mining зумовлені різними факторами, серед яких вирізняємо основні: вдосконалення програмно-апаратного забезпечення, вдосконалення технологій зберігання і запису даних, накопичення великої кількості ретроспективних даних, вдосконалення алгоритмів обробки інформації.

Сутність і мету технології Data Mining можна описати так: це технологія, призначена для пошуку у великих інформаційних масивах даних неочевидних, об'єктивних, корисних на практиці закономірностей. ІАД здійснюється за допомогою використання технологій розпізнавання шаблонів, а також статистичних і математичних методів.

При розвідці даних багаторазово виконуються операції і перетворення над "сирими"даними (відбір ознак, стратифікація, кластеризація, візуалізація і регресія), що призначені для знаходження:

– структур, які інтуїтивно зрозумілі для людей і краще розкривають суть бізнес-процесів, що лежать в основі їх протікання;

– моделей, які можуть передбачити результат або значення певних ситуацій, використовуючи історичні або суб'єктивні дані.

Інтелектуальний аналіз даних - процес автоматичного пошуку прихованих закономірностей або взаємозв'язків між змінними у великих масивах необроблених даних, що поділяється на задачі класифікації, моделювання і прогнозування. Класичне визначення цього терміна дав у 1996 р. один із засновників цього напряму Г. П'ятецький-Шапіро.

Data Mining - це процес виявлення у необроблених даних раніше невідомих нетривіальних, практично корисних і доступних інтерпретацій знань, необхідних для прийняття рішень у різних сферах діяльності.

За визначенням SAS Institute, Data Mining - це процес виділення, дослідження і моделювання великих обсягів даних для виявлення невідомих до цього структур з метою досягнення переваг у бізнесі.

За визначенням Gartner Group, Data Mining - це процес, мета якого - виявляти нові кореляції, зразки і тенденції у результаті просіювання великого обсягу даних з використанням методик розпізнавання зразків і статистичних та математичних методів.

В основу технології Data Mining покладено концепцію шаблонів (patterns), що є закономірностями, які властиві вибіркам даних і можуть бути подані у формі, зрозумілій людині.

Задачі Data Mining:

1. Класифікація (Classification) - виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних - класи; за цими ознаками новий об'єкт можна віднести до того або іншого класу. Для вирішення задач класифікації можуть використовуватися методи: найближчий сусід (Nearest Neighbor); к-найближчий сусід (k-Nearest Neighbor); байєсовські мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (neural networks).

2. Кластеризація (Clustering) - результатом її є поділ об'єктів на групи.

3. Асоціація (Associations) - знаходять закономірності між пов'язаними подіями у наборі даних. Найбільш відомий алгоритм рішення задачі пошуку асоціативних правил - алгоритм Аргіогі.

4. Послідовність (Sequence), або послідовна асоціація (sequential association), - дає можливість знайти часові закономірності між транзакціями. Завдання послідовності подібне до асоціації, але її метою є встановлення закономірностей між подіями, пов'язаними за часом, тобто послідовність визначається високою ймовірністю ланцюжка пов'язаних за часом подій.

5. Прогнозування (Forecasting) - на основі особливостей історичних даних оцінюються майбутні значення показників. Застосовуються методи математичної статистики, нейронні мережі тощо.

6. Визначення відхилень (Deviation Detection), аналіз відхилень або викидів - виявлення й аналіз даних, що найбільше відрізняються від загальної чисельності даних, виявлення нехарактерних шаблонів.

7. Оцінювання (Estimation) - зводиться до прогнозу безперервних значень ознак.

8. Аналіз зв'язків (Link Analysis) - задача знаходження залежностей у наборі даних.

9. Візуалізація (Visualization, Graph Mining) - створюється графічний образ аналізованих даних. Для вирішення задач візуалізації використовуються графічні методи, що показують наявність закономірностей в даних.

10. Підбивання підсумків (Summarization) - опис конкретних груп об'єктів за допомогою аналізованого набору даних.

Зазначені вище задачі поділяються за призначенням на описові і предиктивні.

Описові, або дескриптивні (descriptive), задачі пов'язані з поліпшенням розуміння аналізованих даних. Ключовий момент у таких моделях - простота і прозорість результатів для сприйняття людиною. До такого типу задач належать кластеризація і пошук асоціативних правил.

Рішення предиктивних (predictive), або прогнозуючих, задач поділяється на два етапи. На першому етапі на підставі набору даних з відомими результатами будується модель. На другому етапі вона використовується для прогнозу результатів на підставі нових наборів даних. Вимагається, щоб побудовані моделі працювали максимально точно. До цього типу задач відносять задачі класифікації і регресії. Сюди можна віднести і задачу пошуку асоціативних правил, якщо результати її рішення можуть бути використані для прогнозу появи деяких подій.

За способами рішення задачі поділяють на такі, що вирішують за допомогою вчителя і без його допомоги. Категорія навчання з учителем представлена такими задачами Data

Mining: класифікація, оцінка, прогнозування, категорія навчання; без учителя - задачею кластеризації.

У випадку рішення з допомогою вчителя задача аналізу даних розв'язується у кілька етапів. Спочатку за допомогою конкретного алгоритму Data Mining будується модель аналізованих даних - класифікатор. Потім класифікатор піддається навчанню. Іншими словами, перевіряється якість його роботи і, якщо вона незадовільна, відбувається додаткове навчання класифікатора. Так продовжується доти, доки не буде досягнуто необхідного рівня якості або не стане зрозуміло, що обраний алгоритм не працює коректно з даними, або дані не мають структури, здатної проявитися. До цього типу задач відносять задачі класифікації і регресії.

Рішення без допомоги вчителя об'єднує задачі, що виявляють описові моделі, наприклад, закономірності в часових рядах макропоказників. Очевидно, якщо ці закономірності існують, то модель має їх проявити. Перевагою цих задач є можливість їх рішення без будь-яких попередніх знань про дані аналізу. До них належать кластеризація і пошук асоціативних правил.

Класифікації і регресії.

Під час аналізу часто необхідно визначити, до якого з відомих класів відносять досліджувані об'єкти, тобто як їх класифікувати. Задачу класифікації розглядають як задачу визначення значення одного з параметрів аналізованого об'єкта на підставі значень інших параметрів. Досліджуваний параметр часто називають залежною змінною, а параметри, що беруть участь у його визначенні - незалежними змінними. Задача класифікації і регресії розв'язується у два етапи. На першому виділяється навчальна вибірка. До неї входять об'єкти, для яких відомі значення як незалежних, так і залежних змінних. На підставі навчальної вибірки будується модель визначення значення залежної змінної. її часто називають функцією класифікації або регресії. Для отримання максимально точної функції до навчальної вибірки пред'являються такі основні вимоги: о кількість об'єктів, що входять до вибірки, має бути досить великою; о до вибірки мають входити об'єкти, що представляють усі можливі класи у задачі класифікації або всю область значень у задачі регресії; о для кожного класу в задачі класифікації або кожного інтервалу області значень у задачі регресії вибірка має містити достатню кількість об'єктів. На другому етапі побудовану модель застосовують до об'єктів аналізу. Задача класифікації і регресії має геометричну інтерпретацію. Пошук асоціативних правил Пошук асоціативних правил є поширеним застосуванням Data Mining. Суть задачі полягає у визначенні наборів об'єктів, що часто зустрічаються, в інформаційних масивах. Ця задача є окремим випадком задачі класифікації. При аналізі потрібною є інформація про послідовність подій, що відбуваються. При виявленні закономірностей у таких послідовностях можна з певною часткою ймовірності передбачати появу подій у майбутньому, що дає змогу приймати правильніші рішення. Така задача є різновидом задачі пошуку асоціативних правил і називається секвенціональним аналізом. Він широко використовується, наприклад, в телекомунікаційних компаніях для аналізу даних про аварії на різних вузлах мережі.

Кластеризації.

Задача кластеризації полягає в поділі об'єктів на групи подібних об'єктів, що нази-

ваються кластерами (cluster), тобто сукупності осіб, предметів. Задачі поділу множини елементів на кластери називають кластерним аналізом.

Кластеризація може застосовуватися практично в будь-якій сфері, де необхідне дослідження експериментальних або статистичних даних. Для задачі кластеризації характерна відсутність яких-небудь відмінностей між змінними і об'єктами. Кластерний аналіз дає змогу розглядати досить великий обсяг інформації і різко скорочувати, стискати великі масиви інформації, робити їх компактними. Слід зазначити деякі особливості, властиві задачі кластеризації.

1) Рішення залежить від природи об'єктів даних (і їх атрибутів), а також від представлення кластерів і передбачуваних відношень об'єктів даних і кластерів. Так, необхідно враховувати такі властивості, як можливість/неможливість приналежності об'єктів кільком кластерам. Необхідне визначення самого поняття приналежності кластеру: однозначна ймовірність приналежності, нечітка ступінь приналежності.

2) Дані деталізуються для подальшої обробки, тобто необхідним є виявлення і використання формалізованих закономірностей або дистиляція шаблонів.

При технології дистиляції шаблонів один зразок (шаблон) інформації витягується з початкових даних і перетворюється у певні формальні конструкції, вид яких залежить від методу Data Mining. Цей процес відбувається на стадії вільного пошуку, у першої групи методів ця стадія - відсутня. На стадіях прогностичного моделювання і аналізу виключень використовуються результати стадії вільного пошуку. Методи цієї групи: логічні методи; методи візуалізації; методи крос-табуляції; методи, засновані на рівняннях.

Логічні методи, або методи логічної індукції, включають нечіткі запити й аналізи, символьні правила, дерева рішень, генетичні алгоритми. Методи цієї групи придатні для інтерпретації. Вони підтримують знайдені закономірності у прозорому вигляді з погляду користувача. Методи крос-табуляції забезпечують пошук шаблонів. Методи на основі рівнянь виражають наявні закономірності у вигляді математичних виразів - рівнянь. Основні методи цієї групи: статистичні методи і нейронні мережі. Статистичні методи найчастіше застосовуються для вирішення задач прогнозування. Є багато методів статистичного аналізу даних, наприклад, кореляційно-регресійний аналіз, кореляція рядів динаміки, виявлення тенденцій динамічних рядів, гармонійний аналіз.

Інша класифікація поділяє все різноманіття методів Data Mining на дві групи: статистичні і кібернетичні методи. Ця схема поділу заснована на різних підходах щодо навчання математичним моделям.

Статистичні методи Data Mining. Ці методи включають: попередній аналіз природи статистичних даних (перевірка гіпотез стаціонарності, нормальності, незалежності, однорідності, оцінка виду функції розподілу, її параметрів); виявлення зв'язків і закономірностей (лінійний і нелінійний регресійний аналіз, кореляційний аналіз); багатовимірний статистичний аналіз (лінійний і нелінійний дискримінантний аналіз, кластерний аналіз, компонентний аналіз, факторний аналіз); динамічні моделі і прогноз на основі часових рядів. Статистичні методи Data Mining поділяються на чотири групи методів: дескриптивний аналіз і опис початкових даних; аналіз зв'язків (кореляційний і регресійний аналіз, факторний аналіз, дисперсійний аналіз); багатовимірний статистичний аналіз (компонентний

аналіз, дискримінантний аналіз, багатовимірний регресійний аналіз, канонічні кореляції); аналіз часових рядів (динамічні моделі і прогнозування).

Кібернетичні методи Data Mining. До цієї групи належать такі методи: еволюційне програмування; асоціативна пам'ять (пошук аналогів, прототипів); нечітка логіка; дерева рішень; системи обробки експертних знань, штучні нейронні мережі (розпізнавання, кластеризація, прогноз); генетичні алгоритми (оптимізація).

Нейронні мережі (Neural Networks)- це клас моделей, що базуються на аналогії з роботою мозку людини і призначаються для вирішення різноманітних задач аналізу даних після проходження етапу навчання на даних. Нейронні мережі - це моделі біологічних нейронних мереж мозку, в яких нейрони імітуються однотипними елементами (штучними нейронами). Нейронна мережа може бути представлена направленим графом зі зваженими зв'язками, у якому штучні нейрони є вершинами, а синаптичні зв'язки - дугами. Серед сфер застосування нейронних мереж - автоматизація процесів розпізнавання образів, прогнозування показників діяльності підприємства, медична діагностика, прогнозування, адаптивне управління, створення експертних систем, організація асоціативної пам'яті, оброблення аналогових і цифрових сигналів, синтез й ідентифікація електронних систем. За допомогою нейронних мереж можна, наприклад, передбачати обсяги продажу виробів, показники фінансового ринку, розпізнавати сигнали, конструювати самонавчальні системи. Нейронна мережа є сукупністю нейронів, з яких складаються шари. У кожному шарі нейрони пов'язані з нейронами.

1 Методи вибіркових спостережень

Щоб вивчити довільну сукупність, треба її охарактеризувати різного роду зведеними ознаками. Зведені ознаки характеризують не окремі ознаки, а сукупність в цілому або певної її частини. У цьому зв'язку масове спостереження, тобто реєстрація кожної конкретної одиниці, має зміст лише як проміжний етап для одержання зведених ознак.

Статистиків давно цікавить питання, як спростити цей проміжний етап, тобто як перейти від фіксації кожної ознаки, що входить до досліджуваної сукупності, до часткової. Тобто мова йде про перехід від суцільного спостереження до часткового.

Отже, дослідження можна здійснювати двома шляхами: вивчати всі одиниці спостереження всього масиву або лише їх частину, відібрану за певними науковими принципами. У першому випадку здійснюється суцільне спостереження, в другому -несуцільне, яке називають вибірковим.

Вибірковими даними користуються досить широко в різних сферах людської діяльності. Наприклад, для оцінки якості зерна або молока немає необхідності в обстеженні всього обсягу продукції, досить лише взяти певну кількість проб. Незначна кількість дослідів виявляється достатньою, наприклад, для встановлення зараження зерна шкідниками, встановлення якості борошна, олійності соняшнику і т.ін.

Строгі означення генеральної сукупності (всієї досліджуваної сукупності або популяції), вибірки (вибіркової сукупності), характеристик вибірки (параметрів або характеристик популяції) вивчались в курсі математичної статистики.

Схема процесу проведення вибіркового обстеження може бути поділена на декілька послідовних етапів, починаючи від попередньої підготовки і закінчуючи перевіркою всього процесу реалізації вибіркового плану.
Розглянемо ці стадії детальніше:

1. Ціль дослідження.
   Тобто спершу чітко потрібно визначити, яку інформацію необхідно одержати і які параметри треба оцінити під час проведення спостереження.

2. План процедури одержання вибірки.
   Тобто потрібно спланувати роботу таким чином, щоб одержана вибірка містила всю необхідну інформацію, а також не була більшою ніж необхідний мінімум, оскільки це затрата лишніх ресурсів і коштів. Результати обстеження, тобто одержані оцінки параметрів, завжди містять деякі відхилення від невідомих дійсних значень. Ці похибки можуть бути зменшені і підвищено точність оцінок за рахунок використання інших, більш складних методів відбору і збільшення об'єму вибірки для обстеження. Але таке рішення має приймати замовник обстеження, оскільки підвищення точності вимагає збільшення витрати як часу так і коштів.

3. Збір та обробка даних.
   Важливим є вибір форми реєстрації даних, практичного методу їх одержання, що може суттєво вплинути на якість та швидкість процесу збору даних. Наприклад,

соціологічне обстеження може бути виконано за допомогою прямого опитування, розсилки листів з питаннями, телефонування і т.ін. Обробка даних полягає у знаходженні оцінок параметрів і одержання висновків щодо надійності і точності цих оцінок.

Одержання вибірки із всієї сукупності може бути здійснене одним із чотирьох способів або частіше всього застосуванням комбінації цих способів. Кожний спосіб містить свої підвиди, тому розглянемо детальніше їх.

Простий випадковий вибір (ПВВ)

Нехай маємо N-елементну генеральну сукупність, з якої потрібно вибрати n елементів. Причому з курсу математичної статистики відомо, що вибірка буде репрезентативною, якщо ймовірність вибору всіх n-елементних підмножин буде однаковою. Тому це обов'язково треба забезпечити при відборі.

ПВВ отримується послідовним випадковим вибором об'єктів по одному, поки не наберемо n елементів. Але це можна здійснити двома способами: без повернення об'єкта до генеральної сукупності, і – з поверненням.

Всі елементи сукупності нумеруються числами від 1 до N і випадковим чином вибирається n з них по одному, причому якщо вибраний елемент не повертається назад, то одержана множина називається ПВВ без повернення. Якщо вибраний на кожному кроці елемент повертається назад, то отримана вибірка називається ПВВ з поверненням (але на практиці він рідко застосовується).

Розглянемо вибір Бернуллі, як приклад алгоритму реалізації ПВВ без повернення. Розглянемо генеральну сукупність $x_1, x_2, \ldots, x_N$, яка є реалізацією рівномірно розподіленої величини на проміжку $(0, 1)$. Розглянемо деяке фіксоване число $p \in (0, 1)$. Якщо $x_k < p$, то даний елемент вибрано, $k = 1, ..., N$. При такій побудові вибірки число вибраних елементів має біноміальний розподіл з параметрами N і p.

Межі, в яких змінюється випадковий об'єм вибірки n для відбору Бернуллі, можна оцінити за допомогою наступного довірчого інтервалу (γ – надійність):

$$(Np - z_{1-\gamma/2}\sqrt{Np(1-p)};\quad Np - z_{1-\gamma/2}\sqrt{Np(1-p)}).$$

Наприклад, для сукупності $N = 10000$ елементів і ймовірністю включення $p = 0.2$ 95% довірчий інтервал для n дорівнюватиме 2000 ± 78.

Систематичний вибір (СВ).

Вибіркове обстеження з систематичним відбором іноді розглядають як деяке наближення ПВВ, коли не існує повного переліку або списку всієї сукупності, або коли цей список не є впорядкованим за якоюсь ознакою, тобто коли елементи записано у довільному випадковому порядку.

Алгоритм класичного СВ:

1. Нехай маємо сукупність з N елементів. Для одержання СВ вибираємо бажаний об'єм вибірки n, вибиремо $a = \left[\frac{N}{n}\right]$.

2. Випадковим чином вибираємо ціле число $r \in (1, a)$.

3. Формуємо вибірку: $x_r, x_{r+a}, x_{r+2a}, \ldots, X_{r+(n-1)a}$.

Але потрібно пам'ятати, що вибірка буде репрезентативною, якщо сукупність не буде записана в деякому періодичному або циклічному порядку, тобто сукупність має бути випадковим чином змішана.

Алгоритм СВ за методом дробного інтервалу:

1. Вибираємо $a = \frac{N}{n}$ – деяке дробове число.

2. З рівномірного розподілу вибираємо випадкове число $\lambda \in (0, a)$.

3. Формуємо вибірку: $x_{[\lambda]}, x_{[\lambda+a]}, x_{[\lambda+2a]}, \ldots, X_{[\lambda+(n-1)a]}$.

Стратифіковані випадкові вибірки (СВВ).

Стратифікована вибірка одержується наступним чином. Вся генеральна сукупність ділиться на h множин (їх називають стратами або кластерами). Страти взаємно не перетинаються і їх об'єднання утворює всю сукупність. З кожної страти вибирається незалежна випадкова вибірка за будь-яким з попередніх методів, всі дані об'єднуються і отримана інформація використовується для обчислення оцінок для всієї сукупності.

Стратифіковане дослідження може бути більш зручним у практичному виконанні і мати меншу вартість проведення і дає більш точні оцінки для всієї сукупності, тобто оцінки з меншою дисперсією.

1.1 Завдання для лабораторної роботи №1

Використовуючи один з методів випадкового відбору, зробити m виборок з генеральної сукупності деякої ознаки з файлу student-mat.csv. Оцінити швидкість роботи відповідного алгоритму. Обчислити середнє значення ознаки всієї генеральної сукупності та порівняти його з середніми значеннями зроблених виборок, знайти стандартне відхилення цієї ознаки.

Вар.	Метод	Досліджувана ознака	m	об'єм вибірки
1	ППВ Бернуллі	absences (кількість шкільних пропусків)	10	$p = 0.05$
2	Класичний СВ	G1 (рейтинг за перший курс)	11	8%
3	Метод дробного інтервалу	G2 (рейтинг за другий курс)	12	5%
4	ППВ Бернуллі	G3 (рейтинг за третій курс)	13	$p = 0.04$
5	Стратифікований ВВ (ПВВБ) стратами є школи	traveltime (час добирання до школи)	5; 5	$p_1 = 0.05$ $p_2 = 0.07$
6	СВВ (СВ) стратами є місцевість проживання	health (стан здоров'я)	5; 5	5% з кожної страти
7	СВВ (метод дроб. інт.) стратами є стать	G2 (рейтинг за другий курс)	5; 5	7% з кожної страти

2 Регресійні моделі та їх методи як складові елементи методів ідентифікації та прогнозування у задачах штучного інтелекту. Метод найменших квадратів. Парна лінійна регресія.

Методи ідентифікації та прогнозування найчастіше не є самостійними методами, що застосовуються при розв'язанні слабкоструктуризованих та важкоформалізованих задач штучного інтелекту. Водночас більшість методів, які використовуються при розв'язуванні таких задач, базуються на регресійних моделях та методах або використовують їх в якості складових елементів. Важливою їх особливістю є розвинений математичний апарат, за допомогою якого можна оцінювати якість побудованих моделей, зокрема їх точність та адекватність.Побудова та дослідження трьох видів моделей: парної лінійної регресії, множинної лінійної регресії і деяких типів нелінійної парної та множинної регресії базується на використанні методу найменших квадратів.

Метод найменших квадратів є тим класичним методом, з якого раціонально починати представлення і обговорення методів прогнозування. Він призначений для оцінки невідомих величин за результатами вимірювань чи експериментів , що містять випадкові помилки і застосовується для наближеного представлення заданої функції іншими (більш простими) функціями при обробці спостережень. МНК запропонований К.Гауссом і А. Лежандром.

2.1 Проста лінійна регресія.

Регресійний аналіз вивчає характер залежності між величинами. За допомогою регресійного аналізу створюються математичні моделі економічних (фізичних, біологічних та ін.) процесів на основі спостережуваних (статистичних) значень відповідних показників. Задача регресійного аналізу ставиться наступним чином. Нехай є два економічних показника X і Y, які характеризують економічний процес. Наприклад, Y – рентабельність продукції, X – рівень інфляції. Виникає питання чи рентабельність залежить від інфляції.

Взагальному, показник Y називають вихідною або ендогенною змінною, а X – вхідною або екзогенною змінною. При побудові математичної моделі того чи іншого процесу деякий параметр процесу (ендогенна змінна) подають як функцію деяких зовнішніх факторів (екзогенних змінних). Деякі з цих факторів є суттєвими і чинять значний вплив на параметр процесу, а інші є несуттєвими, бо їх вплив є незначним. Як правило, суттєвих факторів є всього кілька, а несуттєвих — досить багато.Тому не можна повністю нехтувати впливом багатьох несуттєвих факторів на результуючі показники процесу. Позначають результуючий показник процесу через Y, набір суттєвих факторів як $X = (x_1, x_2, \ldots, x_m)$ і набір несуттєвих факторів як $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k)$. Загальний вигляд регресійної залежності:

$$Y = F(x_1, \ldots, x_m, \varepsilon_1, \ldots, \varepsilon_k),$$

де F — функція регресії ; або ще представляють так:

$$Y = f(X) + u(t), \tag{1}$$

де $f(X)$ — деяка функція, яка виражає залежність змінної Y від фактора X, $u(t)$ — випадкова функція, що характеризує вплив неврахованих факторів, t — час спостереження. Зазвичай, вважають, що $u(t) \sim N(0, \sigma)$, де $Du(t) = \sigma^2 = const$ і $cov(u(t),\ u(t+s)) = 0$, $s > 0$.

Слід розрізняти крос-секційну регресію та регресію часових рядів. Крос-секційна регресія відображає зв'язок ендогенною та екзогенними змінними в один і той же момент часу. Тому дані для всіх змінних вимірюють одномоментно. А регресія часових рядів відображає зв'язок між змінними протягом певного часу і тому дані для кожної змінної вимірюють періодично, у певні послідовні моменти часу, протягом деякого часового інтервалу. Рівняння

$$\hat{Y} = f(X), \tag{2}$$

називають рівнянням простої регресії, а функцію $f(X)$ — функцією регресії.

Із (1) і висунутих припущень випливає, що $Y(X)$ — є випадковою функцією з $M(Y(X)) = f(X)$. Якщо розглядати X як випадкову величину, то $M(Y/X) = f(X)$, тобто умовне математичне сподівання Y за умови, що X прийняла певне значення, рівне $f(X)$.

Якщо функція $f(X)$ – є лінійною і залежною тільки від однієї екзогенної змінної, то рівняння (2) матиме вигляд:

$$\hat{Y} = a_0 + a_1 x \tag{3}$$

і називають рівнянням простої парної лінійної регресії. Якщо функція $f(X)$ – є нелінійною функцією, то рівняння (2) називають рівнянням нелінійної регресії. При розгляді моделі (3) коефіцієнти a_0 і a_1 вибирають так, щоб функція (3) найкращим чином наближала значення y_i за спостережуваними даними $(x_i,\ y_i)$.

Найкращим методом оцінки коефіцієнтів моделі (3) є метод найменших квадратів, при якому коефіцієнти a_0 і a_1 шукаються із задачі:

$$min \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2.$$

Розв'язком цієї задачі є коефіцієнти:

$$\begin{aligned} a_0 &= \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{n \sum x_i^2 - (\sum x_i)^2} \\ a_1 &= \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}. \end{aligned} \tag{4}$$

Або використовуючи вибіркові середні $\bar{x}$ і $\bar{y}$ і позначаючи $\overline{xy} = \frac{1}{n}\sum x_i y_i$, $\overline{x^2} = \frac{1}{n}\sum x_i^2$, $cov(x,\ y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$, $\bar{\sigma}_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$. Тоді (4) матиме вигляд:

$$a_0 = \bar{y} - a_1 \bar{x}, \qquad a_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{cov(x,\ y)}{\bar{\sigma}_x^2}. \tag{5}$$

Коефіцієнти a_0 і a_1 визначені по формулам (4) або (5) називають коефіцієнтами простої лінійної регресії.

Величини $e_i = y_i - \hat{y}_i = y_i - a_0 - a_1 x_i$, $i = \overline{1,n}$ називають залишками регресії. Вони показують на яку величину регресія $\hat{y}_i$ відрізняється від реального значення y_i. При оцінюванні коефіцієнтів за методом найменших квадратів значення суми квадратів залишків є мінімальним.

Однією з основних задач регресійного аналізу є отримання прогнозних значень показників X і Y. Якщо значення x_j, $j = n+1,\ n+2, \ldots, N$ являються очікуваними значеннями фактора X, то значення $\hat{y}_i$, що визначаються із рівняння (3), будуть відповідними прогнозованими значеннями показника Y.

2.2 Оцінка якості моделі простої парної лінійної регресії

Для оцінки точності моделі (3) використовується ряд критеріїв, у яких застосовуються такі характеристики:

1. коефіцієнт кореляції;
2. коефіцієнт детермінації;
3. стандартна похибка регресії;
4. довірчі інтервали для коефіцієнтів регресії та для прогнозних значень;
5. усереднений коефіцієнт еластичності.

Розглянемо кожний з критеріїв більш детально.

1. Коефіцієнт кореляції r_{xy} використовують для оцінки тісноти зв'язку між показниками x і y:

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}. \tag{6}$$

Відомо, що $|r_{xy}| \leq 1$. Чим ближчим $|r_{xy}|$ є до 1, тим сильнішим є статистичний зв'язок між показниками. Якщо $r_{xy} = 0$, то зв'язок між показниками x і y відсутній. Якщо $r_{xy} > 0$, то кореляція є додатною, тобто при зростанні x статистично зростає і y. Якщо $r_{xy} < 0$, то кореляція є від'ємною, тобто при зростанні x показник y статистично спадає.

Вважається, що якщо $|r_{xy}| > 0.7$, то зв'язок між показниками x і y є високим і можна будувати просту регресію, якщо $|r_{xy}| < 0.4$, то зв'язок є слабим і замість x необхідно вибирати інший фактор для побудови простої регресії показника y або збільшити кількість спостережень.

Значущість обчисленого значення r_{xy} визначається за допомогою t-критерію Стьюдента:

$$t_{\text{сп}} = \sqrt{\frac{r_{xy}^2}{1 - r_{xy}^2}(n-2)}, \tag{7}$$

$t_{\text{кр}} = t_{\text{табл}}(\alpha,\ n-2)$, зазвичай беруть $\alpha = 0.05$ ($\beta = 1-\alpha$ — рівень довірчої ймовірності).

Якщо $t_{\text{сп}} > t_{\text{кр}}$, то основна гіпотеза H_0: $r_{xy} = 0$ відкидується, тобто значення r_{xy} є значущим і приймається гіпотеза про існування статистичного зв'язку між показниками, у протилежному випадку потрібно вибрати інший показник x.

2. Коефіцієнт детермінації R^2 служить для оцінки степені відповідності моделі фактичним даним:

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n} e_i^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum\limits_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}, \tag{8}$$

Величина E_{ss} (error sum of square) $E_{ss} = \sum\limits_{i=1}^{n}(\hat{y}_i - y_i)^2 = \sum\limits_{i=1}^{n} e_i^2$ означає розкид за рахунок випадкових відхилень від функції регресії.

Величина $R_{ss} = \sum\limits_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (regression sum of square) означає величину розкиду, яка обумовлена регресією; її називають варіацією регресії.

Величина $T_{ss} = \sum\limits_{i=1}^{n}(y_i - \bar{y})^2$ (total sum of square) означає величину розкиду значень y_i відносно середнього.

Має місце нерівність $0 < R^2 < 1$. Коефіцієнт детермінації R^2 показує, яку частину фактичної варіації змінної Y складає варіація регресії.

Чим ближчим R^2 є до 1, тим точнішою є модель лінійної регресії. Якщо $R^2 > 0.8$, то модель лінійної регресії вважається точною, якщо $R^2 < 0.5$, то модель є незадовільною, потрібно будувати нелінійну регресію або вибирати інший фактор x.

3. Стандартна похибка регресії обчислюється за формулою: $\bar{S} = \sqrt{\frac{1}{n-2}\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

Перевірка значущості простої лінійної регресії здійснюється за F-критерієм Фішера: $F = \frac{R^2(n-2)}{1-R^2}$. Якщо $F_{\text{сп}} > F_{\text{кр}}$, то приймається гіпотеза про існування лінійної регресії між показниками x і y, інакше потрібно будувати нелінійну регресію або вибирати інший фактор x.

4. Довірчі інтервали коефіцієнтів регресії при заданому рівні значущості визначаються за формулами:

$$\begin{aligned} &\left(a_0 - S_{a_0}\cdot t(\frac{\alpha}{2},\ n-2);\ a_0 + S_{a_0}\cdot t(\frac{\alpha}{2},\ n-2)\right), \\ &\left(a_1 - S_{a_1}\cdot t(\frac{\alpha}{2},\ n-2);\ a_1 + S_{a_1}\cdot t(\frac{\alpha}{2},\ n-2)\right) \end{aligned} \tag{9}$$

Стандартні похибки коефіцієнтів рівні:

$$S_{a_0} = \sqrt{\frac{\sum x_i^2 \sum(y_i - \bar{y})^2}{n(n-2)\sum(x_i - \bar{x})^2}}, \qquad S_{a_1} = \sqrt{\frac{\sum(y_i - \bar{y})^2}{(n-2)\sum(x_i - \bar{x})^2}}.$$

Довірчий інтервал для прогнозних значень регресії при заданому рівні значущості визначається за формулою: $(\hat{y}_i - V_i;\ \hat{y}_i + V_i)$, де $V_i = \bar{S} \cdot t(\alpha,\ n-2)\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$, $t(\alpha,\ n-2)$ – табличне значення критерію Стьюдента.

5. Усереднений коефіцієнт еластичності показує вплив змінної x на змінну y і визначається за формулою $\varepsilon = a_1 \frac{\bar{x}}{\bar{y}}$. Коефіцієнт еластичності показує на скільки процентів в середньому зміниться y при зміні x на 1%.

2.3 Завдання для лабораторної роботи №2

1. За спостережуваними даними вологості, що залежить від висоти побудувати рівняння парної лінійної регресії методом найменших квадратів. Графічно представити результати лінії регресії і спостережувані дані. Виконати інтерпретацію коефіцієнтів моделі, оцінити якість побудованої моделі , використовуючи різні характеристики (коефіцієнт кореляції; коефіцієнт детермінації; стандартну похибку регресії; довірчі інтервали для коефіцієнтів регресії та для прогнозних значень; усереднений коефіцієнт еластичності.) Дані брати з файлів

1	2	3	4	5
2610.txt	2705.txt	2606.txt	2607.txt	2504.txt
6	7	8	9	10
2402.txt	2306.txt	2312.txt	2209.txt	2302.txt

3 Лінійний множинний регресійний аналіз. Оцінювання якості та адекватності побудованої моделі.

3.1 Лінійний множинний регресійний аналіз

При розв'язуванні задач економічного аналізу і прогнозування часто потрібно визначити вплив на показник y більше ніж одного зв'язаних з ним факторів x_1, ..., x_m. При проведенні експеременті у такій множинній ситуації дослідник записує показники приладів про стан функції відгуку y і всіх факторів від яких вона залежить x_i. Результатами спостережень являються вже не два вектори-стовпці, а матриця результатів спостережень:

$$\begin{pmatrix} y_1 & x_{11} & x_{12} & \cdots & x_{1m} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2m} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ y_n & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

де n - число дослідів. Задача множинного регресійного аналізу полягає у побудові такого рівняння площини у $(m+1)$-мірному просторі, відхилення результатів спостереження y_i від якої були б мінімальними. Або, іншими словами, потрібно обчислити значення коефіцієнтів b_0, b_j у лінійному поліномі:

$$\hat{y} = b_0 + \sum_{j=1}^{m} b_j x_j \tag{10}$$

, таким чином, щоб сума $\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum\limits_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_m x_{im})^2$ була мінімальною.

Для відшукання мінімуму необхідно знайти часткові похідні по всім невідомим b_0, b_1, . . . , b_m і прирівняти їх до нуля. Одержимо систему рівнянь:

$$\begin{cases} nb_0 + b_1 \sum\limits_i x_{i1} + b_2 \sum\limits_i x_{i2} + \ldots + b_m \sum\limits_i x_{im} = \sum\limits_i y_i; \\ b_0 \sum\limits_i x_{i1} + b_1 \sum\limits_i x_{i1}^2 + b_2 \sum\limits_i x_{i1}x_{i2} + \ldots + b_m \sum\limits_i x_{i1}x_{im} = \sum\limits_i y_i x_{i1}; \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ b_0 \sum\limits_i x_{im} + b_1 \sum\limits_i x_{i1}x_{im} + b_2 \sum\limits_i x_{i2}x_{im} + \ldots + b_m \sum\limits_i x_{im}^2 = \sum\limits_i y_i x_{im} \end{cases}$$

або у матричній формі: $X^T X B = X^T Y$, де

$$B = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix}; \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix}; \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Для розв'язання системи рівнянь у матричній формі потрібно домножити її зліва на матрицю, обернену до матриці системи рівнянь, якщо така існує (тобто при $det\ X^T X \neq 0$): $(X^T X)^{(-1)} \cdot (X^T X)B = (X^T X)^{(-1)} X^T Y$.

Таким чином, розв'язком системи рівнянь у матричній формі буде $B = (X^T X)^{(-1)} X^T Y$.

Кожний коефіцієнт рівняння регресії можна знайти по формулі $b_j = \sum_{i=0}^{m} c_{ij} \sum_{i=1}^{n} y_i x_{ij}$, де c_{ij} – елементи матриці $(X^T X)^{(-1)}$.

У результаті проведення всіх цих операцій одержуємо поліном першого степеня $\hat{y} = b_0 + \sum_{j=1}^{m} b_j x_j$ з відомими коефіцієнтами b_0, b_j, $j = \overline{1, m}$. Цей поліном є апроксимацією функції $y = f(x_1, x_2, \ldots, x_m)$, вид якої невідомий. Вже при $n > 20$ і $m > 3$ розрахунки коефіцієнтів рівняння регресії вручну дуже важкі.

3.2 Оцінка якості моделі множинної регресії

Аналогічно до парної регресії, для оцінки точності моделі (10) використовується ряд критеріїв, у яких застосовуються такі характеристики:

1. парні коефіцієнти кореляції двох типів, їх значущість;
2. часткові коефіцієнти кореляції, їх значущість;
3. коефіцієнт множинної кореляції;
4. скорегований коефіцієнт детермінації;
5. оцінка значущості моделі;
6. оцінки значущості коефіцієнтів множинної регресії;
7. коефіцієнти еластичності, β- коефіцієнти, Δ-коефіцієнти;
8. аналіз залишків.

Розглянемо кожний з критеріїв більш детально.

1. Розрахунки зазвичай починають з обчислення парних коефіцієнтів кореляції, що характеризують тісноту зв'язку між двома величинами. У множинній ситуації обчислюють два типи коефіцієнтів кореляції:

   - r_{yx_j} — коефіцієнти, які визначають тісноту зв'язку між функцією відгуку y і одним із факторів x_j
   - $r_{x_i x_j}$ — коефіцієнти, які показують тісноту зв'язку між факторами x_i і x_j, $i, j = \overline{1, m}$.

   Формула для обчислення r_{yx_j} відрізняється від формули, що призначена для обчислення коефіцієнта парної кореляції, тільки індексом при x:

   $$r_{yx_j} = \frac{cov(y,\ x_j)}{\sigma_y \sigma_{x_j}} = \frac{\overline{x_j y} - \bar{x}_j \bar{y}}{\sqrt{\overline{x_j^2} - \bar{x}_j^2}\sqrt{\overline{y^2} - \bar{y}^2}}$$

   Коефіцієнт парної кореляції для показників x_i, x_j рахується за формулою:

   $$r_{x_i x_j} = \frac{cov(x_i,\ x_j)}{\sigma_{x_i} \sigma_{x_j}} = \frac{\overline{x_i x_j} - \bar{x}_i \bar{x}_j}{\sqrt{\overline{x_i^2} - \bar{x}_i^2}\sqrt{\overline{x_j^2} - \bar{x}_j^2}}$$

Значення парних коефіцієнтів кореляції змінюються від -1 до 1. Якщо $r_{yx_j} < 0$, то це значить, що x_j зменшується із збільшенням y. Якщо $r_{yx_j} > 0$, то x_j збільшується із збільшенням y.

Значущість парних коефіцієнтів кореляції можна перевірити за критерієм Стюдента: $t = \frac{r}{\bar{S}_r}$, де $\bar{S}_r$ — середньоквадратична похибка вибіркового парного коефіцієнта кореляції: $\bar{S}_r = \frac{\sqrt{1-r^2_{yx_j}}}{\sqrt{n-2}}$. Якщо $t_{\text{сп}} > t_{\text{кр}}(\alpha, n-2)$, то значення r_{yx_j} є значущим і казатимемо, що між функцією відгуку y і одним із факторів x_j прослідковується статистичний зв'язок.

Якщо один з коефіцієнтів $r_{x_ix_j} = 1$, то це значить, що фактори x_i і x_j функціонально (не статистично) зв'язані між собою і тоді доцільно один з них виключити з розгляду, причому залишити той фактор, у якого коефіцієнт r_{yx_j} більший.

2. Після обчислення всіх парних коефіцієнтів кореляції і виключення того чи іншого фактора, можна побудувати матрицю коефіцієнтів кореляції виду:

$$\begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} & \cdots & r_{yx_m} \\ r_{x_1y} & 1 & r_{x_1x_2} & \cdots & r_{x_1x_m} \\ r_{x_2y} & r_{x_2x_1} & 1 & \cdots & r_{x_2x_m} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{x_my} & r_{x_mx_1} & r_{x_mx_2} & \cdots & 1 \end{pmatrix} \tag{11}$$

Використовуючи матрицю (11) можна обчислити часткові коефіцієнти кореляції, які показують степінь впливу одного з факторів x_j на функцію відгуку y при умові, що інші фактори закріплені на постійному рівні. Формула для обчислення часткових коефіцієнтів кореляції така: $r^j_{y/x_1,x_2,\ldots,x_m} = \frac{D_{1j}}{\sqrt{D_{11}\cdot D_{jj}}}$, де D_{1j} — визначник матриці, який утворюється з матриці (11) викресленням 1-го рядка, j-го стовпця. Визначники D_{11} і D_{jj} обчислюють аналогічно.

Як і парні коефіцієнти, часткові коефіцієнти кореляції змінюються від -1 до 1.

Значущість і довірчий інтервал для коефіцієнта часткової кореляції визначаються так само, як для коефіцієнтів парної кореляції, але число степенів вільності рахують як $\nu = n - k - 2$, $k = m - 1$ - порядок часткового коефіцієнта парної кореляції.

3. Для вивчення тісноти зв'язку між функцією відгуку y і кількома факторами x_1, x_2, ... , x_m використовують коефіцієнт множинної кореляції R. Коефіцієнт множинної кореляції служить і для оцінки якості передбачення; Загалом $0 < R < 1$. Чим ближчим до 1 є R, тим якіснішим є передбачення (прогноз) даної моделі за експериментальними даними.

Для обчислення коефіцієнта множинної кореляції використовують матрицю (11): $R = \sqrt{1 - \frac{D}{D_{11}}}$.

Значущість коефіцієнта множинної кореляції перевіряють за t-критерієм Стьюдента $t_R = \frac{R}{\bar{S}_R}$, де $\bar{S}_R = \sqrt{\frac{1-R^2}{n-m-1}}$ – середньо-квадратична похибка коефіцієнта множинної кореляції.

4. Величину R^2 називають множинним коефіцієнтом детермінації; він показує яка частина дисперсії функції відгуку пояснюється варіацією лінійної комбінації вибраних факторів. Значення R^2 зростає з ростом числа змінних (факторів) у регресії, що не означає покращення якості прогнозу, тому вводять скорегований (adjusted) коефіцієнт детермінації: $R^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1}$. Його використання є більш коректним для порівняння регресій при зміні числа змінних (факторів).

5. Стандартна похибка регресії обчислюється за формулою: $\bar{S} = \sqrt{\frac{1}{n-m-1}\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

   Оцінка значущості моделі, тобто оцінка того наскільки вірною є гіпотеза про лінійність регресії між y і факторами x_j, здійснюється за F-критерієм Фішера. За спостережуваними значеннями визначається статистика: $F_{\text{сп}} = \frac{R^2(n-m-1)}{(1-R^2)m}$. Якщо $F_{\text{сп}} > F_{\text{кр}}(\alpha, n-1, n-m-1)$, то гіпотезу H_0: $\quad R = 0$ відкидаємо і кажемо, що модель є значущою і лінійний зв'язок між досліджуваними показниками є статистично значущим.

6. Оцінки значущості коефіцієнтів множинної регресії (крім вільного члена) здійснюються за допомогою t-критерію Стьюдента: $t_j = \frac{b_j}{\bar{S}\sqrt{a_{jj}}}$, де a_{jj} – діагональний елемент матриці $(X^TX)^{(-1)}$. Якщо $t_j > t_{\text{табл}}$, то j-вий коефіцієнт вважають значущим, інакше фактор, що відповідає даному коефіцієнту потрібно виключити з моделі.

   Довірчий інтервал для прогнозних значень лінії регресії визначається за формулою: $(\hat{y}_i - V_t;\ \hat{y}_i + V_t)$, де $V_t = \bar{S} \cdot t(\alpha, n-m-1)\sqrt{x_n^T(l) \cdot (X^TX)^{(-1)}x_n(l)}$, $x_n(l)$ – вектор-стовпець факторів для прогнозних значень часу ($l = n+1, n+2, \ldots$). Матриця $(X^TX)^{(-1)}$ відповідає спостережуваним значенням факторів.

7. Вплив факторів x_j на показник y оцінюється з допомогою коефіцієнтів еластичності ε_j і β- коефіцієнтів: $\varepsilon_j = b_j\frac{\bar{x}_j}{\bar{y}}$; $\beta_j = b_j\frac{S_j}{S_y}$, $j = \overline{1,m}$, де $S_j = \sqrt{\frac{1}{n-2}\sum\limits_{i=1}^{n}(x_{ij} - \bar{x})^2}$; $S_y = \sqrt{\frac{1}{n-2}\sum\limits_{i=1}^{n}(y_j - \bar{y})^2}$ – незміщені середньоквадратичні відхилення факторів x_j і y. Коефіцієнти еластичності ε_j показують на скільки відсотків зміниться значення змінної y при зміні x_j на 1%.

   Бета коефіцієнти β_j показують на яку частину середньоквадратичного відхилення зміниться y при зміні x_j на величину свого середньоквадратичного відхилення.

   Долю впливу j-го фактора у сумарному впливі всіх факторів на показник y оцінюють з допомогою Δ-коефіцієнтів: $\Delta_j = r_{yx_j}\frac{\beta_j}{R^2}$.

8. Аналіз залишків проводиться за допомогою статистики Дарбіна-Уотсона. Цей критерій дозволяє дослідити залежність між залишками. Якщо залишки суттєво корелюють між собою, то модель є неадекватною (порушено важливе припущення про незалежність похибок у регресійній моделі).

3.3 Тестування і усунення мультиколінеарності

Однією з умов правомірності застосування МНК є умова незалежності екзогенних змінних (або вхідних факторів). Існування залежності між ними називають мультиколінеарністю. Оскільки ця умова часто не виконується, то необхідно визначити рівень впливу специфікації залежності на оцінку параметрів моделі.

Алгоритмом повного дослідження мультиколінеарності є алгоритм Фаррара-Глобера. З його допомогою тестують три види мультиколінеарності:

1. У сукупності всіх факторів (критерій χ^2 Пірсона)
2. Кожного фактора з іншими (критерій Фішера).
3. Кожної пари факторів (критерій Ст'юдента).

Усі ці критерії при порівнянні з їх критичними значеннями дають змогу робити конкретні висновки щодо наявності чи відсутності мультиколінеарності незалежних змінних.

Для оцінки параметрів моделі, в яку входять мультиколінеарні змінні, також використовують метод головних компонент.

Опишемо ***алгоритм Феррара — Глобера.***

1. Стандартизація (нормалізація) змінних. Позначимо вектори незалежних змінних економетричної моделі через $X_1, X_2, \ldots, X_m$. Елементи стандартизованих векторів розрахуємо за формулою:
$$x_{ik}^* = \frac{x_{ik} - \bar{X}_k}{\sqrt{n\sigma_{X_k}^2}}, \tag{12}$$
де $\bar{X}_k$ – вибіркове середнє, а $\sigma_{X_k}^2$ – дисперсія k-ї екзогенної змінної, x_{ik} -елементи матриці:
$$X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \cdot & \cdot & \cdot & \cdot \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{pmatrix};$$
2. Знаходження кореляційної матриці (матриці моментів стандартизованої системи нормальних рівнянь):
$$R = (X^*)^T X^*, \tag{13}$$
3. Визначення критерію χ^2 Пірсона:
$$\chi^2 = -(n - 1 - \frac{1}{6}(2m + 5))ln|det(R)|, \tag{14}$$
тут m - кількість факторів, n - кількість спостережень. Порівнюємо це спостережуване значення із табличним критичним значенням при $\frac{m(m-1)}{2}$ ступенях вільності і рівні значущості α. Якщо $\chi^2 < \chi^2_{\text{кр}}$, то в масиві незалежних змінних не існує мультиколінеарності.

4. Визначення оберненої матриці:

$$C = R^{-1}. \tag{15}$$

5. Розрахунок F- критеріїв:

$$F_k = |c_{kk} - 1|\frac{n-m}{m-1}, \tag{16}$$

де c_{kk} — діагональні елементи матриці C. Фактичні значення критеріїв F_k порівнюються з табличними при $n-m$ і $m-1$ ступенях свободи і рівні значущості α. Якщо $F_k > F_{\text{табл}}$, то відповідна k-ва екзогенна змінна мультиколінеарна з іншими.

Коефіцієнт детермінації для кожної змінної розраховується таким чином: $R^2_{X_k} = 1 - \frac{1}{c_k k}$.

6. Знаходження часткових коефіцієнтів кореляції:

$$r_{kj} = \frac{-c_{kj}}{\sqrt{c_{kk}c_{jj}}}, \tag{17}$$

де c_{kj} — елемент матриці C, що заходиться в k-му рядку і j-му стовпці, , а c_{kk} і c_{jj} — діагональні елементи цієї матриці .

7. Розрахунок t критеріїв:

$$t_{kj} = \frac{r_{kj}\sqrt{n-m}}{\sqrt{1-r^2_{kj}}}. \tag{18}$$

Фактичні значення критеріїв t_{kj} порівнюються з табличними при $n-m$ ступенях свободи і рівні значущості α. Якщо $t_{kj} > t_{\text{табл}}$, то між екзогенними змінними X_k і X_j існує мультиколінеарність.

3.4 Алгоритм методу головних компонент

Метод головних компонент–один з основних способів зменшити розмірність даних, втративши найменшу кількість інформації. Був винайдений Карлом Пірсоном у 1901 році. Іноді МГК називають перетворенням Кархунена-Лоева або перетворенням Хотеллінга.

На практиці часто доводиться мати справу із задачами, в яких кількість факторів перевищує межі адекватного аналізу та інтерпретації. Тому замість множини вихідних факторів $X_1, X_2, \dots X_m$ розглядають іншу множину $Z_1, Z_2, \dots Z_n$, де $n < m$.

Причинами цього є:

– необхідність наочного представлення вихідних даних, що досягається їх проектуванням на спеціальним чином визначений одно-, двох- чи трьохвимірний простір;

– прагненням до лаконізму досліджуваних моделей, що одночасно дозволить спростити розрахунки та інтерпретацію моделей;

– необхідністю стиснення обсягів статистичної інформації.

Процедура визначення факторів $Z_1, Z_2, \dots Z_n$, базується на двох критеріях: перший – максимальне збереження вихідної інформації, що зосереджена у значеннях факторів $X_1, X_2, \dots X_m$, другий – максимальне використання інформації, що зосереджена у цих факторах відносно інших, зовнішніх показників.

Формально задача переходу до нового набору факторів є такою. Нехай $Z = Z(X)$ – деяка k-вимірна вектор-функція початкових факторів і $I_k(Z(X))$ – певним чином задана міра інформативності системи факторів

$$Z(X) = (Z_1(X), Z_2(X), \dots, Z_k(X)).$$

Задача полягає у визначенні такого набору факторів $\hat{Z}$, знайденого у класі F допустимих перетворень початкових факторів X, що $I_m(\hat{Z}(X)) = \max\limits_{Z \in F} I_m(Z(X))$. Припустимо, що перетворення F визначає можливі лінійні ортогональні нормовані комбінації початкових факторів, тобто

$$Z_j(X) = c_{j1}(X_1 - MX_1) + \dots + c_{jm}(X_m - MX_m);$$
$$\sum_{i=1}^{m} c_{ji}^2 = 1,\ j = \overline{1,m}; \quad \sum_{i=1}^{m} c_{ji}c_{ki} = 0,\ j,k = \overline{1,m},\ j \neq k.$$

Мірою інформативності є відношення

$$I_m(Z(X)) = \frac{DZ_1 + \dots + DZ_n}{DX_1 + \dots + DX_m},$$

Тоді вектор $\hat{Z}$ визначається як лінійна комбінація $\hat{Z} = AX$, де рядки матриці A задовільняють умові ортогональності. Конструктивна побудова елементів матриці A розглянута нижче.

Першою головною компонентою $Z_1(X)$ називається така нормовано-центрована лінійна комбінація початкових факторів, яка серед усіх інших таких комбінацій має найбільшу дисперсію.

k-ю головною компонентою досліджуваної системи факторів $X_1, X_2, \dots X_m$ називається така лінійна комбінація цих факторів, яка не корельована з $k-1$ попередніми головними компонентами і серед усіх інших таких комбінацій, що не корельовані з попередніми $k-1$ головними компонентами лінійних комбінацій, має найбільшу дисперсію.

1. Нормуємо та центруємо значення факторів:

$$x_{ik}^* = \frac{x_{ik} - \bar{X}_k}{\sqrt{n\sigma_{X_k}^2}}, \tag{19}$$

   де $\bar{X}_k$ – вибіркове середнє, а $\sigma_{X_k}^2$ – дисперсія k-ї екзогенної змінної.

2. Знаходження кореляційної матриці (матриці моментів стандартизованої системи нормальних рівнянь):

$$R = (X^*)^T X^*, \tag{20}$$

3. Знаходимо характеристичні числа матриці R з рівняння:

$$|R - \lambda E| = 0, \tag{21}$$

4. Впорядковуємо власні числа λ_k за абсолютним внеском головної компоненти в загальну дисперсію.

5. Обчислюємо відповідні власні вектори a_k.

6. Знаходимо головні компоненти-вектори: $Z_k = X^* a_k$, $k = \overline{1,n}$.

   Головні компоненти повинні задовільняти таким умовам: $\sum\limits_{i=1}^{m} z_{k,i} = 0$, $i = \overline{1,m}$,

$$\frac{1}{n} Z_k^T Z_k = \lambda_k, \quad k = \overline{1,n},$$

$$Z_j^T Z_k = 0, \quad j \neq k.$$

7. Визначаємо параметри моделі $\hat{Y} = Z\hat{b}$:

$$\hat{b} = Z^{-1} Y.$$

8. Знаходимо параметри моделі $\hat{Y} = X\hat{\beta}$:

$$\hat{\beta} = a\hat{b}.$$

3.5 Тестування і усунення гетероскедастичності

Застосування МНК веде до негативних наслідків, якщо не виконуються умови незалежності залишків і постійності їх дисперсії. Приклад, наведений на рисунку 1, показує, що прогноз значення показника y_{n+1} у точці x_{n+1} значно відрізняється від істинного значення. Виходячи із критерію мінімуму середньоквадратичної похибки на точках навчальної послідовності, найкращим наближенням експериментальної залежності є пряма. У той же час очевидно, що дисперсії залишків змінюються за деяким законом (квадратичним, або типу квадратичного кореня). У загальному випадку, таке явище призводить до того, що оцінки параметрів за МНК будуть незміщеними, змістовними але не ефективними і формулу для стандартної похибки оцінки адекватно застосовувати не можна.

Якщо дисперсія залишків змінюється для кожного спостереження або групи спостережень, то таке явище називається гетероскедастичністю.

Для перевірки наявності гетероскедастичності найчастіше використовують чотири методи у залежності від природи початкових даних: критерій μ, параметричний тест Гольдфельда-Квандта, непараметричний тест Гольдфельда-Квандта, тест Глейсера. Наведемо алгоритми кожного з цих методів та зазначимо особливості їх застосування.

Критерій μ (застосовується у випадку значної сукупності початкових даних).

1. Значення результуючої характеристики Y розбиваються на k груп відповідно до змін рівня величини (наприклад, за збільшенням).

2. Для кожної групи даних обчислюємо суму квадратів відхилень $S_r = \sum\limits_{i=1}^{n_r} (y_{ir} - \bar{y}_r)^2$, $r = \overline{1,k}$, де n_r – кількість елементів у r-вій групі.

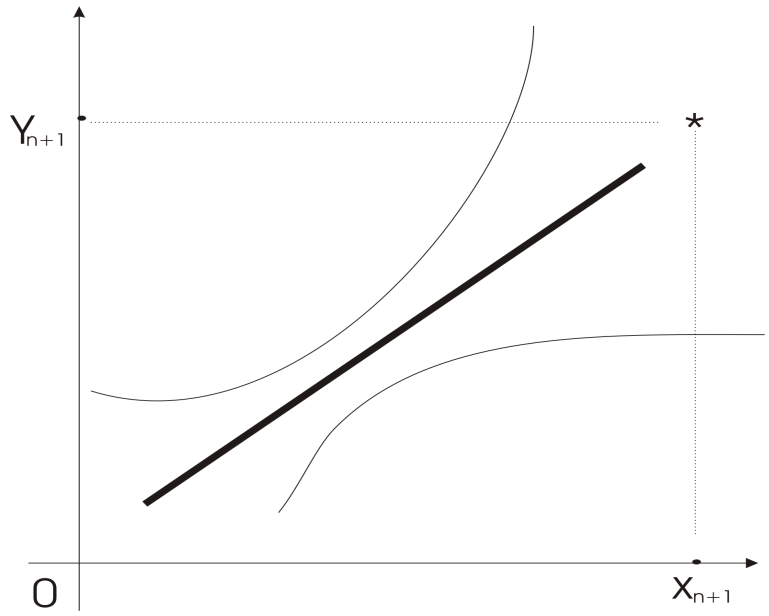


Рис. 1: Приклад гетероскедастичності

3. Визначимо суму квадратів відхилень в цілому по сукупності спостережень: $S = \sum\limits_{r=1}^{k} S_r = \sum\limits_{r=1}^{k} \sum\limits_{i=1}^{n_r} (y_{ir} - \bar{y}_r)^2$, де n_r – кількість елементів в r-вій групі.

4. Обчислимо параметр
$$\alpha = \prod_{r=1}^{k} \frac{\left(\frac{S_r}{n_r}\right)^{\frac{n_r}{2}}}{\left(\frac{S}{n}\right)^{\frac{n}{2}}}$$
, де n – кількість спостережень.

5. Обчислимо значення критерію $\mu = -2 \ln \alpha$, який приблизно відповідає розподілу χ^2 із ступенем вільності $k - 1$, якщо дисперсія всіх спостережень однорідна.

Таким чином, якщо значення μ не менше табличного значення χ^2 при вибраному рівні довіри і ступені вільності $k-1$, то приймається гіпотеза про наявність гетероскедастичності.

Параметричний тест Гольдфельда-Квандта (застосовується, якщо кількість спостережень невелика і зроблено припущення про те, що дисперсія залишків зростає пропорційно квадрату однієї з незалежних змінних.)

1. Упорядкувати спостереження відповідно до величини елементів вектора X_k, для якого ймовірно виконується вищенаведена умова.

2. Виходячи із співвідношення $\frac{c}{n} = \frac{4}{15}$, запропонованого авторами методу, де n – кількість елементів X_k, вилучити c спостережень, які знаходяться в середині вектора.

3. Згідно МНК побудувати дві економетричні моделі по двох отриманих сукупностях спостережень розмірністю $\frac{n-c}{2}$, природно за умови, що $\frac{n-c}{2} > m$, де m – кількість незалежних факторів (екзогенних змінних) присутніх в моделі.

4. Знайти суму квадратів залишків для першої і другої моделей:

$$S_1 = \sum_{i=1}^{\frac{n-c}{2}} (y_i^1 - \hat{y}_i^1)^2 \quad i \quad S_2 = \sum_{i=1}^{\frac{n-c}{2}} (y_i^2 - \hat{y}_i^2)^2$$

5. Обчислити значення критерію $R* = \frac{S_2}{S_1}$, який відповідає F-критерію із $(\frac{n-c-2m}{2}, \frac{n-c-2m}{2})$ ступенями свободи. Таким чином, якщо $R* \leq F_{\text{табл}}$, то гіпотеза про відсутність гетероскедастичності приймається.

Тест Глейсера

Крок 1 Згідно із МНК знаходимо параметри лінійної регресії та для кожного окремого спостереження визначаємо помилки ε_i.

Крок 2 Будуємо регресію, яка пов'язує абсолютні значення похибок, знайдених на першому кроці $|\varepsilon_i|$, з незалежною змінною x_i. Форма регресії підбирається з різних форм кривих:

$$|\varepsilon_i| = b_0 + b_1 x_i^2 + u_i, \quad |\varepsilon_i| = b_0 + b_1 x_i^{-1} + u_i, \quad |\varepsilon_i| = b_0 + b_1 x_i^{1/2} + u_i,$$

$$|\varepsilon_i| = \sqrt{b_0 + b_1 x_i} + u_i, \quad |\varepsilon_i| = \sqrt{b_0 + b_1 x_i^2} + u_i.$$

Крок 3 Якщо $b_0 = 0$ і $b_1 \neq 0$, то має місце "чиста"гетероскедастичність, якщо $b_0 \neq 0$ і $b_1 \neq 0$, то така гетероскедастичність називається "змішаною".

Проводимо будь-який тест на значущість параметрів b_0 і b_1. Якщо вони значно відрізняються від нуля, то ε_i є гетероскедастичними.

3.6 Автокореляція

Автокореляція – це взаємозв'язок послідовних елементів часового або просторового ряду даних. В економетричних дослідженнях виникають ситуації, коли дисперсія залишків постійна, але має місце їх коваріація. Це явище називають автокореляцією залишків.

Автокореляція залишків найчастіше спостерігається тоді, коли економетрична модель будується на основі часових рядів. Якщо існує кореляція між послідовними значеннями деякої незалежної змінної, то буде і кореляція послідовних значень залишків.

Автокореляція може бути також і наслідком помилкової специфікації економетричної моделі. Крім того наявність автокореляції залишків може означати, що необхідно ввести в модель нову незалежну змінну. При присутній автокореляції МНК застосовувати не можна.

Якщо нехтувати автокореляцією залишків і оцінити параметри моделі МНК, то прийдемо до таких трьох наслідків.

1. Оцінки параметрів моделі можуть бути незміщеними але неефективними, тобто вибіркові дисперсії вектора оцінок модуть бути невиправдано великими.

2. Оскільки вибіркові дисперсії обчислюються не по уточнених формулах, то статистичні критерії t- і F-статистики, які знайдені для лінійної моделі, практично не можуть бути використані в дисперсійному аналізі.

3. Неефективність оцінок параметрів економетричної моделі призводить, як правило, до неефективних прогнозів, тобто прогнозів з дуже великою вибірковою дисперсією.

Критерій Дарбіна-Уотсона

Крок 1 Розраховуємо значення d-статистики за формулою:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}. \tag{22}$$

Крок 2 Задаємо рівень значущості α і за таблицею Дарбіна-Уотсона для кількості факторів k та кількості спостережень n знаходимо значення d_L і d_U.

Крок 3 Якщо виконується нерівність $0 < d < d_L$, то має місце позитивна автокореляція. Якщо $4 - d_L < d < 4$, то робимо висновок про негативну автокореляцію. У випадку виконання нерівності $d_L < d < d_U$ або $4 - d_U < d < 4 - d_L$ висновку про існування автокореляції зробити не можна. Якщо $d_U < d < 4 - d_U$, то автокореляції немає.

3.7 Завдання для лабораторної роботи №3

1. Користуючись даними файлу Life Expectancy.csv зробити вибірку даних згідно свого варіанту. За певним правилом, замінити пропущені значення відповідними числовими значеннями. Обчислити середні значення всіх екзогенних змінних за відповідний часовий проміжок для кожної з країн. Побудувати графіки залежностей.

Варіанти	Завдання
1	Дослідити країни, що розвиваються за 2000-2007рр
2	Дослідити розвинуті країни за 2000-2007рр
3	Дослідити країни, що розвиваються за 2008-2015рр
4	Дослідити розвинуті країни за 2008-2015рр
5	Дослідити всі країни за 2005рр
6	Дослідити всі країни за 2010рр
7	Дослідити всі країни за 2015рр

2. Дослідити змінні на мультиколінеарність. Вибрати найбільш суттєві фактори і побудувати модель множинної регресії.

3. Використовуючи різні критерії оцінити якість моделі.

4. Застосувати критерій μ (непарні варіанти), тест Глейсера (для парних варіантів) виконати тестування залишків побудованої моделі на гетероскедастичність.

4 Парна та множинна нелінійна регресія. Алгоритм Брандона

4.1 Парна нелінійна регресія

У випадку, якщо кореляційне поле показує нелінійний зв'язок між показниками, або коли (згідно F-критерію) відкинута гіпотеза про лінійний зв'язок між x і y, потрібно вибрати нелінійну регресію. Наведемо приклади деяких базових рівнянь регресії: $y = a_0 + a_1x + a_2x^2 + a_3x^3 + ...a_nx^n$ - поліноміальна регресія.
$y = a_0 + a_1 \ln x$ - логарифмічна регресія
$y = a_0 + a_1e^x$ - експоненціальна регресія
$y = a_0 + a_1x^p$ - степенева регресія
та ряд інших: $y = \frac{1}{a+bx}$, $y = \frac{1}{a+be^{-x}}$, $y = ab^x$, $y = ae^{\frac{b}{x}}$, $y = \frac{x}{a+bx}$, $y = a_0 + \sum_{i=1}^{m}(a_i \cos(ix) + b_i \sin(ix))$. Коефіцієнти всіх цих рівнянь можна визначити, використовуючи МНК.

Поліноміальна регресія вибирається, коли має місце немонотонна залежність між x і y. Якщо на кореляційному полі є тільки одна точка max або min, то вибирається квадратична регресія.

У випадку квадратичної регресії $y = a_0 + a_1x + a_2x^2$ коефіцієнти знаходять методом найменших квадратів і визначаються за даними із системи лінійних рівнянь.

$$\begin{cases} a_0 + a_1\bar{x} + a_2\bar{x^2} = \bar{y} \\ a_0\bar{x} + a_1\bar{x^2} + a_2\bar{x^3} = \bar{xy} \\ a_0\bar{x^2} + a_1\bar{x^3} + a_2\bar{x^4} = \bar{x^2y} \end{cases} \tag{23}$$

Якщо детермінант системи не рівний нулю, то існує єдиний розв'язок для коефіцієнтів квадратичної регресії.

В інших випадках нелінійної регресії її зводять до лінійної за допомогою заміни змінних. Нехай виходячи з економічний суджень або з виду кореляційного поля, вибрана степенева регресійна модель: $y = ax^b$. Логарифмуючи, одержимо: $\ln y = \ln a + b \ln x$. Проводимо заміну змінних: $v = \ln y$, $z = \ln x$, будуємо лінійну регресійну модель:

$$v = a_0 + a_1z = a_0 + a_1 \ln x,$$

де $a_0 = \ln a$, $a_1 = b$. Звідки нелінійна регресія буде мати вигляд: $y = e^{a_0}x^{a_1}$.

Цей же метод використовується при побудові інших видів нелінійної регресії. На практичних задачах зазвичай будується лінійна і декілька нелінійних моделей. А тоді за максимальним коефіцієнтом детермінації вибирається одна з них.

4.2 Множинна нелінійна модель

Перший етап нелінійного множинного регресійного аналізу – це одержання, так званої, квадратичної форми. Для цього визначають коефіцієнти регресії b_0, b_j, b_{jm}, b_{jj} у поліномі:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + ... + b_mx_m + b_{11}x_1^2 + b_{22}x_2^2 + ... + b_{mm}x_m^2 + b_{12}x_1x_2 + ...$$

Степінь регресійної залежності можна підвищувати до тих пір, поки зменшується залишкова дисперсія. Починаючи з 2-го кроку кожному підвищенню степеня полінома передує заміна змінних, що лінеаризує функції $t_1 = x_1^2$, $t_2 = x_2^2, ...$, після чого коефіцієнти нового лінійного полінома визначають за згаданими раніше формулами.

Іншою формою проведення нелінійного регресійного аналізу є використання так званих "внутрішньо-лінійних"форм рівнянь, тобто форм, які легко лінеаризуються логарифмуванням та іншими перетвореннями. До таких моделей відносять мультиплікативну модель: $\hat{y} = b_0 x_1^{b_1} x_2^{b_2} \cdot \ldots \cdot x_m^{b_m}$.

Логарифмуючи за основою e, приводять її до лінійної:

$$\ln \hat{y} = \ln b_0 + b_1 \ln x_1 + b_2 \ln x_2 + \ldots + b_m \ln x_m.$$

Іншим прикладом "внутрілінійних"форм рівнянь є наступні експоненціальні моделі:

$$1)\ \hat{y} = exp\{b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m\}$$
$$2)\ \hat{y} = b_0 e^{b_1 x_1} \cdot e^{b_2 x_2} \cdot ... \cdot e^{b_m x_m}$$
$$3)\ \hat{y} = \frac{1}{1+e^{b_0+b_1 x_1+b_2 x_2+...+b_m x_m}}$$

І, накінець, можна згадати про обернену модель: $\hat{y} = \frac{1}{b_0+b_1 x_1+b_2 x_2+...+b_m x_m}$

4.3 Алгоритм Брандона

Американський економіст запропонував свій алгоритм побудови оптимальної нелінійної регресійної моделі. Припустимо початкові дані подані у таблиці:

x_1	x_2	$\cdots$	x_m	y
x_{11}	x_{12}	$\cdots$	x_{1m}	y_1
x_{21}	x_{22}	$\cdots$	x_{2m}	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_{n1}	x_{n2}	$\cdots$	x_{nm}	y_n

1. Обчислити середнє значення вихідної характеристики $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad y_i > 0.$

2. Виконати перетворення $y_{0i} = \frac{y_i}{\bar{y}}, \quad i = \overline{1,n}.$

3. Для пари змінних (y_0; x_1) побудувати всі залежності, які наведено вище, і за критерієм Дарбіна-Уотсона, або за значенням кореляційного відношення η, або за їх композицією (для лінійної залежності беруть коефіцієнт кореляції) вибирається оптимальна залежність: $\tilde{y}_0 = f_1(x_1)$.

4. Виконати перетворення $y_{1i} = \frac{y_{0i}}{\tilde{y}_{0i}}, \quad i = \overline{1,n}.$

5. Для пари змінних (y_1; x_2) вибрати вид залежності, що має максимальний рівень специфікації як сказано вище: $\tilde{y}_1 = f_2(x_2)$. Процес обчислень продовжують до вичерпання всіх факторів, що впливають на ендогенну змінну. Після визначення $\tilde{y}_{m-1} =$

$f_m(x_m)$ будуємо загальну формулу множинної регресії:

$$\tilde{y} = \bar{y} \prod_{k=0}^{m-1} \tilde{y}_k.$$

Кореляційне відношення розраховуємо за формулою:

$$\eta = \sqrt{1 - \frac{\sum\limits_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y}_i)^2}}.$$

Якщо, наприклад, $\eta = 0.7$, то це означає, що середня відносна похибка апроксимації дорівнює 30%.

Нехай $e_i = y_i - \tilde{y}_i$. Тоді значення критерію Дарбіна-Уотсона визначають за формулою:

$$DW = \frac{\sum\limits_{i=1}^{n}(e_i - e_{i-1})^2}{\sum\limits_{i=1}^{n} e_i^2}.$$

Якщо $DW = 2$, то автокореляція відсутня, якщо $DW = 0$ або $DW = 4$, то має місце повна автокореляція. Проміжні результати перевіряють за допомогою спеціальних таблиць.

4.4 Завдання для лабораторної роботи №4

1. За табличними даними методом Брандона побудувати рівняння нелінійної регресії, оцінити її адекватність та виконати її інтерпретацію.

5 Еволюційне моделювання і методи самоорганізації. Основні поняття і терміни

Розглянемо аналіз методів розв'язання задач, що базуються на еволюційних принципах. Покажемо їх переваги і недоліки при розв'язуванні задач оптимізації у порівнянні з класичними методами. Розглянемо аспекти наукових досліджень еволюційних технологій у відомих світових школах.

Існують різні варіанти класифікації класичних методів, що використовуються для прийняття рішень. Значну їх частину складають методи оптимізації, які застосовуються при розв'язанні задач лінійного, нелінійного, цілочисельного, опуклого, динамічного, стохастичного, геометричного програмування тощо. До сьогоднішнього дня не розроблені методи, які були б інваріантними до розмірності і змісту області даних, структури і параметрів цільової функції. Рухаючись у цьому напрямку, незалежно різними вченими були запропоновані парадигми, що базуються на ідеях і принципах природної еволюції. До них відносять відомі методи еволюційного моделювання, які ще називають еволюційними алгоритмами (ЕА):

- еволюційне програмування (ЕП);
- еволюційні стратегії (ЕС);
- генетичні алгоритми (ГА);
- генетичне програмування (ГП).

Розглянемо особливості кожного з вказаних еволюційних алгоритмів:

- методи ЕП орієнтовані на оптимізацію неперервних функцій без використання рекомбінацій;
- ЕС орієнтовані на оптимізацію неперервних функцій з використанням рекомбінацій;
- ГА призначені для оптимізації функцій дискретних змінних, в них акцентується увага на рекомбінаціях геномів;
- ГП використовує еволюційний метод для оптимізації комп'ютерних програм.

Така класифікація запропонована професором В.Г.Редьком. Проте сьогоднішні реалії вказують на те, що кожний з еволюційних алгоритмів застосовується для розв'язання й інших задач. Крім того, з'явилися методи, які також вважають представниками еволюційної парадигми. Це мурашині і меметичні алгоритми, програмування генетичних виразів тощо.

Оскільки кожний еволюційний алгоритм є ітераційним методом, то для його реалізації необхідно застосовувати обчислювальну техніку. Неминуче виникають питання збіжності кожного з них, швидкості збіжності, проведення препроцесінгу даних. Ефективний вибір і використання еволюційного алгоритму залежать від правильного співвідношення формалізованої задачі, сутності методу її розв'язання та очікуваних результатів.

Елементи еволюційного підходу присутні і у методі групового врахування аргументів. Цей метод застосовується у різних областях знань, що використовують структурну, параметричну ідентифікацію і прогнозування. Він базується на самоорганізації моделей і на відміну від регресійного аналізу, де структура моделі задана, спрямований на визначення структури моделі оптимальної складності. Тобто базується на переборі моделей, які поступово ускладнюються, і виборі найкращого розв'язку згідно із мінімумом значення зовнішнього критерію. Базовими моделями найчастіше є не лише поліноми але й інші нелінійні функції. За допомогою перебору різних розв'язків в індуктивному підході до моделювання намагаються мінімізувати роль впливу аналітика на результати моделювання. Комп'ютер знаходить структуру моделі і закони, за якими функціонує об'єкт, і використовується як порадник для відшукання нових розв'язків у задачах штучного інтелекту. Академік НАН України О.Г. Івахненко запропонував використовувати принцип зовнішнього доповнення. Базуючись на теоремі Вейєрштрасса про те, що будь-яку неперервну функцію можна як завгодно точно наблизити поліномом, він запропонував наступну схему.

5.1 Багаторядний метод групового врахування аргументів

Нехай початкові дані зосереджені у матриці $A = (X_1, X_2, ..., X_m, Y)$, де X_i, і Y – вектор-стовпчики розмірністю n. Задача полягає в ідентифікації залежності

$$Y = F(X_1, X_2, \ldots, X_m) \tag{24}$$

поліномом Колмогорова-Габора

$$Y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i<j} a_{ij} x_i x_j + \sum_{i<j<k} a_{ijk} x_i x_j x_k + \ldots \tag{25}$$

Відомо, що при збільшенні порядку полінома точність наближення ним функції $F(x)$ зростає, а потім спадає. Якщо точність є максимальною, то цей процес закінчується. Особливістю МГВА є те, що він може бути застосованим у випадку малої кількості точок експериментів, навіть значно меншої, ніж кількість членів полінома.

На першому етапі реалізації МГВА вибирається опорна функція. Найчастіше використовуються залежності виду:

1. $y = a_0 + a_1 x_i x_j$;
2. $y = a_0 + a_1 x_i + a_2 x_j$;
3. $y = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j$;
4. $y = a_0 + a_1 x_i + a_2 x_j + a_3 x_i^2 + a_4 x_j^2 + a_5 x_i x_j$.

Позначимо $y_k = f(x_i,\ x_j)$, де f – одна з вказаних залежностей або, можливо, подібна.

На наступному кроці за допомогою МНК визначають коефіцієнти рівнянь

$$y_1 = f(x_1, x_2), \ldots y_{n-1} = f(x_1, x_n), y_n = f(x_2, x_3), \ldots y_p = f(x_{n-1}, x_n), \tag{26}$$

де $p = C_n^2$.

Після того, як усі залежності y_i, $i = \overline{1,p}$ ідентифіковані, за зовнішнім критерієм відбирають кращі. Визначення їх кількості відносять на свободу вибору, найчастіше це 40-60%. Ті залежності, які залишилися, перенумеровують і одержують $y_1, y_2, \dots y_s$. Перший крок селекції закінчено.

На наступному кроці за допомогою МНК визначають коефіцієнти таких залежностей: $z_1 = f(y_1, y_2),\ z_2 = f(y_2, y_3), \dots, z_r = f(y_{s-1}, y_s),\ r = C_s^2$. Подальша процедура аналогічна вищевикладеній. Якщо значення зовнішнього критерію покращуються, то селекція продовжується, в іншому випадку модель оптимальної складності одержана.

5.2 Критерій регулярності

Опишемо зовнішні критерії, використання яких базується на принципі зовнішнього доповнення (або принципі регуляризації). У залежності від типу задачі О.Г.Івахненко запропонував розглядати такі критерії: регулярності, незміщеності та балансу змінних. Відомі два критерії регулярності:

– мінімум середньоквадратичної помилки на нових точках окремої контрольної послідовності;

– максимум коефіцієнта кореляції на тих же точках.

Розглянемо процедуру їх застосування. Початкові дані знаходяться у таблиці

x_1	x_2	$\cdots$	x_m	y	y_1	y_2	$\cdots$	y_p
x_{11}	x_{12}	$\cdots$	x_{1m}	y_1	y_{11}	y_{12}	$\cdots$	y_{1p}
x_{21}	x_{22}	$\cdots$	x_{2m}	y_2	y_{21}	y_{22}	$\cdots$	y_{2p}
.	.	.	.	.	.	.	.	.
x_{n1}	x_{n2}	$\cdots$	x_{nm}	y_n	y_{n1}	y_{n2}	$\cdots$	y_{np}

Розділимо її рядки на дві частини (приблизно 60% на 40%), тоді $n = l + k$, де l– кількість точок експерименту у навчальній вибірці, k – у другій – контрольній вибірці. Значення l повинне бути більшим числа доданків в опорній функції $f(x)$.

Використовуючи елементи навчальної вибірки, визначаємо коефіцієнти залежностей (26). Далі розраховуємо значення критерію регулярності на точках контрольної вибірки, після чого впорядковуємо y_i за збільшенням значення критерію і вибираємо із них певну кількість із найменшим значенням. Після перенумерації вони складуть множину функцій наступного ряду селекції. Умови закінчення ітерацій не "канонізовані"і можуть бути, наприклад, такими:

– середнє значення помилки для наступного ряду селекції є більшим ніж найбільше (середнє) значення помилки для попереднього ряду;

– мінімальне значення помилки наступного ряду більше мінімального значення помилки попереднього ряду;

– максимальне значення помилки наступного ряду більше максимального значення помилки попереднього ряду;

– модуль відхилення помилок наступного і попереднього ряду менше деякого числа.

Критерій регулярності, що полягає у мінімізації середньоквадратичної помилки на точках контрольної послідовності, є таким:

$$\varepsilon = \frac{\sum\limits_{i=1}^{N_k}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{N_k} y_i^2} \longrightarrow min,$$

де N_k – кількість точок у контрольній послідовності, $\hat{y}_i$ –розраховане значення шуканої залежності в i -вій точці контрольної послідовності. Інший варіант критерію регулярності полягає у максимізації коефіцієнта кореляції.

Щоб можна було одержані результати для різних моделей порівняти між собою, то величини y_i і $\hat{y}_i$ потрібно нормувати, тобто відняти середнє і ділити на стандартне відхилення або на середнє значення.

Перевагою критерію регулярності є плавність зміни його значення при збільшенні складності моделі. Недоліком його використання є низька точність при розв'язанні екстраполяційних задач. Тому критерій регулярності раціонально застосовувати для ідентифікації та короткострокового прогнозу.

5.3 Алгоритм поділу початкової вибірки даних

Реалізація МГВА, у більшості випадків, пов'язана із необхідністю поділу генеральної сукупності даних на дві вибірки – навчальну і контрольну. Найбільш поширеним, але не єдиним, є підхід, при якому в навчальну послідовність вибирають точки експериментів з більшим значенням дисперсії, а у контрольну – з меншим. Це пояснюється тим, що область навчання повинна бути якнайширшою, а контрольні точки, в більшості своїй, знаходиться всередині неї.

Алгоритм поділу є таким:

1. Визначити відсоткове співвідношення між кількістю елементів у навчальній і контрольній послідовності.

2. Для кожного стовпчика X_i, $i = \overline{1,m}$ знайти вибіркове середнє $\bar{X}_i$.

3. Знайти вибіркові дисперсії для кожного рядка: $D_j = \frac{1}{m-1}\sum\limits_{i=1}^{m}(x_{ij} - \bar{X}_i)^2$, $j = \overline{1,n}$

4. Для впорядкування таблиці переставити рядки так, щоб першим був рядок з найбільшим значенням дисперсії, а останній з найменшим.

5. У відповідності до результату кроку 1 розділити дані у таблиці на навчальну і контрольну послідовності.

Якщо розв'язується задача короткострокового прогнозу (на один такт часу вперед), то шукають ще і оптимальне співвідношення кількості образів у навчальній послідовності до кількості образів у контрольній з метою отримання найпростішої і достовірної моделі.

5.4 Генетичний алгоритм. Еволюційні стратегії .

З біології відомо, що генетичний код організму називається його генотипом, а фізична реалізація коду – фенотипом. Ці та інші визначення є базовими в термінології ГА, що не означає точного наслідування біологічних процесів, і лише в деякому наближенні ГА можна вважати їх моделлю.

До базових операторів ГА відносять кроссовер (рекомбінації, кроссинговер), мутації і інверсії. З їх допомогою здійснюється домінуюче розмноження краще адаптованих до зовнішнього середовища індивідів, а також одержання індивідів з характеристиками, які були відсутні у індивідів попередніх поколінь. В оптимізаційних задачах, таким чином, реалізується наближення до оптимального розв'язку і вибивання цільової функції з локальних екстремумів.

ГА є одним із методів знаходження екстремумів складних функцій. ГА – складова частина еволюційного моделювання як наукового напряму, що базується на принципах природного відбору за Дарвіним.

Розглянемо базовий ГА:

1. Ініціалізувати початковий момент часу $t = 0$.
2. Випадковим чином сформувати початкову популяцію, що складається з k індивідів: $B_0 = \{A_1, A_2, ..., A_k\}$.
3. Обчислити пристосованість кожного індивіда $F_{A_i} = fit(A_i)$, $i = \overline{1,k}$ і популяції в цілому $F_t = fit(B_t)$ (її ще називають фітнес-функцією). Значення цієї функції вказує на те, наскільки оптимальним є індивід, що описується даною хромосомою, для розв'язання задачі.
4. Вибрати індивіда A_{c_1} з популяції: $A_{c_1} = Get(B_t)$.
5. Вибрати другого індивіда з популяції $A_{c_2} = Get(B_t)$ і з певною ймовірністю (ймовірністю кроссоверу P_c) виконати оператор кроссоверу.
6. З ймовірністю 0.5 з A_{c_1} і A_{c_2} відібрати одного індивіда $A_c = Get(A_{c_1}, A_{c_2})$.
7. З певною ймовірністю (ймовірністю мутації P_m) виконати оператор мутації: $A_c = mutation(A_c)$.
8. З певною ймовірністю (ймовірністю інверсії P_i) виконати оператор інверсії: $A_c = inversion(A_c)$.
9. Помістити отриману хромосому в нову популяцію: $insert(B_{t+1}, A_c)$.
10. Виконати операції, починаючи з кроку 3, k раз.

11. Збільшити номер поточної епохи: $t = t + 1$.

12. Якщо виконується умова зупинки, то завершити роботу, інакше перейти на крок 3.

Нехай S – деяка система або процес. Її атрибутами є X – вектор вхідних і внутрішніх параметрів, Y – вектор результуючих характеристик. Припустимо, що перетворення $Y = f(X)$ є ідентифікованим і залежність f достатньо складна. Відомі також границі можливих змін значень складових вектора X. Необхідно знайти такі значення вектора X, щоб значення Y було оптимальним.

Не можна стверджувати, що така задача не може бути розв'язаною іншими методами, але у разі достатньо складної, можливо розривної, поліекстремальної функції f робити це дуже і дуже складно.

Як розв'язується задача з використанням ГА? Функція f є фітнес-функцією. Можливі значення елемента вектора X є його фенотипом. Двійковим представленням фенотипу є генотип (наприклад, 34 $\longrightarrow$ 100010). Генотип має певну кількість елементів (генів, бітів). Один або декілька генотипів (по кількості елементів в X) утворюють хромосому. Кросовером називають поділ двох хромосом і обмін частинами (наприклад, батьки – 1100 і 1010 $\longrightarrow$ нащадки – 1110 і 1000). Мутація – інвертування одного з елементів хромосоми (наприклад, 0000 $\longrightarrow$ 0100). Інверсія – зміна порядку місцезнаходження частин хромосоми (наприклад, 1100 $\longrightarrow$ 0011).

5.5 Завдання для лабораторної роботи №5

Необхідно одержати математичну модель за методом МГВА, здійснити прогнозування та порівняти результат МГВА із результатом, що одержаний будь-яким іншим методом, використовуючи контрольні точки відомої функції.

6 Методи кластеризації. Постановка задачі. Характеристика методів кластерного аналізу

Процес поступального руху до створення інформаційного суспільства супроводжують проблеми, пов'язані із зберіганням і обробкою великих масивів даних. Необхідною умовою їх вирішення є інтелектуальний аналіз даних, технології якого формуються на перетині теорії штучного інтелекту, статистики та теорії баз даних. Однією з технологій є data mining ("розкопка даних"). Великим масивам інформації властива присутність шумових ефектів, їх обробка призводить до накопичення сукупної помилки. Для подолання вказаної проблеми необхідно встановити значущі фактори і здійснити їх аналіз. Зменшення інформаційної ентропії може бути також досягнуто шляхом групування об'єктів і виявлення знань в менших, функціонально зв'язних сукупностях. Такі процедури направлені на послідовне подолання невизначеності. Їх першим кроком є розв'язання задачі кластеризації.

Важливо зауважити, що кластеризація і класифікація є основою як повсякденної діяльності людини, так і фундаментальним процесом наукової практики, оскільки навіть діти класифікують об'єкти навколишнього світу, а система класифікацій містить поняття, які є необхідною умовою розробки теорій та методів науки. Найбільш часто методи кластеризації використовуються у соціології, маркетингових дослідженнях, економіці, біології, медицині, археології.

Складність задач кластерного аналізу полягає в тому, що реальні об'єкти є багатовимірними, тобто описуються не одним, а кількома параметрами, і об'єднання об'єктів у групи проводиться у n-вимірному просторі, що є зовсім не тривіальним.

6.1 Постановка задачі

Задача кластеризації полягає у визначенні груп об'єктів (процесів), які є найближчими один до іншого за деяким критерієм. При цьому ніяких припущень про їх структуру, як правило, не робиться. Більшість методів кластеризації базується на аналізі матриці коефіцієнтів схожості, до яких належать відстань, кореляція та інші. Якщо критерієм або метрикою виступає відстань, то кластером називають групу точок Ω, таку, що середній квадрат внутрішньогрупової відстані до центру групи менше середньої відстані до загального центру в початковому наборі об'єктів, тобто $\bar{d}^2_\Omega < d^2$, де $\bar{d}^2_\Omega = \frac{1}{N}\sum\limits_{X_i\in\Omega}(X_i - \bar{X}_\Omega)^2$, $\bar{X}_\Omega = \frac{1}{N}\sum\limits_{X_i\in\Omega} X_i$, N – кількість точок в кластері Ω.

Вихідними даними для аналізу можуть бути власне об'єкти і їх параметри. Дані для аналізу можуть бути також зображені матрицею відстаней між об'єктами, у якій на перетині i-го рядка з j-вим стовпцем записана відстань між i-вим і j-вим об'єктами.

Якщо відстані не дані зразу, то агломеративні алгоритми починаються з обчислення відстаней між об'єктами. Розглянемо кілька способів обчислення відстаней між об'єктами:

1. Відстань Евкліда $d(X_k, X_l) = \left(\frac{1}{n}\sum\limits_{j=1}^{n}(X_{kj} - X_{lj})^2\right)^{1/2}$.

2. Максимальна відстань за ознаками (відстань Чебишова) $d(X_k, X_l) = \max\limits_{1\le j\le n} |X_{kj} - X_{lj}|$.

3. Відстань Махаланобіса $d(X_k, X_l) = \left((X_k - X_l)R^{-1}(X_k - X_l)^T\right)^{1/2}$, де R^{-1} – обернена до коваріаційної матриця.

4. Відстань Хеммінга $d(X_k, X_l) = \frac{1}{n}\sum\limits_{j=1}^{n} |X_{kj} - X_{lj}|$.

Розв'язання задачі мінімізації відстані між об'єктами рівносильне розв'язанню задачі мінімізації відстані до об'єкту, що має усереднені характеристики, оскільки, наприклад, для відстані Хеммінга

$$\sum_{\substack{j=1\\k<l}}^{n} |X_{kj} - X_{lj}| = \sum_{\substack{j=1\\k<l}}^{n} |X_{kj} - \bar{X}_j + \bar{X}_j - X_{lj}| \le$$
$$\le \sum_{\substack{j=1\\k<l}}^{n} |X_{kj} - \bar{X}_j| + \sum_{\substack{j=1\\k<l}}^{n} |X_{lj} - \bar{X}_j| \le 2 \max_{p\in\{k,l\}} \sum_{j=1}^{n} |X_{pj} - \bar{X}_j|,$$

Задачі кластеризації супроводжують дві проблеми: визначення оптимальної кількості кластерів і розрахунок їх центрів. Початковими даними для задачі кластеризації є значення параметрів об'єктів дослідження. Найчастіше визначення оптимальної кількості кластерів є прерогативою дослідника. Припустимо, що число кластерів K задано і $K << n$, де n – кількість об'єктів. Отримаємо задачу

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} ||X_j - \bar{X}_i|| \longleftrightarrow min, \tag{27}$$

де n_i, $i = \overline{1, K}$ – кількість об'єктів в i-му кластері, $\bar{X}_i$ –середнє значення в кластері, $||X_j - \bar{X}_i||$ – відстань між об'єктами. Розв'язком задачі (27) є центри кластерів $\bar{X}_i$, які можуть міститися серед даних об'єктів, що є достатньо строгою умовою, і можуть бути представленими будь-якими точками області дослідження.

Перехід від об'єктів до відстаней між ними – важливий момент. Відстані між об'єктами – це одна з мір схожості. Існує наступна класифікація мір схожості за Снітом і Сокелом:

1) коефіцієнт кореляції;

2) відстані між об'єктами;

3) коефіцієнт асоціативності;

4) ймовірнісні коефіцієнти схожості.

Кількісне оцінювання схожості між об'єктами виходить з поняття метрики. При цьому об'єкти представляються точками координатного простору, причому зафіксовані схожості і відмінності між точками знаходяться у відповідності з метричними відстанями між ними.

Існують чотири стандартних критерія, яким повинна задовільняти міра схожості, щоб бути метрикою:

1) симетрія (якщо x і y об'єкти, то $\rho(x,y) = \rho(y,x) = 0$) ;

2) нерівність трикутника: $\rho(x,y) \leq \rho(x,z) + \rho(z,y)$;

3) відмінність нетотожних об'єктів (якщо $\rho(x,y) \neq 0$, то $x \neq y$);

4) відсутність відмінності між ідентичними об'єктами ($\rho(x,x') = 0$, де x, x' – ідентичні об'єкти).

Зазвичай, застосування тільки евклідової метрики не розв'язує всіх проблем. Наприклад, при проведенні масового опитування ми маємо відповіді типу "так-ні "гірше-краще ... До такого типу задач застосовуємо іншу міру схожості. Важливою мірою схожості, яка традиційно використовується у соціологічних науках, являється статистичний коефіцієнт кореляції, наприклад, коефіцієнт кореляції Пірсона.

Для бінарних даних часто застосовують такий метод обчислення коефіцієнтів подібності, які називають коефіцієнтами асоціативності. Розглянемо два індивіда (об'єкта) з бінарними характеристиками (наприклад: високий чи низький, палить чи ні, соціально активний чи ні, і т. д.) Тоді для двох індивідів можна утворити таблицю 2×2, у кожній клітинці якої стоїть число, що показує, скільки відповідних пар значень існує:

		індивід 1		
		1	2	Усього
інди-	1	a	b	a+b
-від 2	2	c	d	c+d
		a+c	b+d	a+b+c+d

Ось список коефіцієнтів подібності, які часто використовують на практиці для такої ситуації: 1) $\frac{a+d}{a+b+c+d}$; (цей коефіцієнт називають простим коефіцієнтом зустрічі, оскільки у чисельнику записано кількість параметрів, які співпадають у об'єктах). 2) $\frac{a}{a+b+c}$; 3) $\frac{2a}{2a+b+c}$; 4) $\frac{2(a+d)}{2(a+d)+b+c}$; 5) $\frac{a}{a+2(b+c)}$.

До яких даних застосовувати той чи інший коефіцієнт, визначається природою реальної ситуації (задачі).

Наприклад, нехай маємо двох індивідів, які описуються десятьма показниками:

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Тоді

		індивід		
		1	2	Усього
інди-	1	2	1	3b
-від	2	2	5	7
		4	6	10

Тоді значення коефіцієнтів будуть: 1) 0.7; 2) 0.4; 3) 0.57; 4) 0.82; 5) 0.25.

6.2 Основні сімейства кластерного аналізу

Розроблені вченими кластерні методи утворюють сім основних сімейств:

1) ієрархічні агломеративні методи;

2) ієрархічні дивізивні методи;

3) ітеративні методи групування;

4) методи пошуку модальних значень щільності;

5) факторні методи;

6) методи скупчень;

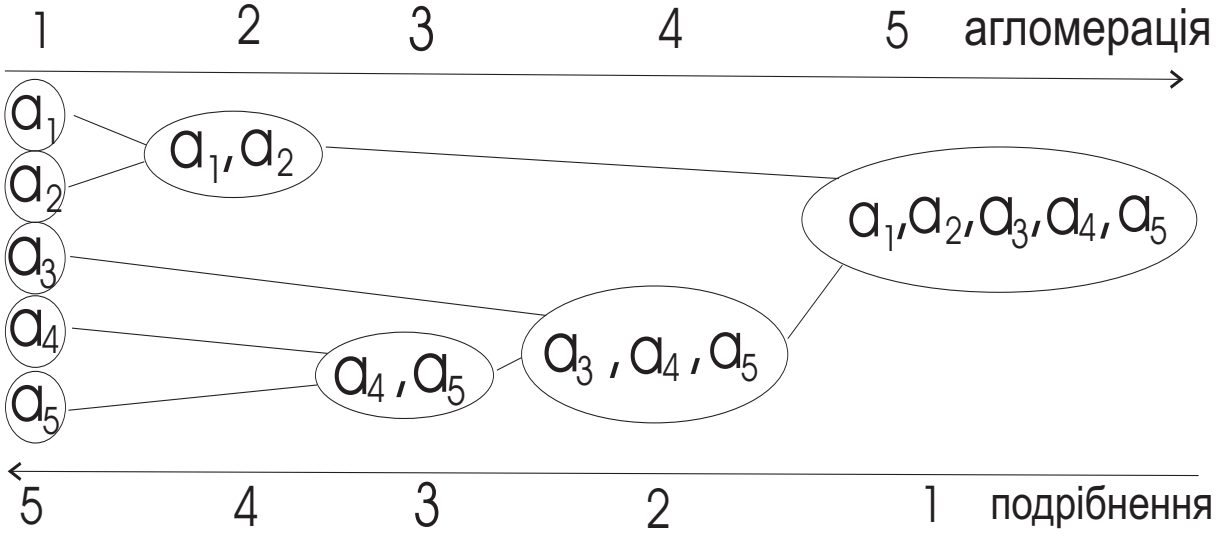
7) методи, які використовують теорію графів.

Ці сімейства відповідають різним підходам до утворення груп, і застосування різних методів до одних і тих же даних можуть привести до результатів, які значно відрізняються між собою. У конкретних галузях науки можуть виявитись особливо корисними певні сімейства методів. Так ієрархічні агломеративні методи найчастіше використовуються у біології, тоді як факторні аналітичні методи великим успіхом користуються у психології.

Взагальному із семи сімейств кластерних методів найбільш часто застосовують ієрархічні агломеративні та дивізивні методи (агломерат – скупчення, division – ділення, розбиття).

В агломеративних, або об'єднуючих методах виконується послідовне об'єднання найбільш близьких об'єктів в один кластер. Такий процес можна показати на графіку у виді дендрограми, або "дерева об'єднання". Це зручне представлення дозволяє наглядно зобразити кластеризацію агломеративними алгоритмами.

Приклад дендрограми для множини з п'яти об'єктів:



На першому кроці, коли кожен об'єкт є окремим кластером, відстані між цими об'єктами визначаються вибраним способом. Проте коли зв'язуються разом декілька об'єктів, виникає питання, як слід визначити відстані між кластерами? Іншими словами, необхідне

правило об'єднання або зв'язку для двох кластерів. Тут є різні можливості: наприклад, ви можете зв'язати два кластери разом, коли будь-які два об'єкти в двох кластерах ближче один до одного, ніж відповідна відстань зв'язку. Іншими словами, ви використовуєте "правило найближчого сусіда"для визначення відстані між кластерами; цей метод називається **методом одиничного зв'язку**. Це правило будує кластери, "зчеплені разом"тільки окремими елементами, що випадково опинилися ближче за інших один до одного. Недоліком цього методу є утворення занадто великих "довгастих"кластерів. Як альтернативу ви можете використовувати сусідів у кластерах, які знаходяться далі за решту всіх пар об'єктів один від одного. Цей метод називається **методом повного зв'язку** (або методом найвіддаленіших сусідів). Крім даних, існує багато інших методів об'єднання об'єктів у кластери:

Метод середнього зв'язку має декілька варіантів застосування. У першому з них обчислюється середня подібність об'єкта, який потрібно кластеризувати, з усіма об'єктами кластеру. Якщо знайдене середнє значення подібності досягає або перевищує деякезадане порогове значення, об'єкт приєднується до кластеру. В іншому варіанті обчислюють коефіцієнт подібності між центрами ваги кластерів, що підлягають об'єднанню.

Метод Уорда. Цей метод відрізняється від інших методів, оскільки він використовує методи дисперсійного аналізу для оцінки відстаней між кластерами. Метод мінімізує суму квадратів для будь-яких двох (гіпотетичних) кластерів, які можуть бути сформовані на кожному кроці. Суть методу пояснюють ще й так: він полягає у тому, що при переході від одного кроку розбиття до наступного об'єднують ті два кластери, при об'єднанні яких відбувається мінімальне збільшення загальної втрати інформації. За втрату інформації для однієї групи беруть зазвичай середньоквадратичне відхилення, а для кількох груп – суму всіх групових відхилень.

Наприклад, нехай маємо 10 осіб, для яких отримано значення деякої характеристики: 2, 6, 5, 6, 2, 2, 2, 0, 0, 0. Якщо кожен кластер містить лише одного індивіда, то оскільки відхилення в кожному кластері рівно 0, то загальне відхилення рівне теж 0.

Нехай маємо розбиття, яке складається лише з одного кластера, що містить усіх індивідів. Середнє значення характеристики тоді: $\frac{2+6+5+6+2+2+2+0}{10} = 2.5$ Відхилення в цьому разі: $(2-2.5)^2 + (6-2.5)^2 + ... + (0-2.5)^2 = 50.5$.

Якщо ж кластеризувати індивідів таким чином:

$$\{0,\ 0,\ 0\},\ \{2,\ \ 2,\ \ 2,\ \ 2\},\ \{5\},\ \{6,\ \ 6\},$$

то відповідні середні груп будуть: $\bar{x_1} = 0$, $\bar{x_2} = 2$, $\bar{x_3} = 5$, $\bar{x_4} = 6$. Відхилення для кожної групи нульове і тому загальне відхилення також рівне нулю.

Залежно від напряму, в якому розглядаємо створення кластерів, маємо або процес агломерації, або процес подрібнення.

В ієрархічних дивізимних методах можлива реалізація двох варіантів: монотетичного і політетичного поділу. У першому варіанті кластери визначаються за однаковістю або близькістю значень однієї ознаки, у другому – до кластеру належать об'єкти, що мають певні співвідношення значень із деякої множини ознак

На відміну від ієрархічних агломеративних методів, ітеративні методи групування кластерного аналізу не мали широкого застосування.

В основі застосування ітеративних методів групування лежить *базовий алгоритм:*

1. починають з вихідного розбиття даних на деяке задане число кластерів і обчислюють центр кожного з цих кластерів;

2. поміщають кожну точку даних у кластер з ближчим центром ваги;

3. обчислюють нові центри ваги кластерів; кластери не змінюють на нові до тих пір поки не будуть проглянуті повністю всі об'єкти;

4. кроки 2 і 3 повторюють до тих пір, поки не перестануть змінюватись кластери.

Застосування ітеративних методів групування пов'язане з рядом проблем обчислювального характеру, тому використовують підготовчі процедури для вибору вихідного поділу, типу ітерацій, статистичних критеріїв.

Вибір початкового розбиття здійснюється двома шляхами: перший полягає в тому, що визначаються початкові точки – центри кластерів і далі обчислюється відстань від кожного об'єкту до центрів кластерів. Об'єкт належить тому кластеру, відстань до якого є найменшою. Згідно другого способу об'єкти довільно розбивають на кластери і знаходять їх центри як середні значення.

Існує два основні типи ітерацій: за принципом "k-середніх"і за принципом "сходження на гору". Ітерації за принципом "k-середніх"полягають у переміщенні об'єкта в кластер з найближчим центром ваги. Вони можуть бути комбінаторними або некомбінаторними. У першому випадку перерахунок центра кластеру здійснюється після кожної зміни його складу, в іншому – лише після того, як буде завершено перегляд усіх даних. Крім того, ітерації за принципом "k-середніх"є включаючими та виключаючими. У включаючих ітераціях після обчислення центру кластера об'єкт включається до складу коастера, у виключаючих – вилучається із кластера. В ітераціях, що реалізуються за принципом "сходження на гору переміщення об'єктів відбувається, виходячи із того, чи буде таке переміщення оптимізувати значення дечкого статистичного критерію.

До функцій, що визначають якість кластеризації (статистичні критерії), належать trW, $trW^{-1}B$, $detW$ та найбільше власне число матриці $W^{-1}B$, де W – об'єднана внутрішньогрупова коваріаційна матриця, B – об'єднана міжгрупова коваріаційна матриця. Використовуючи кожний із статистичних критеріїв знаходять кластери визначеного вигляду. Так, критерій trW орієнтований на утворення гіперсферичних однорідних кластерів. За критерієм $detW$ передбачається, що у кластерів буде однакова форма, не обов'язково гіперсферична.

Проблема всіх ітеративних методів – при розв'язуванні задачі кластеризації мають місце і субоптимальні розв'язки.

Факторні методи кластерного аналізу дуже популярні у психології. Найбільш відомими є: варіанти факторного аналізу, обернений факторний аналіз або факторизація Q-типу. Робота методів починається з формування кореляційної матриці схожості між

об'єктами. За кореляційною матрицею визначаються факторні навантаження, і об'єкти розподіляються по кластерам в залежності від їх факторних навантажень. До недоліків таких методів належать: необхідність обгрунтування застосування лінійної множинної регресії; проблема множинних факторних навантажень, оскільки існує проблема прийняття рішень, якщо об'єкт має високе навантаження більше ніж для одного фактора.

У **методах пошуку модальних значень щільності** кластер визначають як область простору із високою щільністю образів у порівнянні з навколишнім середовищем. Розрізняють два підходи: у першому – орієнтуються на методі одиночного зв'язку, у другому – на поділі "сумішей"багатовимірних ймовірнісних розподілів. Особливістю першого підходу є те, що при появі нового образу він з певним пріоритетом утворює новий кластер, ніж приєднується до вже існуючого. Друрий підхід базується на статистичній моделі, в якій елементи різних груп чи кластерів мають різний розподіл значень ознак.

Такі методи є чутливими до проблеми субоптимальних розв'язків, компоненти "сумішей"є багатовимірними нормальними розподілами, що визначає їх недоліки.

Методи згущення дозволяють утворювати кластери, які перекриваються. Вони вимагають обчислення матриці подібності між образами і встановлення оптимального значення стратегічного критерію. Оскільки ці методи на початку утворюють лише дві групи, то раціонально пропонувати декілька конфігурацій, кожна з яких оцінюється на придатність. Недоліком методів є те, що через невдалу пошукову процедуру відбувається повторне знаходження одних і тих же груп, що не надає нової інформації.

Порівняно новим напрямком у розробці методів кластеризації є методи, що базуються на теорії графів. Розвинений математичний апарат є альтернативою численним евристичним методам. Значного поширення набули методи, які базуються на пірамідальних мережах, що ростуть. Такі мережі дозволяють виконувати кластеризацію в режимі реального часу.

6.3 Алгоритми, що базуються на гіпотезі компактності

Гіпотеза компактності полягає в тому, що реалізації одного і того ж образу відображаються в просторі ознак у геометрично близькі точки, утворюючи "компактні"згустки у припущенні, що проведена попередня обробка образів. Міра компактності може бути різною, найчастіше цю роль грає Евклідова відстань. Розрізняють унімодальну, полімодальну та локальну компактність. Гіпотеза унімодальної компактності лежить в основі численних алгоритмів таксономії, за допомогою якої одержуються кластери у вигляді гіперсфер чи гіперпаралелепіпедів. Використання гіпотези локальної компактності пов'язано з критерієм інформативності, яким є означення кількості опорних точок, необхідних для безпомилкового розпізнавання навчальної послідовності. Для прогнозування значень пропущених елементів у таблицях "об'єкт-ознака"застосовується гіпотеза полімодальної компактності.

Алгоритм Forel

1. Ознаки об'єктів нормуються так, щоб їх значення знаходились на відрізку $[0,\ 1]$,

наприклад,

$$x_{ij} = \frac{x_{ij} - x_{min\ j}}{x_{max\ j} - x_{min\ j}}. \quad k = 0.$$

2. Будуємо гіперсферу мінімального радіуса, що охоплює всі m точок. При нормуванні, запропонованому на кроці 1, такий радіус дорівнює $R_k = \frac{\sqrt{n}}{2}$, де n – кількість факторів або ознак об'єкта.

3. Зменшуємо радіус гіперсфери за формулою $R_{k+1} = R_k - \frac{k+1}{10}R_k$ і центр сфери розміщуємо в одній із точок (вибраної випадково).

4. Визначаємо точки, відстань від яких до центра гіперсфери менше R_{k+1} і обчислюємо координати їх центра ваги.

5. Переносимо центр сфери в центр ваги і знову визначаємо внутрішні точки. Ця операція циклічно повторюється.

6. Якщо склад множини внутрішніх точок і, як наслідок, координати центра ваги не змінюються, то сфера зупинилась в області локального максимуму щільності точок в просторі ознак.

7. Вилучаємо з розгляду точки, що належать сфері (таксону 1).

8. Якщо ще залишились "вільні"точки, то кроки 2-7 повторити, інакше – перейти на крок 9.

9. Якщо кількість таксонів є меншою, ніж задана, то $k = k + 1$ і перейти на крок 3, інакше – на крок 10.

10. Закінчення алгоритму.

Існує ще модифікація наведеного базового алгоритму Forel, яку називають Forel-2. У ній передбачена зміна радіуса на певну величину на кожній ітерації. Разом із тим, відзначимо значний суб'єктивізм вибору як радіуса, так і його приросту, що часто призводить до неоптимальної кластеризації. Алгоритм Forel-2 дозволяє отримати точно задане число кластерів.

6.4 Алгоритми, що базуються на гіпотезі лямбда-компактності

У багатьох задачах важливу роль відіграють не самі відстані між об'єктами, а відношення між ними. Алгоритми, які розглядаються нижче, мають у своїй основі особливість людського сприйняття кластерів – увагу звертають не на абсолютні відстані, а на відношення відстаней між декількома сусідніми точками. Зробимо попередні припущення. Нехай усі точки генеральної сукупності з'єднані між собою ребрами повного графа. Позначимо довжину між точками A і B індексом α. Серед усіх ребер, які є суміжними цьому ребру, знайдемо найкоротше і його довжину позначимо β_{min}. Відношення $\lambda = \frac{\alpha}{\beta_min}$ називають λ-довжиною ребра (A, B). Очевидно, що більші значення λ мають ребра, які з'єднують

віддалені одна від іншої точки, оточені близькими сусідами. Саме такі локальні сплески щільності точок найкраще помічає людське око при емпіричній кластеризації.

Алгоритм λ-KRAB

1. Знайти пару точок із мінімальним значенням λ - відстані між ними і з'єднати їх ребром нового графа.

2. З'єднати наступні найбільш λ - близькі точки із тих, що ще не приєднані до побудованого графа.

3. Якщо всі точки вичерпані, то перехід на крок 4, якщо ні – то на крок 2.
   Зауваження 1 Отриманий граф не має петель і сумарна довжина всіх його ребер буде мінімальною. Граф із такими властивостями називають найкоротшим незамкненим шляхом (ННШ) і позначають λ-ННШ. Розв'язуємо задачу розбиття вихідної множини точок на два кластери.

4. Для кожного ребра знайдемо характеристику його напруженості
   $$C = \lambda \frac{2m_i}{m} \frac{2m_j}{m},$$
   де m_i, m_j – кількість точок, що знаходяться по різні сторони від даного ребра.

5. Позначимо довжину розірваного ребра d, розрахуємо середнє значення довжини внутрішніх ребер таксонів ν. Якщо кластер містить один об'єкт, то $\nu = 0$, два об'єкти – $\nu = 1$, при об'єднанні всіх точок в один кластер – $d = 0$.
   Зауваження 2 Розрив найбільш напруженого ребра забезпечує виконання таких умов:

   – границя між кластерами проходить по найбільш напружених ребрах λ-ННШ;

   – середня напруженість внутрішніх ребер у кластерах буде мінімальною;

   – кластери матимуть однакове число точок.

   Зауваження 3. Критерієм якості кластеризації вважають величину $F = \frac{cd}{cr+V}$, де cd – середня напруженість граничних ребер, cr – середня напруженість внутрішніх ребер кластерів. Коефіцієнт V є більшим або рівним нулю для того, щоб при збільшені числа одиничних кластерів значення F не прямувало до нескінченності.

Раціонально прирівняти V, наприклад, середньому значенню напруженості (c') повного λ-ННШ і тоді $F = \frac{cd}{cr+c'}$. Якщо всі точки об'єднані в один кластер, то $F = 0$. У проміжку між цими крайностями значення F може бути як меншим, так і більшим 1, але завжди є більшим 0. Характеристика F інваріантна по відношенню до абсолютних значень довжин ребер графа λ-ННШ, що дозволяє порівнювати між собою якість кластеризації різних множин при різній кількості об'єктів m, різному числі кластерів k, різній середній λ – відстані між об'єктами.

Якщо бажане число кластерів задано діапазоном від k_{min} до k_{max}, то, спостерігаючи за значеннями функції $F = f(k)$, можна знайти таке число кластерів, при якому F досягає максимуму, що відповідає оптимальній кластеризації.

Важливим етапом реалізації алгоритму KRAB є побудова λ-ННШ. Якщо кількість початкових точок перевищує декілька сотень, то етап побудови є дуже трудомістким. Для прискорення виконання процедури необхідно здійснити попередню підготовку даних.

6.5 Пірамідальні мережі, що ростуть

Ефективним інструментарієм розв'язання задач класифікації, прогнозування та діагностики є пірамідальні мережі, що ростуть (ПМР), які були запропоновані професором В.П. Гладуном в Інституті кібернетики у Києві. Мережі ПМР реалізують гіпотезу про закономірності структурування інформації при її сприйнятті. Застосування ПМР в різних областях науки і техніки підтвердило їх репутацію ефективного засобу структуризації великих обсягів інформації.

Означення 1 Ациклічний орієнтований граф, в якому немає вершин із однією вхідною дугою (IN), називається пірамідальною мережею, що росте.

Означення 2 Вершини, що не мають IN-дуг, називаються рецепторами, інші вершини – концепторами.

Означення 3 Підграф ПМР, що включає вершину А і вершини, від яких є шлях до вершини А, називається пірамідою вершини А.

Означення 4 Вершини, що входять в піраміду вершини А, утворюють її субмножину. Множина вершин, до яких є шляхи від вершини А, називається її супермножиною.

Означення 5 Вершини, які зв'язані з вершиною А у субмножині та супермножині, називаються, відповідно, O-субмножиною та O-супермножиною.

Рецептори ПМР відповідають ознакам об'єктів, концептори – описам об'єктів у цілому та перетину понять. У початковому стані мережа складається лише із рецепторів, концептори формуються в процесі роботи алгоритму її побудови.

Розробка і використання ПМР складаються з декількох етапів:

1. Побудова ПМР.
2. Формування в ПМР структур, що представляють поняття.
3. Формування кластерної бази даних.

Розглянемо їх детальніше. **Етап 1** Наведемо алгоритм побудови мережі із можливістю включення в існуючі піраміди об'єктів нових ознак в режимі реального часу, без заміни пірамід у цілому. Позначимо: F_A – підмножина збуджених вершин O-субмножини вершини A; G– множина збуджених вершин мережі, що не мають інших збуджених вершин у своїх супермножинах. При введенні опису ознак об'єкта відповідні рецептори збуджуються. Концептор збуджується, якщо збуджуються всі вершини його O-субмножини. Введення нових вершин відбувається за такими правилами:

A1. Якщо вершина A не збуджена і множина F_A містить більше ніж один елемент, то дуги, що з'єднують вершини з множини F_A з вершиною A, ліквідовуються і у мережу вводиться новий концептор, який з'єднується IN-дугами з вершинами множини

F_A та вихідною (OUT) дугою з вершиною A.

Інтерпретація A1. Умовою введення нової вершини є ситуація, коли деяка вершина мережі є не повністю збудженою (тобто збуджено не менше двох вершин її O-субмножини, але не всі). Нові вершини вводяться у субмножини не повністю збуджених вершин. Вони репрезентують у мережі перетини описів об'єктів (рис. 2.)

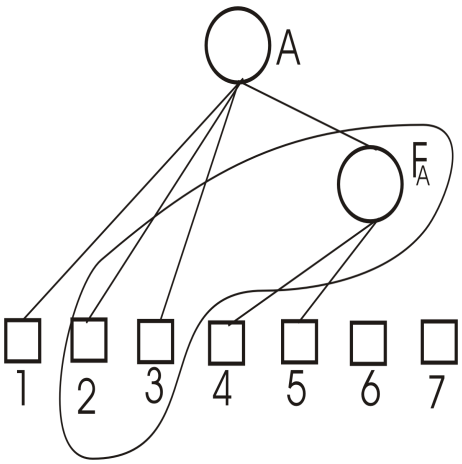


Рис. 2:

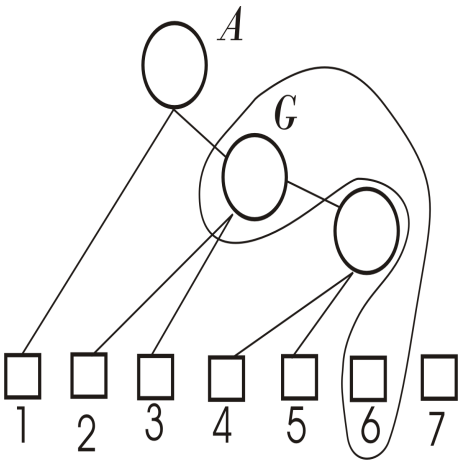


Рис. 3:

Мережа на рис.3 формується після збудження в мережі на рис.2 рецепторів 2,3,4,5.

A2. Якщо множина G містить більше ніж один елемент і не включає вершини, помічені іменем введеного об'єкта, до мережі приєднується новий концептор, який з'єднується IN-дугами з усіма вершинами множини G. Нова вершина знаходитиметься у збудженому стані. Мережа на рис.4 формується із мережі на рис.3 після збудження рецепторів 2,3,4,5,6.

A3. Якщо підмножина G містить більше ніж один елемент і включає вершину, помічену ім'ям введеного об'єкта, то ця вершина з'єднується IN- дугами з тими вершинами з множини G, які з нею не з'єднані. Мережа на рис.5 виникає із мережі на рис.4 за умови, що збуджені рецептори 2,3,4,5,6,7 і вони відповідають опису об'єкта B.

Означення 6 Елемент системи знань, що є узагальненою логічною ознаковою моделлю класу об'єктів, за допомогою якої реалізуються процеси розпізнавання і генерації моделей конкретних об'єктів, називається поняттям.

Етап 2 Означення 7 Множина узагальнених в понятті об'єктів складає його обсяг.

Розглянемо задачу індуктивного формування понять. Нехай V_i, $i = \overline{1,\ n}$ – множина об'єктів, $V_i \wedge V_j = \phi$, $i \neq j$. Позначимо L – множину об'єктів, яка є навчальною вибіркою. Мають місце співвідношення $L \bigcap V_i = \phi$ і $V_i \nsubseteq L\ \forall i = \overline{1,\ n}$, які свідчать про те, що з кожної множини хоча б один об'єкт представлено у навчальній вибірці і жодна множина

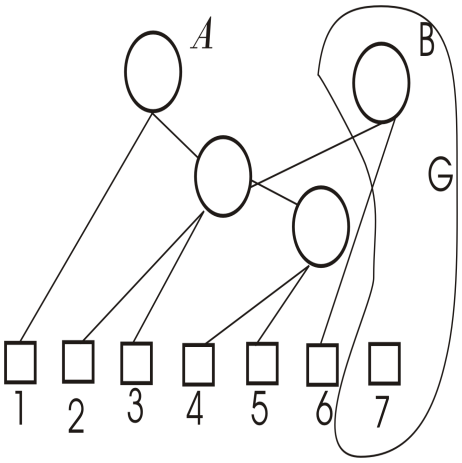


Рис. 4:

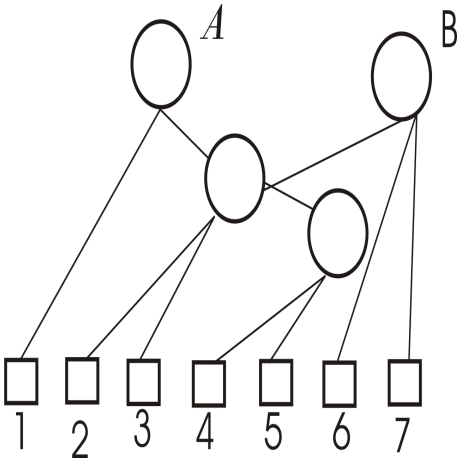


Рис. 5:

не входить повністю до навчальної вибірки. Шляхом аналізу необхідно сформувати n понять з обсягами $V_1, V_2, \ldots, V_n$, достатніх для правильного розпізнавання всіх об'єктів із L.

Нехай є ПМР, яка представляє всі об'єкти навчальної вибірки L. Для формування понять $P_1, P_2, \ldots, P_n$, які відповідають множинам $V_1, V_2, \ldots, V_n$, послідовно проглядають піраміди всіх об'єктів навчальної вибірки. У ПМР виділяють спеціальні вершини, за допомогою яких повинно здійснюватись розпізнавання об'єктів із обсягу поняття. Їх називають контрольними вершинами даного поняття. При виборі використовують дві характеристики вершин мережі: $m_1, m_2, \ldots, m_n$, де m_i – число об'єктів обсягу поняття P_i, в піраміди яких входить дана вершина; k – число рецепторів у піраміді, що відповідає вершині.

При перегляді піраміди виконуються перетворення за такими правилами:

B1 Якщо в піраміді об'єкта із обсягу поняття P_i вершина, що має найбільше k з усіх вершин з найбільшим m_i, не є контрольною вершиною поняття P_i, то вона помічається як контрольна вершина поняття. Якщо в групі вершин з найбільшим m_i значення k всіх вершин рівні, в якості контрольної вершини поняття P_i позначається будь-яка з них.

B2 Якщо в піраміді об'єкта із обсягу поняття P_i є контрольні вершини інших понять, які не містять у своїх супермножинах збуджених контрольних вершин поняття P_i, у кожній з цих супермножин вершина, яка має найбільше k із усіх збуджених вершин з найбільшим m_i, позначається як контрольна вершина поняття P_i.

Приклади застосування правил $B1$ і $B2$ зображені на рис. 6, 7, 8. Так на рис. 6 при збудженні піраміди вершини 2 контрольною вершиною є вершина 6, оскільки вона має найбільше k із усіх вершин, які мають найбільше m_i $\{6, 12, 13\}$. Числа в кружечках є значеннями m_i для концепторів, а в квадратиках – значення m_i для рецепторів.

У відповідності до правила B2 збудження піраміди вершини 2 (рис.7) за умови, що вона представляє об'єкт із обсягу поняття P_i, приводить до виділення в якості контрольної вершини поняття P_i вершини 5 (рис.8).

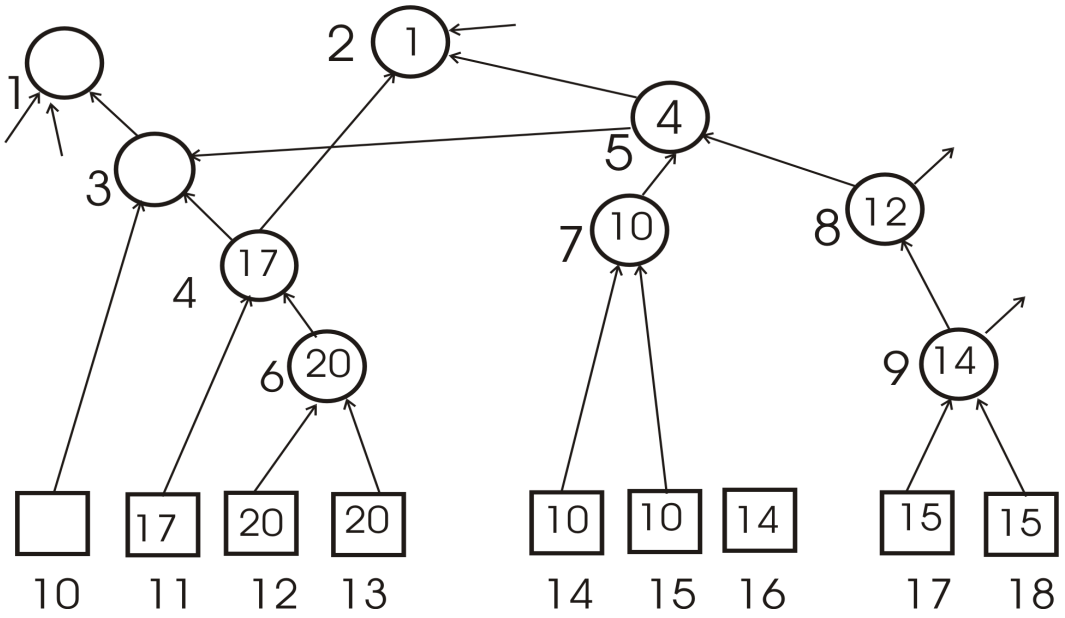


Рис. 6:

Якщо при розгляді всіх об'єктів навчальної вибірки з'явилась хоча б одна нова контрольна вершина, тобто хоча б один раз виконувались умови, які містяться в правилах В1 і В2, здійснюється новий перегляд всіх об'єктів навчальної вибірки. Робота алгоритму закінчується, якщо при черговому перегляді навчальної вибірки не виникає жодної нової контрольної вершини.

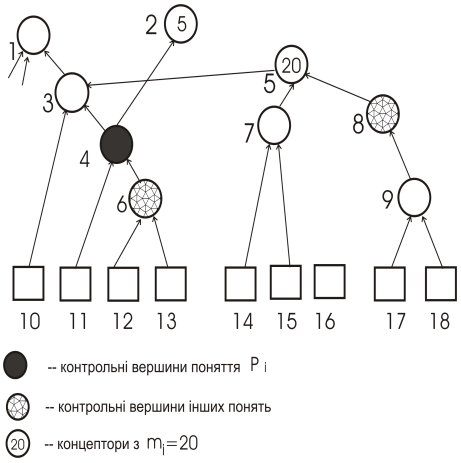


Рис. 7:

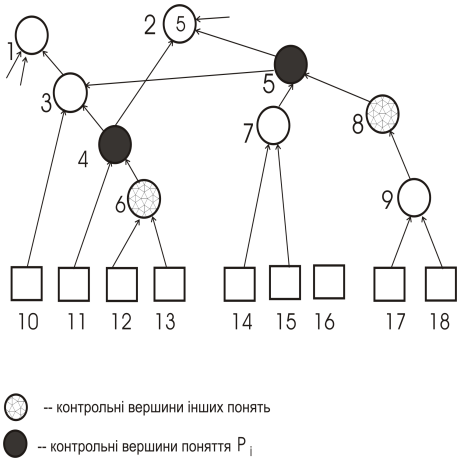


Рис. 8:

На наступному кроці застосовується таке правило розпізнавання. Об'єкт входить в обсяг P_i, якщо в його піраміді є контрольні вершини поняття P_i і не має жодної контрольної вершини будь-якого іншого поняття, яка не містить збуджених контрольних вершин поняття P_i у своїй супермножині. Якщо ця умова не виконується для всіх понять, об'єкт вважається невизначеним.

Існує певна аналогія між ПМР та нейронними мережами. Перевагою ПМР є те, що структура мережі апріорі не задана і формується у залежності від вхідних даних, чим зменшується інформаційна надлишковість. Крім того, знання одержані у результаті функціонування ПМР, є явно представленими і допускають інтерпретацію.

6.6 Еволюційна кластеризація. Алгоритм EvoClast

Альтернативним методом розв'язання задачі кластеризації є використання ідей, які лежать в основі еволюційного моделювання і зокрема генетичного алгоритму. Базовою операцією є формування фітнес-функції. Нагадаємо, що початковими даними задачі кластеризації є значення характеристик об'єктів. Після їх нормування, значення всіх параметрів належатимуть одиничному гіперкубу $[0.\ 1]^n$.

Реалізація фітнес-функції здійснюється за алгоритмом EvoClast:

1. Значення фітнес-функції покласти рівним нулю (F=0).
2. Задати кількість кластерів k і вказати значення n– кількості об'єктів.
3. Виконати ініціалізацію матриці належності елементів до кластерів.
4. Для всіх об'єктів виконати наступні кроки. Нехай $i = 1$.
5. Обчислити відстані від i-го об'єкта до центрів всіх k-кластерів, які є індивідами вибіркової популяції.
6. Серед усіх відстаней d_j, $j = \overline{1,k}$, вибрати мінімальну d_q і віднести i-вий об'єкт до q-го кластера. Внести відповідний запис в матрицю T_k.
7. $F = F + d_q$, $i = i + 1$.
8. Якщо кроки 5-7 виконані для всіх об'єктів, то отримано значення фітнес функції F, в іншому випадку перейти на крок 5.
9. Закінчення алгоритму.

Очевидно, що алгоритм отримання фітнес функції можна оптимізувати. Підвищення ефективності є його внутрішньою властивістю. Різноманіття варіантів операцій генетичного алгоритму представляють множину зовнішніх властивостей процесу отримання фітнес-функції. Можливість розв'язання задачі її оптимізації також припускає двійкове і десяткове представлення початкових даних. І якщо в першому випадку в процедурах генетичного алгоритму домінуючим є рівномірний розподіл, то у другому – при пошуку оптимального розв'язку перевага віддається значенням, що мають нормальний розподіл з математичним сподіванням, яке співпадає з центром кластеру. Визначення оптимальної дисперсії – одна задача, яка залишається нерозв'язаною.

Запропонований метод еволюційного моделювання, що базується на використанні генетичного алгоритму, ефективно функціонує при обробці масивів великої розмірності,

оскільки в ньому оптимально поєднуються цілеспрямований пошук і елементи випадковості, направлені на вибивання цільової функції з локальних мінімумів. Ніяких попередніх умов для його використання не вимагається. Головною умовою оптимізації обчислень є правильна алгоритмізація розрахунку значень цільової функції. Багатовекторність процесу покращення швидкості алгоритму (для генетичних алгоритмів це особливо необхідно) і його точності (пошуку глобального мінімуму фітнес-функції), а також його актуальність свідчать про необхідність розв'язання задачі оптимізації еволюційного методу.

6.7 Завдання для лабораторної роботи №6

1. Використовуючи методи класичного кластерного аналізу, виконати кластеризацію 56 об'єктів, кожний з яких має 8 ознак-характеристик (див. табл.). Визначити інформативні фактори. Порівняти точність результатів, наповнення кластерів, центри кластерів і значення цільової функції.

№	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	6	150	1.8	24	30	120	3.4	15
2	7	150	1.8	24	30	120	9.7	5
3	6	170	1.8	24	30	120	7.4	23
4	7	170	1.8	24	30	120	10.6	8
5	6	150	2.4	24	30	120	6.5	20
6	7	150	2.4	24	30	120	7.9	9
7	6	170	2.4	24	30	120	10.3	13
8	7	170	2.4	24	30	120	9.5	5
9	6	150	1.8	36	30	120	14.3	23
10	7	150	1.8	36	30	120	10.5	1
11	6	170	1.8	36	30	120	7.8	11
12	7	170	1.8	36	30	120	17.2	5
13	6	150	2.4	36	30	120	9.4	15
14	7	150	2.4	36	30	120	12.1	8
15	6	170	2.4	36	30	120	9.5	15
16	7	170	2.4	36	30	120	15.8	1
17	6	150	1.8	24	42	120	8.3	22
18	7	150	1.8	24	42	120	8	8
19	6	170	1.8	24	42	120	7.9	16
20	7	170	1.8	24	42	120	10.7	7
21	6	150	2.4	24	42	120	7.2	25
22	7	150	2.4	24	42	120	7.2	5
23	6	170	2.4	24	42	120	7.9	17
24	7	170	2.4	24	42	120	10.2	8
25	6	150	1.8	36	42	120	10.3	10
26	7	150	1.8	36	42	120	9.9	3
27	6	170	1.8	36	42	120	7.4	22

№	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
28	7	170	1.8	36	42	120	10.5	6
29	6	150	2.4	36	42	120	9.6	24
30	7	150	2.4	36	42	120	15.1	4
31	6	170	2.4	36	42	120	8.7	10
32	7	170	2.4	36	42	120	12.1	5
33	6	150	1.8	24	30	130	12.6	32
34	7	150	1.8	24	30	130	10.5	10
35	6	170	1.8	24	30	130	11.3	28
36	7	170	1.8	24	30	130	10.6	18
37	6	150	2.4	24	30	130	8.1	22
38	7	150	2.4	24	30	130	12.5	31
39	6	170	2.4	24	30	130	11.1	17
40	7	170	2.4	24	30	130	12.9	16
41	6	150	1.8	36	30	130	14.6	38
42	7	150	1.8	36	30	130	12.7	12
43	6	170	1.8	36	30	130	10.8	34
44	7	170	1.8	36	30	130	17.1	19
45	6	150	2.4	36	30	130	13.6	12
46	7	150	2.4	36	30	130	14.6	14
47	6	170	2.4	36	30	130	13.3	25
48	7	170	2.4	36	30	130	14.4	16
49	6	150	1.8	24	42	130	11	31
50	7	150	1.8	24	42	130	12.5	14
51	6	170	1.8	24	42	130	8.9	23
52	7	170	1.8	24	42	130	13.1	23
53	6	150	2.4	24	42	130	7.6	28
54	7	150	2.4	24	42	130	8.6	20
55	6	170	2.4	24	42	130	11.8	18
56	7	170	2.4	24	42	130	12.4	11

2. За значеннями цільової функції виконати порівняльний аналіз ефективності кожного із класичних методів кластеризації та еволюційного методу.

3. Використовуючи еволюційний метод кластеризації, виконати кластеризацію об'єктів. Здійснити перевірку алгоритму еволюційного методу на стійкість, використовуючи методи регуляризації даних. Дослідити точність результату як залежності від параметрів генетичного алгоритму.

7 Відновлення інформації

Сучасні наукові та практичні дослідження базуються на обробці поточної та ретроспективної інформації. Від того наскільки вона є якісною (інформативною, точною, достовірною і т.д.), залежить точність результатів. Особливий інтерес становить ситуація, коли частина даних відсутня. Це можливо із-за відмов обладнання, втрати інформації з технічних причин, а також суб'єктивних обставин.

Задача відновлення пропусків має декілька варіантів постановки, що визначається структурою пропущених даних. Зокрема, пропуски можуть бути серед значень вхідних факторів, результуючих характеристик, вхідних факторів і їх результуючих характеристик, а також серед значень ознак певного об'єкта, де такі фактори і характеристики явно не виділені.

Методи, якими розв'язуються задачі відновлення пропусків у даних, теж мають свою класифікацію. Розглядають методи, що базуються на елементарних обчисленнях, статистичні методи, ймовірнісні методи, нейромережні, еволюційні методи. Визначаючи, який метод застосовувати, необхідно знати особливості та обмеження його використання. Так, методи, що базуються на елементарних обчисленнях, раціонально застосовувати тоді, коли кількість пропусків є незначною. Одержані оцінки, найчастіше, є незміщеними.

Статистичні методи застосовують, якщо передбачається існування лінійної залежності між вхідними факторами та результуючими характеристиками. Необхідність апріорних знань про ймовірнісні характеристики визначає аспекти застосування ймовірнісних методів. Нейромережні методи, в загальному випадку, дозволяють обробляти різні структури пропусків, але точність одержаних оцінок визначатиметься інформативністю та повнотою даних для навчання нейромереж, а також їх архітектурою та законами функціонування. Оскільки задача відновлення пропусків має оптимізаційний характер, то для її розв'язання запропоновано використовувати еволюційні методи, які інтегрують в собі нейромережну ідентифікацію та генетичну оптимізацію. Результати експериментів засвідчили порівняно високу точність результатів. Недоліком є необхідність використання значних обчислювальних ресурсів.

7.1 Постановка задачі відновлення пропусків у таблицях даних

Нехай $X = (X_1, X_2, \ldots, X_n)$ – вектор вхідних факторів, $Y = (Y_1, Y_2, \ldots, Y_m)$ – вектор результуючих характеристик, p – кількість експериментів або періодів ретроспективи, $A = (a_{ij})$, $i = \overline{1,p}$, $j = \overline{1,n+m}$ –матриця вхідної інформації. Вона має пропуски, які позначені зірочками (табл. 2).

Припустимо, що між вхідними факторами і результуючими показниками існують залежності

$$\hat{Y}_i = F_i(X_1, X_2, \ldots, X_n), \quad i = \overline{1,m} \tag{28}$$

Тоді задача відновлення пропусків у даних полягає в знаходженні

$$\min_{*} \|Y - F(X)\|, \tag{29}$$

Табл. 2: Структура вихідної інформації

№	X_1	X_2	X_3	$\cdots$	X_{n-1}	X_n	Y_1	Y_2	$\cdots$	Y_m
1	a_{11}	a_{12}	a_{13}	$\cdots$	$*$	a_{1n}	a_{1n+1}	a_{1n+2}	$\cdots$	a_{1n+m}
2	a_{21}	a_{22}	$*$	$\cdots$	a_{2n-1}	a_{2n}	a_{2n+1}	$*$	$\cdots$	a_{2n+m}
3	a_{31}	$*$	a_{33}	$\cdots$	a_{3n-1}	a_{3n}	a_{3n+1}	a_{3n+2}	$\cdots$	$*$
p-1	a_{p-11}	a_{p-12}	a_{p-13}	$\cdots$	a_{p-1n-1}	$*$	a_{p-1n+1}	a_{p-1n+2}	$\cdots$	a_{p-1n+m}
p	a_{p1}	a_{p2}	a_{p3}	$\cdots$	a_{pn-1}	a_{pn}	a_{pn+1}	a_{pn+2}	$\cdots$	a_{pn+m}

де $F = (F_1, F_2, \ldots, F_m)$ і $Y = (Y_1, Y_2, \ldots, Y_m)$ – вектори значень, що одержані за ідентифікованими залежностями і наведені в табл. 2, відповідно. Задачу (29) деталізуємо і перепишемо у вигляді

$$\min_{*} \frac{1}{pm} \sum_{i=1}^{p} \sum_{j=1}^{m} (Y_{ij} - F_j(X_1^i, X_2^i, \ldots, X_n^i))^2, \tag{30}$$

або

$$\min_{*} \frac{1}{pm} \sum_{i=1}^{p} \sum_{j=1}^{m} (\hat{Y}_{ij} - a_{ij})^2, \tag{31}$$

Якщо припустити, що залежності (28) лінійні, тобто

$$\hat{Y}_i = b_{i0} + b_{i1}X_1 + b_{i2}X_2 + \ldots + b_{in}X_n, \tag{32}$$

тоді задача відновлення пропусків полягає у знаходженні:

$$\min_{*} \|Y - BX\|, \tag{33}$$

де $Y = (a_{ij})$, $i = \overline{1,p}$, $j = \overline{n+1, n+m}$, $B = (b_{ij})$, $i = \overline{1,m}$, $j = \overline{0,n}$, $X = (a_{ij})$, $i = \overline{1,p}$, $j = \overline{1,n}$.

Розв'язання задач (29) – (32) має перший етап, який, у загальному випадку, полягає в ідентифікації залежностей F_i, $i = \overline{1,m}$. Зауважимо, що в задачі відновлення пропусків у таблицях даних процедури ідентифікації та оптимізації ітеративно повторюються.

7.2 Евристичні методи обробки некомплектних даних

Наведемо та виконаємо аналіз методів відновлення втрачених даних, що застосовуються найчастіше. Матриця, рядок та стовпчик, що мають пропуски даних, називаються некомплектними.

1. Метод заповнення середнім значенням.
   Згідно із цим методом відсутнє значення на перетині i- го рядка і j-го стовпчика розраховується за формулою:

$$a_{ij}^* = \frac{1}{q} \sum_{i=1}^{p} a_{ij}, \quad a_{ij} \neq *, \tag{34}$$

де q – кількість заповнених елементів у j-му стовпчику. Перевагою методу є простота. Недоліки: в одному стовпчику може бути велика кількість пропусків і всі вони будуть заповнені однаковими значеннями; не враховується зв'язок некомплектного рядка з іншими рядками, що веде до зміщеної і недостовірної оцінки невідомого значення. Цей висновок справедливий і для випадку здійснення перерахунку з урахуванням кожного заповненого пропуску.

2. Метод виключення некомплектних рядків.
   Застосовується у випадку незначної кількості пропусків. Метод є простим, але вилучення даних збільшує ентропію прогнозних значень та веде до зміщеності параметрів моделі.

3. Метод підстановки
   Метод має декілька модифікацій. Розглянемо одну із них. Припустимо, що на перетині i-го рядка та j-го стовпчика є відсутнє значення. Тоді, серед усіх інших рядків вибираємо ті, в яких лише у j-му стовпчику пропуск або рядки, які є некомплектними. Знаходимо їх відстань до цільового рядка за формулою:

$$d_{ki} = \sqrt{\sum_{\substack{l=1 \\ l \neq i}}^{n+m} (a_{kl} - a_{il})^2}, \quad k = \overline{1, z}, \tag{35}$$

   де z –кількість рядків, що визначаються згаданою вище умовою. Впорядковуємо значення d_k за спаданням і задаємо деяке число $d > 0$. Серед всіх d_k, $k = \overline{1, z}$, вибираємо перші h, для яких $d_k > d$. Знаходимо значення пропуску

$$a_{ij} = \frac{\sum\limits_{l=1}^{h} a_{lj} C_l}{\sum\limits_{l=1}^{h} C_l} = \frac{\sum\limits_{l=1}^{h} a_{lj} \frac{1}{1+d_{li}}}{\sum\limits_{l=1}^{h} \frac{1}{1+d_{li}}}. \tag{36}$$

   Ідея методу базується на гіпотезі існування залежностей між факторами, що, найчастіше, не відповідає дійсності. Значні обчислювальні затрати зменшують і так низьку ефективність методу, оскільки для адекватних обчислень відстаней між рядками дані необхідно нормувати.

4. Метод множинної лінійної регресії.
   Застосовується у припущенні, що залежність (28) є лінійною. Для її ідентифікації використовують лише комплектні дані і з використанням МНК одержують (32). Очевидно, що надалі (32) використовуються для відновлення пропусків, але адекватно це можна робити лише у випадку одного пропуску серед значень рядка $(X_1^i, X_2^i, \ldots, X_n^i, Y_i)$. Якщо таких пропусків два і більше, то задача розв'язується за додаткових припущень та обмежень. Метод вимагає виконання ряду застережень та перевірок вхідних факторів на мультиколінеарність, гетероскедастичність, автокореляцію і застосування модифікованих версій МНК.

5. Метод множинної нелінійної регресії.
   На відміну від лінійної регресії його застосовують лише у випадку пропуску значень вихідної характеристики.

7.3 Відновлення пропусків значень залежної змінної

Розглянемо випадок, коли проводиться активний експеримент і значення факторів $X = (X_1, X_2, \ldots, X_n)$ задані дослідником, а Y – залежна від цих факторів змінна. Очевидно, що тоді пропуски серед значень вхідних факторів містяться набагато рідше, ніж серед значень вихідної характеристики. **Метод Бартлетта заповнення пропусків**
Зробимо припущення про те, що пропущені значення є лише серед значень результуючої характеристики Y і рядки, які їм відповідають, знаходяться вгорі таблиці вихідних даних. Кожний пропуск y_i, $i = \overline{1, m_0}$, заповнимо початковими значеннями $\tilde{y}_i$. Побудуємо матрицю Z супутніх значень змінних пропусків. За визначенням i-ва супутня змінна пропусків є індикатором i-го пропущеного значення, тобто завжди є нулем, за виключенням випадку, якщо пропущено i-ве значення, і тоді вона дорівнює одиниці. Перший рядок матриці Z: $Z_1 = (1, 0, 0, \ldots, 0)$, m_0-вий рядок – $Z_{m_0} = (0, 0, 0, \ldots, 1)$. Рядки, починаючи з $m_0 + 1$ до n-го, рівні $(0, 0, \ldots, 0)$. Таким чином, маємо вираз

$$\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_p \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & \ldots & 1 \\ 0 & 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \ldots \\ \gamma_{m_0} \\ \ldots \\ \gamma_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \ldots \\ \tilde{y}_{m_0} \\ \ldots \\ \tilde{y}_n \end{pmatrix},$$

або у матричному вигляді

$$Y = X\beta + Z\gamma + \varepsilon. \tag{37}$$

Тут $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ – вектор залишків, які є незалежними, однаково розподіленими, з нульовим середнім і однаковою дисперсією, β – оцінюваний параметр – вектор довжиною p.

Вважаючи, що виконуються всі передумови застосування МНК, класична оцінка β є такою (Y є комплектним):

$$\hat{\beta} = (x^T x)^{-1} x^T y. \tag{38}$$

Для кожної задачі згідно із МНК цільова функція

$$\Psi(\beta, \gamma) = \sum_{i=1}^{m_0} (\tilde{y}_i - x_i\beta - z_i\gamma)^2 + \sum_{i=m_0+1}^{n} (y_i - x_i\beta - z_i\gamma)^2. \tag{39}$$

Необхідно знайти $\min\limits_{\beta,\gamma} \Psi(\beta, \gamma)$.

Використовуючи визначення матриці Z із (39), одержимо

$$\Psi(\beta, \gamma) = \sum_{i=1}^{m_0} (\tilde{y}_i - x_i\beta - z_i\gamma)^2 + \sum_{i=m_0+1}^{n} (y_i - x_i\beta)^2. \tag{40}$$

Припустимо, що $\hat{\beta}^*$ – оцінка, одержана за МНК за формулою (38) для існуючих значень Y, тобто за останніми $m = n - m_0$ рядками. Вона мінімізує другу суму у виразі (40). Якщо при $\beta = \hat{\beta}^*$ покласти $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_{m_0})^T$, де $\hat{\gamma}_i = \tilde{y}_i - x_i\hat{\beta}^*$, $i = \overline{1, m_0}$, то мінімізується та перетворюється в нуль перша сума в (40) і одержимо функцію

$$\Psi(\beta, \gamma) = \sum_{i=m_0+1}^{n} (y_i - x_i\hat{\beta}^*)^2.$$

Таким чином, $(\hat{\beta}^*, \hat{\gamma})$ мінімізує $\Psi(\beta, \gamma)$ і є оцінкою МНК, одержаною з моделі (37). Рівняння (40) означає також, що точна оцінка МНК відсутнього значення y_i, тобто $\hat{\gamma}_i = x_i\hat{\beta}^*$ є $\tilde{y}_i - \hat{\gamma}_i$ або: прогноз i-го пропущеного значення за МНК є початковим значенням для i-го пропуску мінус коефіцієнт для супутньої змінної i-го пропуску.

Відомі два способи ініціалізації пропущених значень: за першим із них вони дорівнюють 0, за другим – є середнім відомих значень. До переваг методу належить його неітеративність; якщо структура пропусків є виродженою, то результат відсутній.

Resampling-метод

Метод відновлення пропусків Resampling є різновидом відомого методу "bootstrap"запропонованого американським статистиком Бредлі Ефроном. Його сутність полягає у багатократній обробці різних частин одних і тих же даних, що дозволяє здійснювати їх різносторонній аналіз і співставляти одержані результати. Припустимо, що дані пропуски мають ту ж структуру як і для методу Бартлетта. Застосування методу Resampling може бути виконано двома способами:

Resampling-1

1. Формуємо матрицю повних спостережень $H = (X_1, X_2, \ldots, X_p, Y)$, кількість рядків у якій m_0.

2. Випадковим чином вибираємо j-й рядок із матриці H та замінюємо i-вий рядок початкової матриці, $i \in \{m_0+1, m_0+2, \ldots, n\}$. Цей рядок може вибиратись випадково або по порядку, починаючи з $(m_0 + 1)$-го.

3. Якщо всі пропуски заповнені, то за МНК знаходимо коефіцієнти регресійного рівняння β_k, $k = \overline{0, p}$, в іншому випадку виконуємо перехід на крок 2.

4. Якщо одержано r векторів $\beta^q = (\beta_0^q, \beta_1^q, \ldots, \beta_p^q)$, $q = \overline{1, r}$, то знаходимо середні значення коефіцієнтів регресійної моделі: $\bar{\beta}_k = \frac{1}{r}\sum\limits_{q=1}^{r} \beta_k^q$, $k = \overline{0, p}$.

Resampling-2

1. За матрицею H будується регресійна модель і знаходяться оцінки коефіцієнтів $\hat{\beta}_i$, $i = \overline{0, p}$.

2. Розраховуємо оцінку $\hat{Y}_i$ за регресійною моделлю для $i = \overline{1, m_0}$.

3. Знаходимо помилки $\varepsilon_i = Y_i - \hat{Y}_i$, $i = \overline{1, m_0}$.

4. Для кожного пропуску, підставляючи значення супутніх змінних $X_1, X_2, \ldots, X_p$ в одержане регресійне рівняння, знаходимо оцінку $\hat{Y}_i$ $i = \overline{m_0+1, n}$.

5. Значення, якими замінюють пропуски, одержують за формулою: $Y_i = \hat{Y}_i + \varepsilon_i$, $i = \overline{m_0+1, n}$, де ε_i вибирають випадковим чином із результатів кроку 3.

6. За даними, одержаними після заповнення, будуємо регресійну модель і знаходимо оцінки коефіцієнтів $\hat{\beta}_i$, $i = \overline{0, p}$.

7. Аналогічний кроку 4 із resampling-1. У запропонованого методу resampling перевагою є повне використання вихідної інформації, водночас її повторне використання зменшує інформативність даних.

7.4 Локальні методи відновлення пропусків. Алгоритм ZET

Сутність алгоритму полягає у тому, що кожне пропущене значення оцінюють по "компетенцій"матриці, яка складається із визначеного числа рядків і стовпчиків вихідної матриці. Пропущене значення у рядку знаходять, використовуючи обчислення відстаней, у стовпчику – обчислення коефіцієнтів кореляції. Остаточну оцінку знаходять, усереднюючи попередні оцінки з ваговими коефіцієнтами, значення яких визначаються певними параметрами.

1. Виконаємо нормування значень матриці A за формулою $a_{ij} = \frac{a_{ij} - \overline{a_j}}{\sigma_j}$.

2. Припустимо, що $a_{ij} = *$. Задамо коефіцієнт впливу компетентності на результат прогнозування α.

3. Знаходимо відстані від усіх комплектних рядків матриці A до i-го рядка за формулою:
$$d_{ki} = \left(\sum_{\substack{l=1 \\ l \neq j}}^{n+m} (a_{kl} - a_{il})^2 \right)^{\frac{1}{2}}, \quad k = \overline{1, p}, k \neq i. \tag{41}$$

4. Визначаємо q рядків, для яких d_{ki} є найменшими із них та i-го рядка формуємо матрицю A_q (порядок розміщення рядків зберігається, i-й рядок стає ii-м).

5. Знаходимо модулі коефіцієнтів кореляції усіх стовпчиків матриці A_q з j-м стовпчиком:
$$r_{jk} = \frac{\frac{1}{q} \sum_{l=1}^{q} (a_{lj} - \overline{a_j})(a_{lk} - \overline{a_k})}{\frac{1}{q} \sum_{l=1}^{q} (a_{lj} - \overline{a_j})^2 \frac{1}{q} \sum_{l=1}^{q} (a_{lk} - \overline{a_k})^2}, \quad k = \overline{1, n+m}, k \neq j. \tag{42}$$

6. Визначимо ν стовпчиків, для яких значення r_{jk} є найбільшими, і з них та j-го стовпчика формуємо "компетентну"матрицю $A_{q\nu}$, при цьому порядок розміщення стовпчиків зберігається, а j-й стовпчик стає jj-м. Таким чином, $a_{ii\ jj} = *$.

7. Обчислюємо "компетентності"рядків k_l^p, $l = \overline{1,q}$, як величини обернено пропорційні відстанням до рядка, що містить відсутнє значення $k_l^p = \frac{1}{1+d_{l\ ii}}$.

8. Обчислюємо "компетентності"стовпчиків k_l^c, $l = \overline{1,\nu}$, як величини прямо пропорційні (або рівні) модулям r_{jk}.

9. Для кожного i-го рядка, $i = \overline{1,q}$, і ii-го рядка згідно із методом МНК знаходимо рівняння парної лінійної регресії: $a_{ii} = k_i a_i + b_i$ і, прирівнюючи $a_i = a_{i\ jj}$, знаходимо оцінку пропущеного значення $\hat{a}_i^p$, $i = \overline{1,q}$.

10. Аналогічний кроку 9, але для стовпчиків знаходимо оцінку $\hat{a}_j^c$, $j = \overline{1,\nu}$.

11. Для компетентності матриці знаходимо прогнозні величини для рядка і стовпчика:

$$\hat{a}^p = \frac{\sum\limits_{l=1}^{q} \hat{a}_l^p (k_l^p)^\alpha}{\sum\limits_{l=1}^{q} (k_l^p)^\alpha}, \quad \hat{a}^c = \frac{\sum\limits_{l=1}^{\nu} \hat{a}_l^c (k_l^c)^\alpha}{\sum\limits_{l=1}^{\nu} (k_l^c)^\alpha}.$$

12. Обчислення пропущеного значення $a_{ij} = \frac{1}{2}(\hat{a}^p + \hat{a}^c)$.

13. Закінчення алгоритму.

Алгоритм ZET застосовується у припущенні про виконання трьох гіпотез: надмірності, аналогічності та локальної компетентності, згідно з якими припускають, що у таблицях даних є подібні рядки та залежні стовпчики; якщо пара об'єктів має подібні значення $(n-1)$-го параметра, то і значення n-го параметра подібні; немає сенсу використовувати для заповнення відсутнього пропуску усіх рядків і стовпчиків матриці, а достатньо брати лише їх "компетентну частину".

До недоліків алгоритму ZET можна віднести певний "волюнтаризм"дослідника, який полягає у суб'єктивному визначенні розмірів "компетентної"матриці та коефіцієнта "компетентності що впливає на присутність шумового ефекту при передбаченні пропущеного значення та точність обчислень результату.

Алгоритм ZET Braid

На відміну від алгоритму ZET, в ZETBraid реалізована ідея поступового додавання в "компетентну"матрицю рядків і стовпчиків. Сутність алгоритму полягає у специфічному підрахунку відстаней між рядками та стовпчиками.

Відстань між рядками обчислюємо за формулою: $r_{ij} = \sum\limits_{k=1}^{n} b_k (a_{ik} - a_{jk})^2$, де b_k – ваговий коефіцієнт, значення якого залежить від того, чи входить i-й стовпчик в "компетентну"матрицю.

При обчисленні коефіцієнтів b_k, $k = \overline{1,n}$, дотримуються трьох принципів:

1. Всі вагові коефіцієнти стовпчиків, що входять в "компетентну"матрицю, рівні.

2. Всі вагові коефіцієнти стовпчиків, що не входять в "компетентну"матрицю, рівні.

3. Сума вагових коефіцієнтів стовпчиків, що входять в "компетентну"матрицю, поділена на суму вагових коефіцієнтів, що не входять в "компетентну"матрицю, є константою (параметр алгоритму).

Якщо із n стовпчиків p належать "компетентній"матриці, то ваговий коефіцієнт стовпчика:

$$W = \begin{cases} 1, & \text{якщо стовпчик не належить "компетентній"матриці,} \\ 1 + C_n(\frac{n}{p} - 1), & \text{в іншому випадку.} \end{cases}$$

Для знаходження відстані між стовпчиками необхідно будувати рівняння лінійної регресії. Нехай $X = (x_1, x_2, ..., x_m)$ і $Y = (y_1, y_2, ..., y_m)$ – стовпчики, тоді потрібно одержати рівняння $y = a + bx$.

Для знаходження коефіцієнтів a і b мінімізуємо функцію

$$D = D(a, b) = \sum_{i=1}^{m} b_i(y_i - a - bx_i)^2, \tag{43}$$

де вагові коефіцієнти рядків b_i, $i = \overline{1, m}$, знаходяться аналогічно ваговим коефіцієнтам стовпчиків, тобто

$$W = \begin{cases} 1, & \text{якщо стовпчик не належить "компетентній"матриці,} \\ 1 + C(\frac{m}{q} - 1), & \text{в іншому випадку,} \end{cases}$$

де q – кількість рядків, що належать "компетентній"матриці.

Важливою задачею залишається підбір розміру "компетентної"матриці. Критерієм оцінки адекватності цієї матриці є оцінка якості передбачення невідомого елемента. Таким чином, при побудові "компетентної"матриці рядки і стовпчики додаються до тих пір, поки значення критерію (абсолютного відхилення точного і прогнозованого значення) зменшується.

Відомими є два варіанти розрахунку цього критерію. Згідно із першим, методом "хреста за рівнянням лінійної регресії розраховують всі відомі значення рядка і (або) стовпчика, що містять невідомий елемент і знаходять середню помилку. Ця середня помилка і є оцінкою передбачення даної "компетентної"матриці.

Другий варіант, дисперсійний метод, полягає в обчисленні дисперсії передбачень невідомого елемента. Для цього за рівнянням лінійної регресії для кожного стовпчика прогнозують значення невідомого елемента і знаходять дисперсію цих $n - 1$ прогнозів, яка і є шуканою оцінкою.

"Компетентна"матриця за побудовою є квадратною або кількість рядків і стовпчиків відрізняються на одиницю. Для того, щоб розмірність матриці могла бути довільною, але адекватною, виконаємо такі кроки:

1. Знаходимо найбільш близький рядок, що не входить до "компетентної"матриці, до цільового рядка. Якщо додавання цього рядка не погіршує її оцінку, то додаємо його до "компетентної"матриці.

2. Аналогічний кроку 1 для стовпчика.

Кроки 1 і 2 повторюють до моменту погіршення оцінки "компетентної"матриці і для рядка, і для стовпчика.

Для того, щоб уникнути помилок при початковій побудові "компетентної"матриці, на перших К кроках (зазвичай, $K = 6$) додають рядки і стовпчики в "компетентну"матрицю, не зважаючи на її оцінку.

7.5 Еволюційний метод відновлення пропусків

Припустимо, що пропуски є лише серед значень вхідних факторів $X = (X_1, X_2, ..., X_n)$, результуюча характеристика Y одна і існує залежність $Y = F(X)$.

А.М.Колмогоров і В.І.Арнольд довели теорему про те, що кожна неперервна функція n змінних, задана на одиничному кубі n-вимірного простору, може бути представленою у вигляді

$$f(x_1, x_2, ..., x_n) = \sum_{q=1}^{2n+1} h_q \sum_{p=1}^{n} \varphi_q^p(x_p),$$

де функції $h_q(u)$ неперервні, а функції $\varphi_q^p(x_p)$, крім того, ще і стандартні, тобто не залежать від вибору функції f. В термінах теорії нейронних мереж теорема вказує на те, що будь-яка неперервна функція ідентифікується мережею з одним, як мінімум, прихованим шаром нейронів з нелінійними функціями активації. Для ідентифікації $F(x)$ в якості моделі виберемо прямозв'язну нейромережу з пороговим алгоритмом зворотного поширення похибки. Надалі структура мережі та її елементний базис в експериментах залишаються постійними.

Оскільки вхідні образи для навчання нейронної мережі мають пропуски значень, то необхідно розв'язати задачу параметричної оптимізації. Як метод оптимізації запропоновано використати генетичний алгоритм. Для гарантування його збіжності використаємо теорему, доведену Р.Харті.

Якщо використовувати бінарне представлення розв'язків і для формування їх популяції – елітний відбір, то теорема вказує на збіжність ГА за ймовірністю.

Для роботи ГА необхідно сформувати генеральну і вибіркоку сукупність хромосом-розв'язків. Хромосома складається з фрагментів, які відповідають пропускам в таблиці даних: $Xr =<$ пропуск 1, пропуск 2, ..., пропуск k $>$. Дані в таблиці без врахування пропущених значень нормуємо. Якщо активаційною функцією буде вибрано гіперболічний тангенс, то нормування раціональніше здійснювати у відрізок $[-1, 1]$. Кількість хромосом в генеральній сукупності визначається заданою точністю результату, у вибірковій – дослідником.

На наступному кроці формуємо навчальну і контрольну послідовність для навчання нейромережі. Пропонується всі образи з пропусками вважати елементами навчальної послідовності. Для контрольної послідовності їх використання є проблематичним, оскільки неможливо застосувати для верифікації недостовірні значення. Співвідношення потужності множини образів навчальної і контрольної послідовності може бути різним, на що

впливає співвідношення кількості образів з пропусками і без пропусків в початковій таблиці.

Алгоритм відновлення пропусків EvoGap буде таким:

1. Ініціалізація K_{max} хромосом-розв'язків вибіркової послідовності.

2. $K = 1$

3. Навчання нейромережі на точках навчальної послідовності, де значення пропусків заповнені значеннями K-ї хромосоми. При цьому розв'язується задача пошуку

$$M_K = \min_W \frac{1}{2} \sum_{i=1}^{P_0} (\hat{a}_{in+1} - a_{in+1})^2,$$

де W – матриця вагових коефіцієнтів нейромережі, P_0 – кількість образів в навчальній послідовності.

4. Обчислення цільової функції ГА (фітнес функції):

$$G_K = \frac{1}{2} \sum_{i=1}^{P_C} (\hat{a}_{in+1} - a_{in+1})^2,$$

де P_C – кількість образів контрольної послідовності. Якщо $G_K < G_{min}$, то перехід на крок 7.

5. $K = K+1$. Якщо $K > K_{max}$, де K_{max} – кількість елементів у вибірковій послідовності, то перехід на крок 6, інакше перехід на крок 3.

6. Виконання процедур кроссоверу, мутації, визначення і відбір хромосом наступної епохи. Перехід на крок 2.

7. Закінчення алгоритму.

Еволюційний метод відновлення пропусків в даних має ряд переваг. Так, його використання не вимагає виконання обмежень на вихідну інформацію. Таблиця початкових даних може мати довільну розмірність і структуру пропусків. Перспективним є дослідження ефективності використання нейромережі з неітеративними алгоритмами навчання. Необхідно з'ясувати вплив розподілу значень факторів на точність відновлення пропусків. Додаткові дослідження у вказаних напрямках дозволяють сформувати методику відновлення пропусків з використанням еволюційного підходу.

7.6 Завдання для лабораторної роботи №7

1. Відновлення пропусків серед значень результуючої характеристики.
   Відновити пропущені значення серед даних, наведених у таблю 3, використовуючи методи resampling-1, resampling-2 та метод Бартлетта. Порівняти одержані результати. Для їх контролю та верифікації вважати, що у таблиці представлена залежність $Y = 3X_1X_3 + exp(X_2 - X_1) + \sin(X_1 + X_3)$.

Табл. 3:

№	X_1	X_2	X_3	Y
1	1	3	4	18.43013
2	3	2	4	37.02487
3	5	6	4	*
4	8	7	6	145.3585
5	5	4	3	46.35724
6	4	3	4	49.35724
7	5	4	5	*
8	3	4	2	19.75936
9	2	3	1	8.8594
10	5	4	3	46.35724

2. Відновлення пропусків серед значень вхідних факторів.
   Відновити пропущені значення серед значень вхідних факторів. Вихідні дані є статистичною інформацією про виробництво і споживання різних видів енергії. Використати емпіричні методи та еволюційне моделювання. Порівняти одержані результати.

№	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
1	2.97	31.72	1.43	1.45	0.88	1.59	29.0	0.0	2.97	31.98
2	2.98	35.54	1.89	*	0.79	1.47	31.63	0.0	2.98	34.62
3	2.78	40.15	2.75	2.79	1.46	2.29	37.41	0.0	2.78	40.21
4	2.93	*	4.0	4.19	1.02	*	42.14	0.01	2.93	45.09
5	3.40	50.68	5.4	5.89	1.38	1.83	50.58	0.04	3.40	54.02
6	4.08	63.50	7.47	8.34	1.94	2.63	63.52	0.24	4.08	67.84
7	4.27	62.72	8.54	9.53	1.55	2.15	*	0.41	4.27	69.29
8	4.40	63.92	10.3	11.39	1.53	2.12	67.7	0.58	4.40	72.70
9	*	63.58	13.47	14.61	1.43	2.03	70.32	0.91	4.43	75.71
10	4.77	62.37	13.13	14.3	1.62	2.20	67.91	1.27	4.77	73.99
11	4.72	61.36	*	14.03	1.76	2.32	65.35	1.9	4.72	72.0
12	4.77	61.6	15.67	16.76	1.6	2.17	*	2.11	4.77	76.01
13	4.25	62.05	18.76	19.95	1.44	2.05	70.99	2.7	4.25	78.0
14	5.04	63.14	17.82	19.11	1.08	1.92	71.86	3.02	5.04	79.99
15	5.17	65.95	17.93	19.46	1.75	2.86	72.89	2.78	5.17	80.90
16	5.49	*	14.66	15.80	2.42	3.69	69.98	2.74	5.49	78.29
17	5.47	67.01	12.64	13.72	2.94	4.31	67.75	3.01	5.47	76.34
18	5.99	66.57	10.78	11.86	2.79	4.61	64.04	3.13	5.99	73.25
19	6.49	64.11	10.65	11.75	2.04	3.69	63.29	3.20	6.49	73.1
20	6.43	68.83	11.43	12.47	2.15	3.79	66.62	3.55	6.43	76.74
21	6.03	67.65	*	11.78	2.44	4.20	66.22	4.08	6.03	76.47

№	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8 X_9	Y	
22	6.13	67.09	13.20	*	2.25	4.02	66.15	4.38	6.13	76.78
23	5.69	67.61	14.16	15.40	2.09	*	68.63	4.75	5.69	79.23
24	5.49	68.95	15.75	17.30	2.50	4.37	71.66	*	5.49	82.84
25	6.29	69.36	17.16	18.77	*	4.66	73.02	5.60	6.29	84.96
26	6.13	70.77	17.12	18.82	2.77	4.75	72.46	6.10	6.13	84.70
27	6.16	70.41	16.35	18.33	2.85	5.14	72.00	6.42	6.16	84.64
28	5.91	69.98	16.97	19.37	2.68	4.94	73.52	6.48	*	85.99
29	6.16	68.30	18.51	21.27	1.96	4.26	75.05	6.41	6.16	87.62
30	6.06	70.71	19.24	22.39	1.88	4.06	76.48	6.69	6.06	89.28
31	6.67	71.18	18.88	22.26	2.32	4.51	77.49	*	6.67	91.25
32	*	72.50	*	23.70	2.37	4.63	79.98	7.09	7.14	94.26
33	7.08	72.43	21.74	25.22	2.19	4.51	81.09	6.60	7.08	94.77
34	6.56	72.83	22.91	26.58	2.09	4.30	81.59	7.07	6.56	95.19
35	6.60	71.71	23.13	27.25	1.53	3.71	82.65	7.61	6.60	96.84
36	6.16	71.27	*	28.97	1.53	4.01	84.96	7.86	6.16	98.96
37	5.33	71.88	25.40	30.16	1.27	3.77	3.18	8.03	5.33	96.47
38	5.84	70.76	24.68	29.41	1.03	3.66	3.99	8.14	5.84	97.88
39	6.08	70.01	26.22	*	1.12	4.07	4.49	8.96	6.08	98.31

Література

[1] Айвазян С.А., Мхитарян В.С. *Теория вероятностей и прикладная статистика.* - М.: ЮНИТИ-ДАНА, 2001.

[2] Дрейпер Н., Смит Г. *Прикладной регрессионный анализ: В 2-х кн.* - М.: Финансы и статистика, 1987-88. - Т.1. - 366с., Т.2.- 351с.

[3] Злоба Е., Яцкив И. *Статистические методы восстановления пропущеных значений* // Computer Modelling & New Technologies. - 2002. - Vol.6. - №1. - Pp. 51-61.

[4] Ивахненко А.Г. *Долгосрочное прогнозирование и управление сложными системами.* - К.: Техніка, 1975. - 312с.

[5] Литтл Р. Дж. А., Рубин Д.Б.. *Статистический анализ даных с пропусками.* - М.: Финансы и статистика, 1991. - 336с.

[6] Лук'яненко І.,Краснікова Л. *Економетрика.* - К.: Знання, 1998. - 494с.

[7] Мендель И.Д. *Кластерный анализ.* - М.: Финансы и статистика, 1988. - 176с.

[8] Рассел С, Норвиг П. *Искусственный интелект. Современный подход.* - М.: Вильямс, 2006. - 1408с.

[9] Снитюк В.Є. *Прогнозування. Моделі. Методи. Алгоритми. Навчальний посібник.* - К.: "Маклаут 2008. - 364с.