Розділ 1. ЗАГАЛЬНІ ПРИНЦИПИ ПОБУДОВИ МОДЕЛЕЙ СКЛАДНИХ ОБ'ЄКТІВ ТА ВИРІШАЛЬНИХ ПРАВИЛ НА ЦИХ МОДЕЛЯХ

1.1. Сутність моделювання та задачі, що розв'язуються за допомогою моделювання складних об'єктів

1.1.1. Основні поняття моделювання

При дослідженні об'єктів для знаходження розумного рішення, пов'язаного з керуванням або оптимізацією, потрібно скласти математичну модель системи або процесу, що досліджуються.

Моделювання передбачає заміщення одного об'єкта, котрий є оригіналом, іншим, тобто його моделлю, і фіксацію чи вивчення властивостей оригіналу шляхом дослідження властивостей моделі [1]. Заміщення проводиться з метою спрощення, здешевлення, прискорення фіксації чи вивчення властивостей оригіналу.

Під об'єктом-оригіналом можна розуміти будь-яку природну чи штучну, реальну чи уявну систему. Вона має певну множину параметрів A_0 і характеризується певними властивостями. Множина параметрів системи і їх значень відображає її внутрішній зміст, тобто структуру і принципи функціонування чи існування. Кількісною мірою властивостей системи є множина характеристик Y_0. Система виявляє свої властивості під впливом зовнішніх вхідних дій X.

Характеристики системи знаходяться у функціональній залежності від її параметрів. Кожна характеристика системи $y_0 \in Y_0$ визначається в основному чи повністю обмеженою підмножиною параметрів $\{a_{0k}\} \subset A_0$. Інші параметри не впливають, чи практично не впливають, на значення даної характеристики системи. Дослідника, як правило, цікавлять тільки певні конкретні характеристики $\{y_{0k}\} \subset Y_0$ при конкретних зовнішніх діях $\{x_{0k}\} \subset X_0$.

Таким чином, модель є теж системою зі своїми множинами параметрів A_m і характеристик Y_m. Оригінал і модель схожі за одними параметрами і різні за іншими. Заміщення одного об'єкта іншим правомірне, коли характеристика оригіналу і моделі, котрі інтересують дослідника, визначаються однотипними підмножинами параметрів і пов'язані однаковими залежностями з цими параметрами. При однакових зовнішніх діях $\{x_{0n}\}$ за певний час t для оригіналу і моделі характерні залежності

$$y_{0k} = f(\{a_{01}\}, \{x_{0n}\}, t), \tag{1.1}$$

$$y_{mk} = f(\{a_{m1}\}, \{x_{mn}\}, t), \tag{1.2}$$

де y_{mk} – k-та характеристика моделі; x_{mn} – зовнішня дія на модель; t – модельний час, тобто час, на протязі котрого на модель здійснюються зовнішні дії $\{x_{mn}\} \subset X$ і вимірюються характеристики $\{y_{mn}\} \subset Y$.

За допомогою моделювання може розв'язуватись як пряма, так і обернена задача. У першому випадку множина характеристик моделі y_{mk} є відображенням множини характеристик оригіналу y_{ok}. У другому випадку при дослідженні складних природних систем склад елементів і принципи їх взаємодії мало вивчені, тобто відсутня достатня кількість відомостей про множину параметрів $\{a_{oi}\}$, і за допомогою моделювання розв'язується обернена задача. Будують певну модель, визначають її характеристики y_{mk} при еквівалентних зовнішніх діях $\{x_{mn}\}$, і якщо має місце відображення φ: $y_{ok} \rightarrow y_{mk}$ з певною відомою функцією φ, то вважають, що система-оригінал має такі ж параметри.

Теорія моделювання являє собою взаємозв'язану сукупність положень, визначень, методів і засобів створення та вивчення моделей. Ці положення, визначення, методи і засоби, як і самі моделі, є предметом теорії моделювання. Основна задача теорії моделювання полягає в тому, щоб дати в розпорядження дослідників технологію створення таких моделей, котрі б достатньо точно і повно фіксували властивості оригіналів, до яких виявлено інтерес, простіше і швидше піддавалися дослідженню і допускали перенесення його ресурсів на оригінали.

Теорія моделювання є основною складовою загальної теорії систем, де як головний принцип постулюються здійснювані моделі, і де система може бути зображена скінченною множиною моделей, кожна з котрих відображає певну грань її суті [1, 2].

Система – це сукупність взаємозв'язаних елементів, котрі об'єднані в одне ціле для досягнення певної мети. Під метою розуміють сукупність результатів, що визначаються призначенням системи. Наявність мети пов'язує елементи в систему.

Елемент системи – це мінімальний неподільний об'єкт системи. Елемент належить системі тому, що він зв'язаний з іншими елементами системи так, що множина елементів складає систему. Множину елементів, що складає систему, неможливо розбити на дві чи більше непов'язаних підмножин. Видалення елемента з системи обов'язково змінює її властивості в напрямку, відмінному від мети.

Складна система є множиною взаємозв'язаних і взаємодіючих між собою елементів і підсистем різної фізичної природи, що складають нероздільне ціле, котрі забезпечують виконання системою певної складної функції та описуються складною математичною моделлю.

Підсистема є сукупністю елементів, її поняття використовується в тому випадку, коли підсистема є достатньо самостійною частиною складної

системи, але мета її функціонування виходить із загальної мети функціонування системи. Розбиття складних систем на підсистеми називають декомпозицією систем. На сьогоднішній час цей процес неформалізований і носить евристичний характер.

Проста система – система, яка складається з малої кількості елементів, або модель якої можна віднести до розряду простих.

Реальні системи описуються шляхом визначення їх функцій і структур.

Функція системи – це правило одержання результатів, що визначається метою чи призначенням системи. Поведінку системи при цьому описують певною системою понять: відношень між змінними, векторами, множинами та ін. Функція встановлює, що робить система для досягнення поставленої мети і не визначає, як побудована система. Функціонувати – це означає реалізовувати функцію, тобто одержувати результати згідно з призначенням системи. Для опису функцій систем використовуються теорії множин, алгоритмів, випадкових процесів, інформації та ін.

Структура системи – це фіксована сукупність елементів і зв'язків між ними. В загальній теорії систем під структурою розуміють тільки множину зв'язків між елементами. Вона відображає тільки конфігурацію системи безвідносно до елементів, з яких вона складається. На практиці поняття "структура" містить не тільки множину зв'язків, але й множину елементів, між котрими існують зв'язки. Найбільш часто структуру системи відображають у вигляді графа, де елементи зображені вершинами графа, а зв'язки між ними – дугами.

Керування – процес збору, обробки і передачі інформації. За ступенем централізації керування системи розподіляють [2]:

- на централізовані системи керування, в котрих керування зосереджено в єдиному центрі;
- на децентралізовані системи, в котрих функція керування розподілена між головними і периферійними пристроями;
- на змішані системи.

Складні системи мають, як правило, ієрархічну структуру з кількома рівнями керування.

Визначення моделювання. При проектуванні, зокрема автоматизованому, розробник оперує не з самими об'єктами, а з їх моделями. Моделювання в даному випадку виступає і як апарат, і як засіб, за допомогою котрого створюється проект складної системи.

У широкому значенні під моделюванням розуміють процес адекватного відображення найбільш істотних сторін досліджуваного об'єкта чи явища з точністю, яка необхідна для практичних потреб. В загальному випадку моделюванням називають також особливу форму опосередкування, основою котрого є формалізований підхід до дослідження складної системи.

Теоретичною базою моделювання є теорія подібності.

Подібність – це взаємно однозначна відповідність між двома об'єктами, при котрій відомі функції переходу від параметрів одного об'єкта до параметрів іншого, а математичні описи цих об'єктів можуть бути перетворені в тотожні. Теорія подібності дає можливість встановити наявність подібності чи дозволяє розробити спосіб її одержання.

Таким чином, моделювання – це процес представлення об'єкта дослідження адекватною (подібною) йому моделлю і проведення експериментів з моделлю для одержання інформації про об'єкт дослідження. Під час моделювання модель є як засобом, так і об'єктом досліджень, що знаходиться у відношенні подібності до об'єкта, котрий моделюється.

Іншими словами, модель – це фізична чи абстрактна система, котра адекватно відображає собою об'єкт дослідження.

1.1.2. Класифікація моделей

В основу класифікації моделей може бути покладена ступінь абстрагування моделі від оригіналу. В цьому випадку всі моделі можна розділити на дві групи: матеріальні (або фізичні) і абстрактні (або математичні) [1, 2].

Фізичні моделі утворюються із сукупності матеріальних об'єктів. Для їх побудови використовуються різні фізичні властивості об'єктів, причому природа матеріальних елементів, що застосовуються в моделі, не обов'язково така ж, як і в об'єкті, котрий досліджується. Фізичною моделлю взагалі називають систему, котра еквівалентна чи подібна оригіналу, чи процес функціонування котрої такий же, як і оригіналу, і має ту ж фізичну природу. У свою чергу, фізичні моделі можна поділити на натурні, квазінатурні, масштабні та аналогові.

Натурні моделі – це реально досліджувані системи, їх називають макетами і дослідними зразками. Натурні моделі повністю адекватні системі-оригіналу. Це забезпечує високу точність і достовірність результатів моделювання. Зазвичай, випробуванням дослідних зразків завершується процес проектування складних пристроїв та систем.

Квазінатурні моделі – це сукупність натурних та математичних моделей. Цей вид моделей використовується тоді, коли математична модель як складова частини системи не є задовільною (наприклад, модель людини-оператора) чи коли одна складова системи повинна бути досліджена у взаємодії з іншими складовими, але їх на даний момент не існує, чи включення їх в модель ускладнене, або має велику вартість. Прикладами квазінатурних моделей є реальні АСУ, котрі досліджуються сумісно з математичними моделями відповідних виробництв.

Масштабні моделі – це системи тієї ж фізичної природи, що і оригінал,

але відрізняється від нього масштабами. Методологічною основою такого моделювання є теорія подібності, котра передбачає дотримання геометричної подібності оригіналу і моделі та відповідних масштабів для їх параметрів.

Аналогові моделі – це системи, котрі мають фізичну природу, що відрізняється від оригіналу, але подібні до оригіналу процеси функціонування. Обов'язковою умовою при цьому є однозначна відповідність між параметрами об'єкта і його моделі, а також тотожність безрозмірних математичних описів процесів, що протікають в них. Для створення аналогової моделі необхідний математичний опис системи, що вивчається. Як аналогові моделі використовуються механічні, гідравлічні та інші моделі, але найбільшого поширення набули електричні та електронні аналогові моделі. В них величина струму чи напруги є аналогами фізичних величин іншої природи.

Абстрактні моделі – це опис об'єкта дослідження на певній мові. Абстрактність моделі виявляється в тому, що її компонентами є певні поняття, наприклад, словесні описи, креслення, схеми, графіки, таблиці, алгоритми чи програми, математичні описи, а не фізичні елементи. Серед абстрактних моделей розрізняють гносеологічні, інформаційні, сенсуальні (чуттєві), концептуальні та інші.

Гносеологічні моделі направлені на вивчення об'єктивних законів природи, зокрема, наприклад, моделі сонячної системи, біосфери, світового океану, катастрофічних явищ природи та ін.

Інформаційні моделі описують поведінку об'єкта-оригінала, але не копіюють його (приклад – інформаційний опис конкретного мікропроцесора).

Сенсуальні моделі – моделі певних почуттів, емоцій, чи моделі, котрі виявляють дії на почуття людини (наприклад, музика, поезія, живопис).

Концептуальні моделі – це абстрактні моделі, що виявляють причинно-наслідкові зв'язки, котрі притаманні досліджуваному об'єкту. Ці зв'язки істотні в рамках певного дослідження. Основним призначенням концептуальної моделі є виявлення набору причинно-наслідкових зв'язків, врахування котрих необхідне для одержання потрібних результатів. Той самий об'єкт може бути представлений різними концептуальними моделями, котрі будуються залежно від мети дослідження. Наприклад, одна концептуальна модель може відображувати часові аспекти функціювання системи, а інша – вплив відмов на працездатність системи.

Математичні моделі – це абстрактні моделі, представлені на мові математичних відношень. Вони мають форму функціональних залежностей між параметрами, що враховуються відповідними концептуальними моделями. Ці залежності конкретизують причинно-наслідкові зв'язки, котрі виявлені в концептуальній моделі, і характеризують їх кількісно. Математична модель являє собою формалізований опис системи за

допомогою абстрактної мови, зокрема, за допомогою математичних співвідношень, котрі відображують процес функціонування системи. Для складання математичної моделі можна використати будь-які математичні засоби: алгебраїчне, диференціальне, інтегральне числення, теорію множин, теорію алгоритмів, теорію інформації та кодування та ін. Математичні моделі за методом їх дослідження та ознакою подальшого використання поділяють на аналітичні, чисельні, імітаційні, діагностичні та ін.

Аналітичні моделі – такий формалізований опис системи, котрий дозволяє одержати розв'язок рівняння в явному вигляді, використовуючи відомий математичний апарат.

Чисельні моделі – характеризуються залежністю такого виду, котрий допускає тільки часткові чисельні розв'язки для конкретних початкових умов і кількісних параметрів моделі.

Імітаційне моделювання в широкому розумінні являє собою моделювання з використанням комп’ютерів та змістовний опис об'єктів дослідження у формі алгоритмів [3]. Імітаційна модель – це модель, в котрій враховуються такі особливості, як наявність в самій системі елементів неперервної та дискретної дії, нелінійні співвідношення будь-якого характеру, які описують зв'язки між елементами, дію численних факторів складної фізичної природи. Засобами формалізованого опису імітаційних моделей в основному є універсальні і спеціальні алгоритмічні мови.

Для врахування певних особливостей і для визначення характерних рис оригіналу математичні моделі можуть поділятися на детерміновані та імовірнісні.

Детермінована модель – це модель, в котрій у заданий момент часу характеристики стану однозначно визначаються через вказані величини.

Імовірнісна (стохастична) модель – це модель, в котрій за допомогою математичних співвідношень можна визначити лише розподіл характеристик стану системи за заданими імовірнісними характеристиками її параметрів, вхідних сигналів, початкових умов.

Таким чином, модель – це спеціальний об'єкт, котрий в певних відношеннях заміщує оригінал. У принципі не існує моделі, котра була б повним еквівалентом оригіналу. Будь-яка модель відображає лише певні сторони оригіналу. Тому, з метою одержання найбільш повних знань про оригінал доводиться користуватися сукупністю моделей. Складність моделювання як процесу полягає у відповідному виборі такої сукупності моделей, котрі заміщують реальний пристрій чи систему в необхідних співвідношеннях.

1.1.3. Основні етапи моделювання

Для моделювання реальної системи необхідно створити відповідну модель та провести її дослідження. Перед створенням моделі необхідно конкретизувати мету, а після її дослідження провести аналіз результатів моделювання. Процес створення моделі здійснюється в кілька етапів. Він починається з вивчення системи S_0 і зовнішніх дій X і завершується розробкою чи вибором математичної моделі, або програми для обчислювальної системи, якщо моделювання проводиться за її допомогою.

Процес моделювання передбачає такі етапи:

- формулювання мети моделювання;
- вибір засобів моделювання;
- розробка концептуальної моделі;
- підготовка вихідних даних;
- розробка математичної моделі;
- вибір методу моделювання;
- розробка програмної моделі;
- перевірка адекватності та корегування моделі;
- планування машинних експериментів,
- комп'ютерне моделювання;
- аналіз результатів моделювання.

Трудомісткість кожного з етапів для різних систем може бути різною. У процесі моделювання конкретної системи можуть мати місце зміни технології моделювання. Наприклад, в процесі моделювання може бути наперед відомий метод моделювання чи його засоби. Математична модель може виявитись настільки простою, що не буде необхідності в машинних експериментах. Тобто, в процесі моделювання конкретного пристрою чи системи окремі етапи, в залежності від їх важливості, можуть вилучатися.

Розглянемо більш детально перелік задач, що вирішуються на кожному з відмічених етапів.

Формулювання мети моделювання. На цьому етапі повинно бути досягнуте повне розуміння між замовником і розробником моделі. Важливість коректного виконання цього етапу полягає в тому, що наступні етапи проводяться з орієнтацією на дану мету моделювання. На цьому ж етапі конкретизується, в якому вигляді (якісні чи чисельні градації, точність вимірювання та ін.) будуть представлені вихідні дані.

Перевірка адекватності моделі відмічена окремим етапом (див. далі), але адекватність моделі забезпечується якісним виконанням практично всіх етапів. Тому перевірка адекватності моделі повинна проводитись в тому чи іншому вигляді, починаючи з розробки концептуальної моделі і закінчуючи аналізом результатів моделювання.

Під розробкою математичної моделі розуміють створення повністю

формалізованого опису динаміки функціонування пристрою чи системи. Не завжди для цього можна підібрати відомий метод формалізації чи конструктивний математичний апарат. Але це слід намагатися зробити, тобто розробити однозначні залежності вихідних характеристик від параметрів і зовнішніх дій для кожної складової системи, алгоритми взаємодії між складовими, логічні умови зміни станів.

Результати машинного моделювання повинні бути проаналізовані з метою перевірки їх достовірності і вироблення рекомендацій про способи підвищення якості системи, що досліджується.

На всіх етапах моделювання слід звертати особливу увагу на документування рішень, що приймаються, допусків, обмежень і висновків.

Для тієї самої системи можна скласти множину моделей $\{S_m\}$. Моделі будуть відрізнятися ступенем деталізації і врахуванням тих чи інших особливостей і режимів функціонування, відображувати певну грань системи, орієнтуватися на дослідження певної властивості чи групи властивостей системи. Тому перед розробкою моделі необхідно сформулювати цілі дослідження. При створенні чи модернізації будь-якої системи постає задача визначення її ефективності. Якщо розробляється кілька варіантів системи, то з них необхідно вибрати найкращий. Вирішення цих задач і є основною метою моделювання. Взагалі в техніці використовується поняття техніко-економічної ефективності, котре враховує витрати і вимірювані характеристики системи:

$$E = E(Y_0) \tag{1.3}$$

Елементи множини характеристик $y_{0k} \in Y_0$ є частковими показниками якості системи: продуктивність, надійність, вартість, маса, габаритні розміри і т. д. Якщо функція (1.3) відома, тобто виражена аналітично, то показник ефективності E можна обчислити за множиною параметрів системи $\{a_0\}$ при певних зовнішніх діях $\{x_{0n}\}$. У противному разі використовується один з двох підходів: однокритеріальна чи багатокритеріальна оцінка.

При однокритеріальній оцінці обмежуються оцінкою ефективності системи за одним частковим показником якості y_0, а на інші характеристики накладають обмеження на їх допустимі зміни. При цьому можна одержати декілька варіантів систем з однаковим, чи приблизно однаковим, значенням y_0 при суттєво різних інших часткових показниках якості. В цьому випадку не можна з певною впевненістю визначити більш раціональний варіант.

При багатокритеріальній оцінці невідомий вид функції (1.3) штучно представляється у формі узагальненого чи інтегрального критерію. Однією з найбільш поширених форм представлення є адитивний критерій:

$$E = \sum_{i=1}^{n} b_i y_i \ , \tag{1.4}$$

де вагові коефіцієнти b_i узгоджують шкали вимірювань характеристик y_i та задовольняють умову

$$\sum_{i=1}^{n} b_i = 1, \quad \forall b_i > 0 \ . \tag{1.5}$$

Таким чином, визначення мети моделювання полягає, в першу чергу, у виявленні виду критерію ефективності E системи, що досліджується, а це, в свою чергу, передбачає задавання скінченої множини характеристик $\{y_i\}$, їх вагових коефіцієнтів і допустимих меж вимірювань.

Якщо метою моделювання є не просто фіксація властивостей системи, але й оптимізація системи, то перед моделюванням слід виявити ті параметри системи, котрі дослідник може змінювати.

Етап формулювання мети моделювання завершується оцінкою доцільності проведення моделювання. На цьому етапі необхідно зіставити витрати на розробку та модернізацію системи з економічним ефектом від її впровадження.

Вибір засобів моделювання та рівні моделювання. Методика моделювання безпосередньо залежить від рівня моделювання, тобто від ступеня деталізації опису об'єкта. Кожному рівню моделювання ставиться у відповідність певне поняття системи, елемента системи, закону функціонування елементів системи в цілому і зовнішніх дій. Наприклад, залежно від ступеня деталізації опису обчислювальних систем, їх пристроїв та елементів можна виділити три основних рівні моделювання:

1) рівень структурного чи імітаційного моделювання з використанням алгоритмічних моделей системи (моделюючих алгоритмів) із застосуванням спеціалізованих мов моделювання, теорій множин, алгоритмів, формальних граматик, графів, масового обслуговування, статистичного моделювання;

2) рівень логічного моделювання функціональних схем, елементів і вузлів обчислювальних систем, моделі котрих можна представити у вигляді рівнянь безпосередніх зв'язків (логічних рівнянь) і будувати з використанням апарату двозначної чи багатозначної алгебри логіки;

3) рівень кількісного моделювання (аналізу) принципових схем елементів обчислювальних систем, моделі котрих можна представити у виді систем нелінійних алгебраїчних рівнянь чи інтегро-диференціальних рівнянь і котрі можна досліджувати з використанням методів функціонального аналізу, теорії диференціальних рівнянь, математичної статистики.

Сукупність моделей об'єкта на структурному, логічному, кількісному

рівнях моделювання є ієрархічною системою, котра розкриває взаємозв'язок різних сторін опису об'єкта і забезпечує системну зв'язаність його елементів і властивостей на всіх стадіях процесу проектування обчислювальної системи. При переході на більш високий рівень абстрагування здійснюється згортка даних про об'єкт, що моделюється, а при переході до більш детального рівня опису – розгортка цих даних.

На структурному рівні моделюється взаємодія елементів об'єкта більш низького рівня. Топологічною моделлю в цьому випадку слугує орієнтований граф $\boldsymbol{G(V,D)}$, складання котрого базується на змістовому описі складу (множина вершин $\boldsymbol{V}$) і способу дії об'єкта (множина ребер $\boldsymbol{D}$). Вершинами орграфа v_i, є функціонально закінчені блоки об'єкта, а ребрами d_i – інформаційні зв'язки між ними.

Структурні відношення між елементами множини $\boldsymbol{V}$ описуються матрицею суміжності

$$[\boldsymbol{G_{ij}}]_V=[n\text{x}n], \tag{1.6}$$

рядки і стовпці котрої відповідають вершинам орграфа структурної моделі, а її $\boldsymbol{G_{ij}}$-й елемент дорівнює кількості ребер, напрямлених від вершини v_i до вершини v_j.

Відношення між елементами множини $\boldsymbol{V}$ і $\boldsymbol{D}$, тобто між вершинами і ребрами орграфа, описуються у вигляді булевої матриці інцидентності

$$[\boldsymbol{a_{ij}}]_{v,d}=[n\text{x}m], \tag{1.7}$$

рядки якої відповідають вершинам, а стовпці – ребрам орграфа; при цьому її $\boldsymbol{a_{ij}}$-й елемент дорівнює +1, якщо v_i – початкова вершина ребра d_i, і –1, якщо v_i – кінцева вершина d_i.

На логічному рівні моделювання кожній множині, булевій матриці бінарних відношень чи структурному графу відповідають набори логічних відношень між елементами, що в них входять, котрі представлені у вигляді логічних змінних. Множинам $\boldsymbol{V}$ і $\boldsymbol{D}$ також відповідають певні логічні відношення, які відображують причинно-наслідкові зв'язки. Ці зв'язки описують послідовності зміни станів об'єкта з врахуванням станів інших об'єктів і необов'язково суміжних з ним.

При кількісному моделюванні кожному елементу множини булевої матриці чи логічної змінної ставиться у відповідність алгебраїчна чи інша кількісна змінна, а логічні відношення переходять у кількісні відношення, наприклад, рівняння чи нерівності.

На кожному з основних рівнів моделювання можливі описи об'єктів з різним ступенем повноти і узагальнення за причиною існування різних ступенів деталізації структурних, логічних і кількісних властивостей і відношень. Але сама по собі задача побудови необхідної моделі, котра б

достатньо точно відображувала характерні властивості об'єкта чи його елемента (блока) на даному рівні проектування і в той же час була доступною для дослідження, є досить складною.

Розробка концептуальної моделі та технологія моделювання. У концептуальній моделі, зазвичай у словесній формі, наводяться відомості про природу і параметри елементарних явищ досліджуваної системи, про вид і ступінь взаємодії між ними, про місце і значення кожного елементарного явища в загальному процесі функціонування [2].

Спочатку концептуальна модель системи $\boldsymbol{S}_0$ виникає в свідомості дослідника. Модель орієнтується на виявлення певних властивостей системи відповідно до цілей моделювання. Для цього дослідник робить як би подумки зріз системи в "площині" тієї метасистеми $\boldsymbol{M}$, в котрій знаходиться система $\boldsymbol{S}_0$ (рис.1.3). Таку операцію називають $\boldsymbol{M}$-орієнтацією. Потім дослідник виявляє основні ознаки орієнтованої моделі і може додавати ще декотрі ознаки та умови, які полегшать дослідження моделі чи дозволять зобразити її у вигляді певного зрізу модельованої системи.

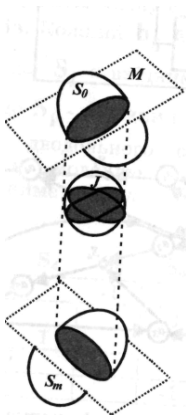

Рис. 1.1. Відображення оригіналу $\boldsymbol{S}_0$ і моделі $\boldsymbol{S}_m$ у свідомості дослідника

Розробка концептуальної моделі потребує достатньо глибоких знань про систему $\boldsymbol{S}_0$. Необхідно обґрунтувати те, що повинне ввійти в модель, а також те, що може бути відкинуте без істотних перекручувань результатів моделювання. Основною проблемою при створенні моделі є знаходження компромісу між простотою моделі та її адекватністю з досліджуваною системою. Розробник моделі, керуючись своїми знаннями системи, оціночними розрахунками, досвідом, приймає рішення про виключення того чи іншого елемента чи явища з моделі без достатньо повної впевненості у тому, що це не внесе істотних похибок в результати моделювання. Напевне, що результат створення концептуальної моделі ніколи не буде повністю формалізованим. Тому інколи кажуть, що моделювання є не тільки наукою, але й мистецтвом.

Наступним кроком при створенні концептуальної моделі є вибір рівня деталізації моделі. Проблема вибору рівня деталізації моделі може бути вирішена шляхом побудови ієрархічної послідовності моделей. На кожному рівні існують характерні особливості системи, змінні, принципи і залежності, за допомогою котрих описується поведінка системи.

Рівні деталізації інколи називають стратами, а процес виділення рівнів – стратифікацією. Вибір страт залежить від цілей моделювання і ступеня попереднього знання елементів.

При побудові стратифікованої концептуальної моделі в неї повинні, в першу чергу, увійти параметри $\{S_{0j}\}$, які забезпечують визначення характеристик, що інтересують дослідника Y_{0k} при конкретних зовнішніх діях $\{X_{0n}\}$ на заданому інтервалі часу t функціонування системи.

Деталізація системи повинна проводитись до того рівня, щоб для кожного елемента були відомі чи могли бути одержані залежності параметрів реакцій елемента, котрі істотні для функціонування системи, а також визначення залежності характеристик системи від параметрів дії, які є вихідними для цього елемента. Якщо за результатами орієнтації, стратифікації і розчленування одержуємо модель великої розмірності, тобто з великою кількістю параметрів, великим числом елементів (сотні і тисячі), то її необхідно спростити. Це можна зробити перетвореннями моделі без зниження ступеня адекватності, у тому числі шляхом декомпозиції системи на підсистеми, інтеграції елементарних операцій і відповідної інтеграції елементів, виключення другорядних технологічних процесів з виключенням елементів, котрі їх забезпечують.

Наступним кроком створення концептуальної моделі є її локалізація, котра здійснюється шляхом представлення зовнішнього середовища у вигляді генераторів зовнішніх дій, які включаються в склад моделі як елементи. При необхідності вони диференціюються на генератори робочого навантаження, котрі подають на вхід системи основні вихідні об'єкти – дані для інформаційних систем, у тому числі для ОС; генератори додаткових об'єктів і енергії; генератори керуючих і активізуючих дій (рис. 1.2.).

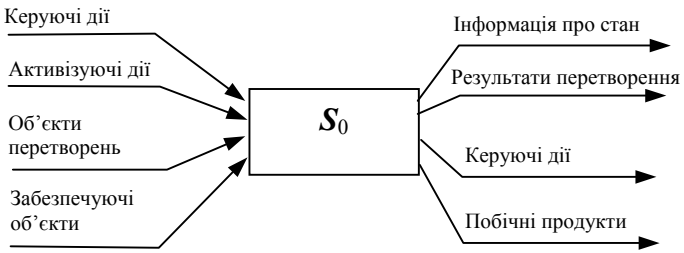

Рис. 1.2. Представлення процесу локалізації моделі

Структуризація та керування. Побудова структури концептуальної моделі завершується показом зв'язків між елементами. Зв'язки розподіляються на речовинні та інформаційні. Речовинні зв'язки відображують можливі шляхи переміщення продукту перетворення від

одного елемента до іншого. Інформаційні зв'язки забезпечують передачу керуючих дій між елементами та інформацію про стан. У концептуальній моделі повинні бути конкретизовані всі правила чи алгоритми керування робочим навантаженням, елементами і процесами.

Виділення процесів. У разі необхідності розгляду динаміки системи модель, що відображає статику системи, доповнюють описом функціонування системи. Функціонування системи полягає у виконанні технологічного процесу перетворення речовини, енергії чи інформації. У складних системах одночасно може протікати кілька технологічних процесів. Як приклад, можна згадати мультипрограмний режим сучасних обчислювальних систем. Технологічний процес являє собою певну послідовність окремих елементарних операцій. Частина операцій може виконуватися паралельно різними елементами системи. Задається технологічний процес маршрутною картою, програмою та ін.

Відображення станів. Для систем зі структурним принципом керування набув поширення інший підхід. Для кожного елемента вибирається певний параметр n, значення котрого змінюється в ході функціонування елемента, і він відображає його стан в поточний момент часу $z(t)$. Множина таких параметрів по всіх $n = \overline{1, N}$, елементах системи $\{z_n\}$ відображає стан системи $Z(t)$. Функціонування системи представляється у вигляді послідовної зміни станів: $Z(t_0)$, $Z(t_1)$,..., $Z(t_m)$. Множину $\{Z\}$ можливих станів системи називають простором станів. Поточний стан системи в момент часу t,($t_0 < t < t_m$) відображається у вигляді координати точки в m-вимірному просторі станів, а вся реалізація процесу функціонування системи за час t_m=T – у вигляді певної траєкторії.

Якщо відомий початковий стан системи $Z(t_0)$, то можна визначити її стан в будь-який момент t, коли відома залежність

$$Z(t) = F(X, Z_0, Z, t) . \qquad (1.8)$$

Тоді характеристики на виходах системи визначатимуться як

$$Y = W(Z, T). \qquad (1.9)$$

Підготовка вихідних даних. При створенні концептуальної моделі виявляються якісні (функціональні) і кількісні параметри системи $\boldsymbol{S}_0$, а також зовнішні дії. Для кількісних параметрів необхідно визначити їх конкретні значення, котрі будуть використані у вигляді вихідних даних при моделюванні. Це відповідальний і трудомісткий етап роботи. Достовірність результатів моделювання однозначно залежить від точності і повноти вихідних даних. Частина параметрів виявляється ще на ранній стадії концептуального моделювання, інші – паралельно з розробкою концептуальної моделі.

Збір вихідних даних ускладнюється через те, що значення параметрів

можуть бути не тільки детермінованими, але й стохастичними. Не всі параметри є стаціонарними. Особливо це стосується параметрів зовнішніх дій. По-друге, мова йде про моделювання неіснуючої системи (система ще тільки проектується, або модернізується), котра повинна функціонувати в нових умовах.

Взагалі більша частина параметрів є випадковими величинами за своєю природою. Але з метою спрощення моделювання вони в більшості випадків представляються детермінованими середніми значеннями. Це можна робити тільки тоді, коли випадкова величина має невеликий розкид. Заміна імовірнісних величин детермінованими повинна здійснюватись обґрунтовано, щоб це не призвело до зміщення середніх значень.

При створенні моделі детерміновані параметри можуть також замінюватись випадковими. Це робиться при інтеграції елементів системи чи зовнішніх дій з метою скорочення розмірності моделі.

Для випадкових параметрів організується збір статистики і наступна обробка. У процесі обробки виявляється можливість представлення параметрів певними теоретичними законами розподілу. Процедура підбору виду закону розподілу виконується наступним чином. За сукупністю чисельних значень параметра будується гістограма відносних частот – емпірична густина розподілу. Гістограма апроксимується плавною кривою. Одержана крива послідовно порівнюється з кривими густини розподілу різних теоретичних законів розподілу. Обирається один із законів за найкращим збіганням виду порівнюваних кривих. За емпіричними значеннями обчислюють параметри цього розподілу. Потім виконують кількісну оцінку ступеня збігання емпіричного і теоретичного розподілів за тим чи іншим критерієм погодження, наприклад, Пірсона, Колмогорова, Смірнова, Фішера чи Стьюдента. Питання підбору виду закону розподілу детально описує математична статистика [4].

Апроксимація функцій. Для кожного елемента системи існує функціональний зв'язок між параметрами вхідних дій і його характеристиками. Вид функціональної залежності для одних елементів є очевидним, а для інших він може бути вияснений з природи функціонування. Але для певних елементів може бути одержана тільки сукупність експериментальних даних про кількісні значення вихідних характеристик при різних значеннях параметрів. У цьому випадку виникає необхідність ввести певну гіпотезу про характер функціональної залежності, тобто апроксимувати її певним математичним рівнянням. Пошук математичних залежностей між двома і більше змінними за зібраними дослідними даними може виконуватись за допомогою методів регресійного, кореляційного чи дисперсійного аналізу [4].

Висування гіпотез. Щодо параметрів, котрі відображають нові елементи

майбутньої системи чи нові умови функціонування, відсутня можливість збору фактичних даних. Для таких параметрів висуваються гіпотези про їх можливі значення. Важливо, щоб гіпотези висували експерти-спеціалісти, котрі достатньо добре уявляють створювану систему чи нові зовнішні дії на систему. Ступінь суб'єктивності зменшується при одержанні відомостей від групи спеціалістів і при застосуванні методик експертних оцінок.

Закінчується етап збору і оцінки вихідних даних їхньою класифікацією на зовнішні і внутрішні, постійні і змінні, неперервні і дискретні, лінійні і нелінійні, стаціонарні і нестаціонарні, детерміновані і недетерміновані (стохастичні). Для змінних кількісних параметрів, котрими можна варіювати в ході моделювання, визначаються межі їх змін, а для дискретних – можливі значення.

Розробка математичної моделі. Концептуальна модель і кількісні вихідні дані є основою для розробки математичної моделі. Створення математичної моделі має дві основні мети:

1) дати формалізований опис структури та процесу функціонування системи для однозначності їх розуміння;

2) намагатися представити процес функціонування у вигляді, що допускає аналітичне дослідження системи.

Розробка єдиної методики створення математичних моделей до цього часу не уявляється можливою. Це обумовлено великою різноманітністю класів систем (статичні, динамічні, із структурним чи програмним керуванням, з постійною чи змінною структурою і т. ін.). За характером вхідних дій і внутрішніх станів системи підрозділяються на неперервні та дискретні, лінійні і нелінійні, стаціонарні і нестаціонарні, детерміновані і стохастичні. Відповідно існує така ж кількість типів моделей, і їх вибір залежить від ступеня стратифікації і деталізації.

Для певних класів систем розроблено формалізовані схеми і математичні методи, котрі дозволяють описати функціонування системи, а в деяких випадках і виконувати аналітичні дослідження. Засобами формального опису процесів функціонування систем з програмним керуванням є певні мови і системи імітаційного моделювання.

Для опису стохастичних систем з дискретними множинами станів, вхідних і вихідний дій, що функціонують в неперервному часі, широко застосовуються стохастичні мережі. Стохастична мережа є сукупністю систем масового обслуговування, в котрій циркулюють заявки, що переходять з однієї системи в іншу.

Велика група мов імітаційного моделювання ґрунтується на формалізованому представленні систем у вигляді стохастичних мереж. За певних умов стохастична мережа може розглядатись як сукупність незалежних систем масового обслуговування. Це дає додаткові можливості

використання досягнень теорії масового обслуговування для проведення аналітичного моделювання.

В основі систем масового обслуговування лежить поняття приладу, котрий може виконувати скінченну множину операцій. Прилад виконує операцію, коли виникає заявка – вимога на виконання операції. Якщо прилад виконує будь-яку операцію, то вважається, що він зайнятий, у противному разі – прилад вільний. Часова послідовність заявок називається потоком заявок. Загальний потік заявок може складатися з кількох потоків. У випадках незалежності потоків, випадкових моментів надходження чи завершення обслуговування заявок в системі можуть виникати черги. Черга – це заявки, котрі чекають обслуговування, коли прилад зайнятий. Прилад може складатися з кількох каналів, кожний з яких здатний обслуговувати будь-яку заявку. Сукупність приладу, потоків заявок і черг до нього називають системою масового обслуговування.

Теорія масового обслуговування добре розроблена і знайшла широке застосування для створення математичних моделей при моделюванні обчислювальних систем.

Системи, стани котрих визначені в дискретні моменти часу t_0, t_1, t_2,…, одержали назву автоматів. У кожний дискретний момент часу (за виключенням t_0) в автомат надходить вхідний сигнал $x(t)$, під дією якого автомат переходить в новий стан відповідно до функції переходів і видає вихідний сигнал, який визначається функцією виходів. Якщо автомат характеризується обмеженими множинами станів z, вхідних сигналів x і вихідних сигналів y, то його називають кінцевим автоматом. Функції переходів і виходів кінцевого автомата задаються таблицями, матрицями чи графіками. Для формалізованого опису функціонування систем використовується також обчислення висловлювань, мережі Петрі та ін.

Таким чином, побудова математичної моделі передбачає аналіз концептуальної моделі і вихідних даних з метою вибору однієї з підходящих формалізованих схем. Якщо це не вдається зробити для всієї системи, то формалізовані схеми можуть бути використані для опису окремих елементів, а вся система описується за допомогою програмного чи структурного підходу.

Вибір методу моделювання. Розроблена математична модель може бути досліджена різними методами – аналітичними, чисельними, якісними, імітаційними.

Аналітичні методи – методи, за допомогою яких можна провести найбільш повне дослідження моделі. В деяких випадках наявність аналітичної моделі робить можливим застосування математичних методів оптимізації. Для використання аналітичних методів математичну модель перетворюють до виду явних аналітичних залежностей між

характеристиками і параметрами системи та зовнішніми діями. Це вдається для простих систем. Застосування математичних методів для більш складних систем пов'язане з більшим, в порівнянні з іншими методами, ступенем спрощення реальності і абстрагуванням. Тому аналітичні методи дослідження використовуються, як правило, для первинної грубої оцінки характеристик всієї системи чи окремих її підсистем, а також на ранніх стадіях проектування систем, коли недостатньо інформації для побудови більш точної моделі.

Чисельні методи. Ряд моделей не піддається розв'язанню відомими аналітичними методами. Для їх дослідження можуть бути використані чисельні методи. Вони прийнятні для більш широкого класу систем, для котрих математична модель представлена у вигляді системи рівнянь, що допускає розв'язання чисельними методами. Застосування чисельних методів особливо ефективне за допомогою швидкодіючих комп'ютерних систем. Результатом дослідження систем чисельними методами є таблиці значень величин, що знаходяться для скінченного набору значень параметрів системи.

Якісні методи. Якщо одержані рівняння не вдається розв'язати аналітичними чи чисельними методами, то вдаються до якісних методів, які дозволяють в ряді випадків оцінити асимптотичні значення величин, що визначаються, стійкість, а також судити про поведінку траєкторії системи в цілому.

Імітаційні методи. При імітаційному моделюванні динамічні процеси системи-оригіналу замінюються процесами, що імітуються в абстрактній моделі, але з дотриманням таких же співвідношень, тривалостей і часових послідовностей окремих операцій (в масштабі модельного часу). Тому метод імітаційного моделювання інколи називають алгоритмічним чи операційним. У процесі імітації, як при експерименті з оригіналом, фіксують певні події і стани чи виміряють вихідні дії, за котрими обчислюють характеристики якості функціонування системи.

Імітаційне моделювання дозволяє розглядати процеси, що проходять в системі, практично на будь-якому рівні деталізації. В імітаційній моделі можна реалізувати будь-який алгоритм керування чи функціонування системи. Моделі, котрі досліджуються аналітичними методами, можуть бути досліджені і імітаційними методами. Все це сприяє застосуванню імітаційних методів моделювання як основних для моделювання складних систем.

Методи імітаційного моделювання класифікуються залежно від класу досліджуваних систем, способу просування модельного часу, виду кількісних змінних параметрів системи і зовнішніх дій. Більш детально імітаційне моделювання розглядається в пп. 1.1.4.

Розділяють методи моделювання дискретних і безперервних систем.

Систему називають дискретною, якщо всі елементи системи мають обмежену множину станів і перехід з одного стану в інший здійснюється миттєво.

Кількісні параметри системи і зовнішніх дій можуть бути детермінованими й випадковими. За цією ознакою розрізняють детерміноване й статистичне моделювання. При статистичному моделюванні для одержання достовірних імовірнісних характеристик процесів функціонування системи необхідне їх багатократне відтворення з різними конкретними значеннями випадкових факторів і статистичною обробкою результатів вимірювань.

При нестаціонарному характері змінних використовуються спеціальні методи моделювання, зокрема метод повторних експериментів.

Ще одним класифікаційним параметром при створенні математичної моделі вважають схему формалізації. При цьому методи розділяють на алгоритмічний (програмний) чи структурний (агрегатний). У першому випадку процеси керують елементами (ресурсами) системи, а в другому – елементи керують процесами, визначають порядок функціонування системи.

Таким чином, вибір того чи іншого методу моделювання повністю визначається математичною моделлю і вихідними даними.

Вибір засобів моделювання та розробка програмної моделі. Для дослідження моделей застосовуються технічні засоби – універсальні (персональні) комп'ютери чи спеціалізовані обчислювальні системи. Для реалізації аналітичного моделювання за допомогою універсальних комп'ютерів не задаються особливі вимоги до технічних засобів. При статистичному моделюванні необхідно досить багато машинного часу, тому бажано використовувати високопродуктивні комп'ютери.

У зв'язку із широким застосуванням імітаційного моделювання все більш актуальними стають розробка і випуск спеціалізованих обчислювальних засобів. До них відносяться стохастичні машини, машини імітаційного моделювання і гібридні обчислювальні комплекси.

Як програмні засоби можуть бути використані процедурно-орієнтовані алгоритмічні мови, програмно-орієнтовані мови чи автоматизовані системи моделювання.

Програмні та технічні засоби моделювання обираються з врахуванням ряду критеріїв. Необхідна умова при цьому – достатність і повнота засобів для реалізації концептуальної і математичної моделі. Серед інших критеріїв можна назвати доступність засобів, наявність у дослідника інформації про ті чи інші засоби. Важливе значення має простота і легкість засвоєння програмних засобів моделювання, швидкість і коректність створення програмної моделі, існування методики застосування засобів для моделювання систем певного класу.

При розробці програм імітаційного моделювання виникають задачі,

загальні для широкого класу моделей. Це: організація псевдопаралельного виконання алгоритмів; динамічний розподіл пам'яті; операції з модельним часом, що відображає астрономічний час функціонування оригіналу; імітація випадкових процесів; введення масиву подій; збирання і обробка результатів моделювання. Щоб полегшити вирішення цих задач, були створені спеціальні мови моделювання. За структурою і правилами програмування мови моделювання подібні до процедурно-орієнтованих алгоритмічних мов високого рівня. Вони мають той чи інший набір операторів, що супроводжуються відповідними операндами. Але оператори мов моделювання визначають виконання більш складних процедур, тому мови моделювання мають більш високий рівень в порівнянні з алгоритмічними мовами, що спрощує складання програм. Мови моделювання є формалізованим базисом створення математичних моделей. На сьогоднішній час відомо більше 500 мов моделювання.

Для спрощення і прискорення процесу створення машинних моделей був реалізований ряд ідей з автоматизації програмування імітаційних моделей. Створено ряд систем, котрі звільняють дослідника від програмування. Програма створюється автоматично за однією з формалізованих схем на основі заданих дослідником параметрів системи, зовнішніх дій і особливостей функціонування. Вихідні дані представляються тією чи іншою канонічною формою, чи в ході діалогу з обчислювальною системою. За результатами машинного експерименту основні вихідні змінні обчислюються і виводяться автоматично, додаткові – за вказівкою дослідника. Такі системи називають універсальними автоматизованими імітаційними моделями чи генераторами імітаційних програм.

Перевірка адекватності і корегування моделі. Суть перевірки адекватності моделі системи полягає в аналізі її відповідності до досліджуваної системи, а також рівнозначності системі. Але модель не може бути повним відображенням системи, інакше губиться сенс її створення. У процесі створення моделі адекватність порушується в результаті орієнтації, стратифікації, деталізації і локалізації. Крім того, адекватність порушується через ідеалізацію зовнішніх умов і режимів функціонування, виключення тих, чи інших параметрів, не врахування деяких випадкових факторів. Відсутність точних відомостей про зовнішні дії, певні нюанси структури системи, прийняті апроксимації, інтерполяції, пропозиції і гіпотези теж ведуть до зменшення відповідності між моделлю і системою. Все це може стати причиною того, що результати моделювання будуть істотно відрізнятися від реальних.

Найпростішою мірою адекватності може бути відхилення певної характеристики y_{ok} оригіналу і y_{mk} моделі:

$$\Delta y = |y_{ok}-y_{mk}|, \qquad (1.10)$$

або відношення відхилення до характеристики оригіналу

$$\Delta y = |y_{ok}\text{-}y_{mk}|/y_{ok}. \quad (1.11)$$

Тоді можна вважати, що модель адекватна системі, якщо імовірність того, що відхилення Δy не перевищує граничної величини Δ, не менше від припустимої імовірності P_Δ:

$$P(\Delta y<\Delta) \geq P_\Delta. \quad (1.12)$$

Але практичне використання даного критерію не завжди можливе за відсутності інформації про значення характеристики y_{ok} для систем, що проектуються чи модернізуються. На практиці оцінка адекватності зазвичай проводиться шляхом експертного аналізу розумності результатів моделювання, при цьому виконуються такі види перевірок:

- перевірка моделей елементів (у сумнівних випадках слід деталізувати елемент чи провести додатковий аналіз);
- перевірка моделі зовнішніх дій;
- перевірка концептуальної моделі функціонування системи (виявляються помилки постановки задачі);
- перевірка формалізованої і математичної моделі;
- перевірка способів вимірювання і обчислення вихідних характеристик (виявляються помилки розв'язування);
- перевірка програмної моделі (аналізується відповідність операцій і алгоритмів функціонування програмної і математичної моделі, проводяться контрольні розрахунки при типових і граничних значеннях змінних, виявляються інструментальні помилки програмування).

Корегування моделі. Якщо за результатами перевірки адекватності виявляються недопустимі розбіжності моделі і системи, то виникає необхідність в корегуванні моделі. При цьому виділяють такі типи змін: глобальні, локальні і параметричні.

Необхідність в глобальних змінах виникає у випадку виявлення методичних помилок в концептуальній чи математичній моделі. Для усунення таких помилок буде потрібна розробка нової моделі.

Локальні зміни пов'язані з уточненням деяких параметрів чи алгоритмів. Вони виконуються шляхом заміни моделей компонентів системи і зовнішніх дій на еквівалентні, але більш точні моделі. Локальні зміни вимагають часткової зміни математичної моделі.

До параметричних відносяться зміни певних параметрів, котрі називаються калібрувальними. Для цього слід наперед виявити калібрувальні параметри і передбачити прості способи варіювання ними.

Стратегія корегування моделі повинна бути спрямована на першочергове введення глобальних змін, потім – локальних і в кінці –

параметричних.

Завершується етап перевірки адекватності і корегування моделі визначенням і фіксацією області придатності моделі. Під областю придатності розуміють множину умов, при дотриманні котрих точність результатів моделювання знаходиться в допустимих межах.

Планування експериментів з моделлю. Цілі моделювання досягаються шляхом дослідження розробленої моделі. Дослідження полягають в проведенні експериментів, в результаті котрих визначаються вихідні характеристики системи при різних значеннях керованих змінних параметрів моделі. Експерименти проводять за певним планом. Це особливо важливо при уявному і імітаційному моделюванні на універсальних комп'ютерах. Грунтується це на тому, що в такому випадку може бути велика кількість можливих сполучень значень керованих параметрів, а кожний машинний експеримент проводиться при певному сполученні значень параметрів. При обмежених обчислювальних і часових ресурсах проведення всіх експериментів неможливе. Виникає необхідність у виборі певних сполучень параметрів і послідовності проведення експериментів. Цей підхід називається стратегічним плануванням. Сукупність методів зменшення тривалості машинного експерименту при забезпеченні статистичної достовірності результатів моделювання називають тактичним плануванням [3].

Аналіз результатів моделювання. Оскільки вихідні характеристики моделі часто є випадковими величинами чи функціями (зокрема, в результаті імітаційного експерименту), то обробка полягає в обчисленні оцінок математичного очікування, дисперсій і кореляційних моментів. Оцінки, одержані в результаті статистичної обробки вимірювань, повинні бути незміщеними, змістовними і ефективними.

Для виключення необхідності зберігання в комп'ютері всіх вимірювань, обробку результатів вимірювань проводять за рекурентними формулами, коли оцінки обраховують в процесі експерименту методом нарощуваного підсумка в міру появи нових вимірювань. Для стохастичних характеристик будують гістограми відносних частот – емпіричну густину розподілу. Для побудованої гістограми підбирають теоретичний закон розподілу. Це також робиться і при підготовці вихідних даних моделювання.

Для випадкових нестаціонарних характеристик період моделювання T_m розбивається на відрізки з постійним кроком Δt (прогони чи перетини), і запам'ятовуються значення характеристик в кінці кожного прогону. Проводиться серія експериментів із різними послідовностями випадкових параметрів моделі. Потім вимірювання кожного перетину обробляються, як при оцінці випадкових величин.

Визначення залежностей характеристик від параметрів системи. За результатами стохастичного моделювання може бути проведено аналіз

залежностей характеристик від параметрів системи і зовнішній дій. Для цього можна скористатися кореляційним, дисперсійним чи регресійним методами.

За допомогою кореляційного аналізу можна встановити наявність зв'язку між двома чи більше випадковими величинами. Коли коефіцієнт кореляції за абсолютною величиною дорівнює одиниці, то нестохастичний лінійний зв'язок між величинами, що аналізуються, наявний, а коли він дорівнює нулю, то зв'язок відсутній. Проміжні значення коефіцієнта кореляції відповідають наявності лінійного зв'язку з розсіюванням чи нелінійної кореляції.

Дисперсійний аналіз використовують для встановлення відносного впливу різних факторів на значення вихідних характеристик. При цьому загальна дисперсія характеристики розкладається на компоненти, котрі відповідають факторам, що розглядаються. За значеннями окремих компонентів роблять висновок про ступінь впливу того чи іншого фактора на характеристику, що аналізується.

Якщо всі фактори в експерименті є кількісними, то можна знайти аналітичну залежність між характеристиками і факторами. Для цього використовується метод регресійного аналізу. Знайдена залежність називається емпіричною моделлю. Регресійний аналіз полягає в тому, що обирається вид співвідношення між залежними і незалежними змінними, за експериментальними даними обчислюються параметри обраної залежності і оцінюється якість апроксимації експериментальних даних моделлю. Якщо якість незадовільна, то обирається залежність іншого виду, і процедура повторюється.

До аналізу результатів моделювання можна віднести задачу аналізу чутливості моделі до варіацій її параметрів. Під аналізом чутливості розуміють перевірку стійкості характеристик процесу функціонування системи до можливих відхилень значень параметрів.

Аналіз результатів моделювання дозволяє уточнити множину інформативних параметрів моделі, що може привести до первинного виду концептуальної моделі; знайти функціональні залежності характеристик параметрів, що інколи дає можливість створити аналітичні моделі системи чи визначити вагові коефіцієнти критерію ефективності.

Використання результатів моделювання. Результати моделювання використовуються для прийняття рішення про роботоздатність системи, для вибору кращого проектного варіанта чи для оптимізації системи. Рішення про роботоздатність системи приймається з того, виходять чи не виходять характеристики системи за встановлені межі при будь-яких допустимих змінах параметрів. При виборі кращого варіанта з усіх роботоздатних варіантів обирається той, у котрого максимальне значення критерію ефективності. Оптимізація системи є найбільш загальною і складною.

Необхідно знайти таку комбінацію значень змінних параметрів системи чи робочого навантаження з множини допустимих, котре максимізує значення критерію ефективності.

Після створення системи доцільна апостеріорна перевірка результатів моделювання і вимірювання характеристик функціонування. Така перевірка допомагає уточнити модель і підвищити ефективність системи. Наявність моделі діючої системи дає можливість прогнозувати якості функціонування при подальшому розвитку системи чи зміні зовнішніх дій.

1.1.4. Аналітичне моделювання

При аналітичному моделюванні використовуються методи формального представлення аналітичних перетворень певних об'єктів, котрі називаються аналітичними виразами. Вони включають операції диференціювання, інтегрування, зведення подібних, розкриття дужок, підстановки рівностей. Це досить великий клас перетворень. Але практично будь-яке аналітичне перетворення зводиться до підстановки однієї чи кількох рівностей в один чи кілька аналітичних виразів. Зведенню подібних і розкриттю дужок, наприклад, відповідає застосування рівностей, що характеризують дистрибутивність операцій додавання і множення.

Щоб надати цим твердженням точного розуміння, визначається область аналітичних перетворень як деяка формальна теорія, що визначається таким [2]:

1) Нехай $x_1, ..., x_n$ – предметні змінні; $a_1, ..., a_n$ – предметні константи; f_i – функціональні символи; A^j_i – предикатні символи, котрі складають алфавіт теорії.

2) Всяка постійна змінна x_i є термом t_i. Вираз типу $f_i\ (t_1, ..., t_n)$ – терм, якщо $t_1, ..., t_n$ – терми. Ніяких інших термів немає.

3) Вираз $A^n_i(t_1, ..., t_n)$ – елементарна формула. Всяка елементарна формула є формулою. Якщо A і B – формули, то $\neg A, A \supset B \forall x_i A$ (x_i – змінна) теж є формули і інших формул немає. ($\vdash$ – ні; $\supset$ – якщо, то; $\forall$ – для всіх).

Визначивши логічні і власні аксіоми, а також правила виведення, одержують певну теорію першого порядку [2]. Представлення аналітичних перетворень в певній формальній теорії дозволяє успішно розв'язувати задачі теоретичного і практичного характеру і дає можливість створювати ефективні системи автоматизованого програмування. Одержані результати є основою при розробці мовного процесора аналітичних перетворень на базі загальноцільових макрозасобів.

Потоки заявок. Під час аналітичного моделювання характеристики системи обчислюються найбільш просто для потоку заявок, котрий називається найпростішим. Найпростіший потік – це потік заявок, який має такі властивості: стаціонарність; відсутність післядії; ординарність.

Стаціонарність означає постійність ймовірності того, що на протязі певного інтервалу часу надійде однакова кількість заявок незалежно від розташування інтервалу на осі часу. Відсутність післядії полягає в тому, що заявки, котрі надійшли, не впливають на майбутній потік заявок, тобто заявки надходять в систему незалежно одна від одної. Ординарність означає, що в кожний момент часу в систему надходить не більше однієї заявки. Будь-який потік, котрий має такі властивості, є найпростішим.

У найпростішого потоку інтервали часу τ між двома послідовними заявками є незалежними випадковими величинами з експоненціальною функцією розподілу

$$F(\tau) = 1 - e^{-\lambda\tau}. \tag{1.13}$$

Такий розподіл має густину

$$f(\tau) = \lambda e^{-\lambda\tau}, \tag{1.14}$$

математичне очікування довжини інтервалу

$$M(\tau) = \int_0^\infty \tau f(\tau) d\tau = \frac{1}{\lambda}, \tag{1.15}$$

дисперсію

$$D(\tau) = \int_0^\infty (\tau - M(\tau))^2 f(\tau) d\tau = \frac{1}{\lambda^2} \tag{1.16}$$

і середньоквадратичне відхилення, що дорівнює математичному очікуванню.

Експоненціальний розподіл характеризується одним кількісним параметром – інтенсивністю λ. Найпростіші потоки заявок мають такі особливості:

1) сума M незалежних, ординарних, стаціонарних потоків з інтенсивностями λ_i (i = 1, ..., M) сходиться до найпростішого потоку з інтенсивністю $\lambda = \sum_{i=1}^{M} \lambda_i$ за умови, що потоки, котрі додаються, виявляють приблизно однаковий малий вплив на сумарний потік;

2) потік заявок, одержаний в результаті випадкового розрідження вихідного стаціонарного ординарного потоку, котрий має інтенсивність λ, коли кожна заявка виключається з потоку з певною ймовірністю p незалежно від того, виключені інші заявки чи ні, утворює найпростіший потік з інтенсивністю $p\lambda$;

3) інтервал часу між довільним моментом часу і моментом надходження чергової заявки має експоненціальний розподіл з тим же математичним очікуванням $1/\lambda$, що і інтервал часу між двома послідовними заявками.

Найпростіший потік набув великого поширення не тільки за аналітичну

простоту пов'язаної з ним теорії, але й за те, що більшість реально спостережуваних потоків статистично не відрізняється від найпростішого.

Пуассонівський потік. Пуассонівським потоком називається ординарний потік заявок з відсутністю післядії, у котрого число заявок, що поступили в систему за проміжок часу τ, розподілено за законом Пуассона:

$$P(k,\tau) = \frac{(\lambda\tau)^k}{k!} e^{-\lambda\tau}, \quad \lambda > 0, \tag{1.27}$$

де $P(k,\tau)$ – ймовірність того, що за час τ в систему надійде точно k заявок; λ – інтенсивність потоку заявок.

Математичне очікування і дисперсія розподілу Пуассона дорівнюють $\lambda\tau$.

Розподіл Пуассона дискретний. Стаціонарний пуассонівський потік є найпростішим. У розподілі Пуассона тривалості інтервалів між двома послідовними заявками – це випадкові величини з експоненціальним розподілом.

1.1.5. Принципи імітаційного моделювання

Метод імітаційного моделювання полягає у створенні логіко-аналітичної (математичної) моделі системи і зовнішніх дій, в імітації функціонування системи, тобто у визначенні часових змін стану системи під впливом зовнішніх дій і в одержанні вибірок значень вихідних змінних, за котрими визначаються їх імовірнісні характеристики. Це визначення справедливе для стохастичних систем. При дослідженні детермінованих систем відпадає необхідність в одержанні вибірок значень вихідних змінних (функцій).

Взагалі імітаційне моделювання – це метод дослідження, котрий ґрунтується на тому, що динамічна система, яка аналізується, замінюється імітатором, і з ним проводяться експерименти для одержання інформації про систему, що вивчається. Роль імітатора виконує спеціальна програма обчислювальної системи.

Моделювання з використанням комп'ютерів, тобто імітаційне моделювання, є в наш час найбільш ефективним засобом дослідження складних систем. Схема такого дослідження включає такі етапи:

- формалізацію системи з метою побудови її математичної моделі;
- розробку і складання моделюючого алгоритму (алгоритмічної моделі) і програми, котра його реалізує;
- відпрацювання моделюючого алгоритму на комп'ютері;
- обробку і аналіз результатів.

Розділення етапів побудови моделі і проведення імітаційних експериментів обумовлено тим, що машинна модель при формуванні плану

експерименту розглядається як чорний ящик. На етапі побудови моделі визначають її параметри. Результати імітаційних експериментів можуть впливати на вид моделюючого алгоритму лише після їх проведення. Наприклад, якщо в процесі експерименту виясниться, що вихідні результати слабо залежать від того чи іншого параметра, то це може бути причиною спрощення моделі, суть котрого полягає в усуненні даного параметра і відповідному зменшенні розмірності моделі.

Залежно від ступеня формалізації системи, що досліджується, і від способу побудови моделюючого алгоритму розрізняють:

- моделювання з використанням чисельних методів;
- імовірнісне чи стохастичне моделювання із застосуванням спеціальних алгоритмічних мов моделювання.

Моделювання з використанням чисельних методів здійснюється в тих випадках, коли систему вдається описати легкоспостережуваними і достатньо строгими математичними співвідношеннями.

Застосування методу стохастичного моделювання робить моделюючий алгоритм за структурою близьким до алгоритму функціонування досліджуваної системи. Зміна умов моделювання не приводить до істотних змін моделюючого алгоритму, котрий може просто доповнюватись новими блоками чи відпрацьовуватись більшу кількість разів на комп'ютері, наприклад, з метою підвищення точності моделювання.

Загальноприйнята схема стохастичного моделювання наведена на рис. 1.3 і містить три блоки:

- блок імітації випадкових процесів, котрі діють на систему;
- блок програми функціонування системи;
- блок статистичної обробки результатів моделювання.

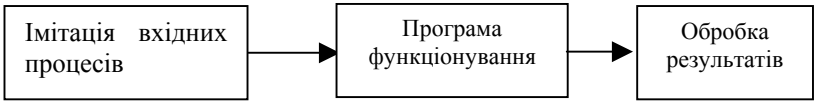

Рис. 1.3. Схема стохастичного моделювання

Статистичні імітаційні моделі є програмним відтворенням структури системи і тих елементарних дій, котрі виконують її окремі елементи.

Основна ідея методу імітаційного моделювання стохастичних систем ґрунтується у багатьох питаннях на методі обчислення випадкових величин, котрий називається методом стохастичних випробувань чи методом Монте-Карло [3]. Він полягає у такому. Нехай необхідно визначити функцію розподілу випадкової величини y. Припустимо, що шукана величина y може бути представлена у вигляді залежності $y = f(\alpha, \beta, ..., \omega)$, де $\alpha, \beta, ..., \omega$ –

випадкові величини з відомими функціями розподілу.

Для розв'язання задачі такого типу використовується наступний алгоритм:

- за кожною з величин α, β, ..., ω здійснюється випадкове випробування, у результаті якого визначається певне конкретне значення випадкової величини α_i, β_i, ..., ω_i;
- використовуючи знайдені величини, визначають одне часткове (певне) значення y_i за наведеною вище залежністю;
- попередні операції повторюються N разів, у результаті чого визначається N значень випадкової величини y_i;
- на основі N значень величини y_i знаходиться її емпірична функція розподілу.

Моделюючий алгоритм, відображаючи елементарні події, котрі відбуваються в системі, слугує для одержання тієї чи іншої інформації про динаміку системи. При моделюванні складних систем, як правило, використовуються імітаційні методи. При цьому домагаються того, щоб моделюючий алгоритм, його структура залежали б не від набору показників роботи системи, а лише від самої математичної моделі. Цього досягають тим, що окремі операції моделюючого алгоритму мають відповідати елементарним явищам, що відбуваються в системі, а послідовність виконання цих операцій повинна відповідати взаємодіям вказаних явищ чи структурі системи. Оскільки моделюючий алгоритм відтворює роботу математичної моделі, то імітаційний підхід вимагає, щоб і математична модель обчислювальної системи структурно і динамічно відповідала реальній системі.

Оператори формування і реалізації випадкових об'єктів (процесів). Вони призначені для імітації різних випадкових факторів (випадкових подій, величин, полів, векторів, функцій), котрі супроводжують процес, що досліджується. Вихідним матеріалом для роботи цих операторів при реалізації на комп'ютері слугують псевдовипадкові числа. Ці числа повинні бути рівномірно розподілені в інтервалі (0,1). Формування чисел можна виділити як окремий оператор, а задачею інших операторів цього ж класу є їх перетворення таким чином, щоб одержати реалізацію заданого закону розподілу з параметрами заданого випадкового об’єкта.

Реалізація модельного часу. Модельний час, котрий відображає час функціонування реальної системи, є одним з основних параметрів при імітаційному моделюванні. Оскільки як моделюючі алгоритми виконуються на комп’ютері, то модельний час є дискретним з кроком квантування Δt.

Важливим є питання просування модельного часу та вибору кроку квантування Δt. Для цього використовують кілька принципів, на основі яких розробляють різні способи апроксимації характеристик стану складних

обчислювальних систем і побудови відповідних моделюючих алгоритмів [3].

Принцип Δt (*принцип приросту часового інтервалу*). Використовується при достатньо малому Δt, для одержання кусочно-задовільної апроксимації характеристик стану системи на кожному кроці дискретизації Δt . Цей принцип є найбільш універсальним і практично придатним для будь-яких систем, але він неекономічний щодо машинного часу. Згідно з принципом Δt, модельний час просувається на деяку величину Δt. Визначаються зміни станів елементів і вихідних дій системи, котрі пройшли за цей час. Після цього модельний час знову просувається на величину Δt, і процедура повторюється до кінця періоду моделювання T_m. Крок Δt в більшості випадків є постійним, але в загальному випадку він може бути і змінним.

Принцип подій. Застосовується при імітаційному моделюванні реальних систем з дискретними вхідними діями та дискретним часом відповідно. При цьому вважається, що стан системи не змінюється між двома сусідніми подіями, а в момент надходження події (особливий стан системи, в якому система знаходиться на протязі дуже короткого інтервалу часу, який при моделюванні вважається нульовим), характеристики системи змінюються стрибком, згідно з алгоритмом реакції системи на вказану подію.

При цьому в поточний момент модельного часу t спочатку аналізуються ті майбутні події – одержання дискретної вхідної дії (заявки), завершення обслуговування і т. ін., для яких були визначені моменти їх настання $t<t_i < T_m$. Обирається найбільш рання подія, і модельний час просувається до моменту настання цієї події. Аналізується реакція системи на поточну подію – зміна станів пов'язаних з цією подією модулів системи та генерація нових типів наступних події, пов'язаних з новими станами. Процедура повторюється до завершення періоду моделювання T_m., або до настання особливого стану системи (неработоздатність, колапс, та ін.).

Принцип послідовного проведення заявок. Є модифікацією принципу подій при моделюванні систем масового обслуговування і полягає в послідовному відтворенні історії кожної із заявок (пріоритету, часу надходження, місця формування) в порядку їх надходження в систему. Алгоритм звертається до інформації про інші заявки тільки в тому випадку, коли це необхідно для вирішення питання щодо долі даної заявки. Алгоритми, побудовані за цим принципом, мають складну логічну структуру, але є найбільш економічними по відношенню до машинного часу.

Нерідко при побудові моделюючих алгоритмів використовують кілька принципів одночасно. Наприклад, загальну структуру моделі будують за принципом особливих станів, а моделюючий алгоритм між особливими станами – за принципом послідовного проведення заявок.

Фіксація і обробка результатів моделювання. При моделюванні складних обчислювальних систем здійснюється багатократна реалізація на

комп’ютері моделюючого алгоритму для одержання статично стійких оцінок шуканих величин. При цьому накопичується значний обсяг статистичної інформації. З метою економії пам'яті комп'ютерної системи необхідно здійснювати фіксацію результатів моделювання та їх статистичну обробку безпосередньо в процесі моделювання з одержанням певних біжучих оцінок для шуканих характеристик.

До якості оцінок, одержаних в результаті статистичної обробки результатів моделювання, ставляться такі вимоги: оцінки повинні бути незміщеними, змістовними (істотними) і ефективними.

Коли серед результатів моделювання присутні випадкові величини, то в якості оцінок для шуканих характеристик розраховують середні значення, дисперсії, кореляційні моменти.

Імітаційне моделювання дає можливість враховувати надійнісні характеристики обчислювальних систем. Зокрема, якщо час напрацювання на відмову і відновлення всіх пристроїв, котрі входять в систему відомий, то визначаються моменти виникнення відмов і моменти відновлення пристроїв протягом періоду моделювання. Якщо в моменти виникнення відмови пристрій зайнятий обслуговуванням заявки, то може прийматися різне рішення залежно від типу пристрою і режиму його роботи: заявка знімається і більше не обслуговується (вибуває з системи) чи заявка ставиться в чергу, а після відновлення пристрою дообслуговується або надходить на повторне обслуговування.

Етапи побудови імітаційної моделі. Розглянемо більш детально розробку моделюючого алгоритму, який реалізує принцип подій. В цьому випадку робота виконується з застосуванням двох основних таблиць – таблиці станів і таблиці подій, які при програмній реалізації можуть бути реалізовані в вигляді структур, об’єктів та ін.

Формування таблиці станів. Система, що моделюється, з заданою точністю деталізації розбивається на блоки (модулі). Всі модулі групуються за типами і нумеруються. В більшості випадків таблиця станів має стільки рядків, скільки модулів в системі (якщо немає динамічно-змінного числа модулів). Число стовпців таблиці станів залежить від типу імітаційної моделі, складності системи, що моделюється та ін., але обов'язковими полями таблиці станів є:

1) унікальний номер модуля;
2) тип модуля, або його ім'я;
3) поточний стан модуля;
4) поля, що показують зв'язок даного модуля з іншими.

Ефективна побудова полів 4-го типу є деякою мірою мистецтвом. Крім того в деяких складних багаторівневих системах дані поля доповнюються різного виду логічними та алгоритмічними перевірками. Головна вимога до

таблиці станів така: вона повинна повністю задавати взаємозв'язок між модулями системи, що моделюється.

Для кожного типу модулів, залежно від призначення імітаційної моделі, задаються можливі типи станів, які кодуються (символами, цифрами, або якось по-іншому). Визначаються початкові стани всіх модулів і заносяться в поле 3 "поточний стан модуля".

Таблиця станів сформована. У подальшому при імітаційному моделюванні поточні стани модулів динамічно змінюються, відображаючи таким чином зміни в реальній системі.

Формування таблиці подій. Подія – ключове поняття імітаційного моделювання – є причиною зміни станів. Якщо всі можливі стани зобразити у вигляді вершин орієнтованого графа, то всі можливі події – це дуги вказаного графа. Таким чином визначається множина можливих подій на всіх можливих станах всіх типів модулів, які аналогічним чином кодуються.

Таблиця подій (майбутніх подій, які повинні виникати в системі) є динамічною таблицею, яку при програмуванні доцільно організувати у вигляді динамічних структур. Кожен запис у таблиці подій (незалежно від типу імітаційної моделі) характеризується трьома атрибутами і складається з трьох полів відповідно:

1) номер модуля, на якому виникає подія;

2) тип (код) події;

3) модельний час виникнення події.

В нульовий момент модельного часу формується початкова таблиця майбутніх подій, при цьому в циклі перебираються всі модулі (проглядається таблиця станів) і залежно від типу модуля та його поточного стану генеруються всі дозволені типи подій (по одній події на кожний тип для кожного модуля). Таким чином поля 1 та 2 задаються явно. А звідки взяти в програмі поле типу 3 – час виникнення події (в реальній системі події виникають самі по собі)?

Для цього використовують статистичні властивості подій (хоча є детерміновані імітаційні моделі, в яких час майбутньої події однозначно визначається залежно від поточного стану модуля).

Події на різних типах модулів підкоряються різним законам розподілу (допускається, що ці закони, а також їх характеристики відомі розробникам імітаційної моделі), тому час наступної події розраховується, виходячи з відомого закону з використанням генератора псевдовипадкових чисел.

Так, для імітаційної моделі надійності, де всі події підкоряються експотенціальному закону, час події розраховується за формулою

$$T = -\frac{1}{\lambda} ln(1 - N), \qquad (1.28)$$

де λ – інтенсивність події даного типу; N – псевдовипадкове число, рівномірно розподілене в інтервалі [0,1].

Псевдовипадкове число N одержується програмою RANDOM. Після заповнення таблиці подій, вона ранжується за зростанням по полю 3 – час виникнення події. Тепер починається сам етап імітаційного моделювання.

Алгоритм імітаційної моделі. При імітаційному моделюванні з таблиці подій вибирається найближча за часом подія (при цьому модельний час дискретно змінюється і прирівнюється до часу настання події), і розраховується реакція моделі на дану подію (яка в кінці приводить до зміни поточних станів модулів). Реакцією системи може бути:

1) зміна стану того модуля, на якому виникла подія (номер модуля є в таблиці подій);

2) зміна станів інших модулів, зв'язаних з даним модулем (обчислюється при аналізі полів таблиці станів);

3) зміна стану всієї системи (обчислюється при перевірці різних критеріїв);

4) генерація нових подій (з врахуванням поточного модельного часу), зумовлених новими станами модулів.

Таким чином, при імітаційному моделюванні першою дією розрахунку реакції є аналіз типу події і реалізація процедур розрахунку реакції для кожного типу події.

Слід відмітити, що після генерації нових подій таблиця подій по-новому ранжирується (можна за простішим алгоритмом).

Імітаційне моделювання виконується до лімітованого часу T_m, або до досягнення конкретного стану всієї системи. Під час імітаційного моделювання формується протокол подій (по кожному модулю, по групі, або по всій системі), на основі якого розраховуються статистичні характеристики системи.

Оскільки вхідними даними для імітаційної моделі є випадкові числа – формула (1.28), то імітаційна модель дає випадкові точечні оцінки, але при достатньо великому часі моделювання (для ергодичних процесів), або при багатократному прорахунку моделі вона дає стійкі статистичні оцінки.

Контрольні питання

1. Задачі, які розв'язуються за допомогою моделювання складних об'єктів.
2. Основні поняття моделювання. Формалізація.
3. Що таке система? Основні визначення та формалізація.
4. Теорія подібності. Основні визначення.
5. Класифікація моделей і області їхнього застосування.
6. Що таке фізична модель і які типи фізичних моделей вам відомі?

7. Що таке абстрактна модель і які типи абстрактних моделей вам відомі?

8. Яке основне призначення концептуальної моделі?

9. Що таке математична модель і які типи математичних моделей вам відомі?

10. Принципи імітаційного моделювання.

11. Етапи розробки моделей.

12. Які задачі розв'язуються на етапі формулювання мети моделювання?

13. Які задачі розв'язуються на етапі вибору засобів моделювання?

14. Які задачі розв'язуються на етапі розробки концептуальної моделі?

15. Які задачі розв'язуються на етапі підготовки вихідних даних?

16. Які задачі розв'язуються на етапі розробки математичної моделі?

17. Які задачі розв'язуються на етапі вибору методу моделювання?

18. Які задачі розв'язуються на етапі розробки програмної моделі?

19. Які задачі розв'язуються на етапі перевірки адекватності та корегування моделі?

20. Які задачі розв'язуються на етапі планування машинних експериментів?

21. Які задачі розв'язуються на етапі комп’ютерного моделювання?

22. Які задачі розв'язуються на етапі аналізу результатів моделювання?

23. Сутність аналітичного моделювання.

24. Основні етапи імітаційного моделювання.

25. Схема стохастичного моделювання при побудові імітаційної моделі.

26. Основні положення методу стохастичних випробувань (методу Монте-Карло).

27. Реалізація модельного часу та принципи побудови моделюючих алгоритмів.

28. Структура таблиці станів при імітаційному моделюванні.

29. Формування таблиці подій при імітаційному моделюванні.

30. Алгоритм імітаційної моделі. Взаємодія таблиць станів та подій.

31. Що таке реакція системи на подію і як вона розраховується?

1.2. Аналіз та попередня обробка множини вхідних та вихідних даних при розробці моделей складних об’єктів

Показники якості моделей складних об'єктів залежать у першу чергу від одержання точних даних і інформації про стан об'єктів, що досліджуються, а також про взаємодію об'єктів з зовнішнім технічним і природним середовищем.

1.2.1. Типи вхідних даних – бінарні, рангові, чисельні та методи їхньої обробки

В задачах класифікації та діагностики вхідні дані одержали назву вхідних ознак. За метрологічною оцінкою вхідні ознаки поділяються на такі типи:

1) Кількісні або числові ознаки. Це заміряні у визначеній шкалі і виражаються числами з визначеною точністю виміру (результати інструментальних досліджень і ознаки, отримані в результаті обробки сигналів та зображень).

2) Якісні, рангові або бальні. Використовуються для вираження експертних оцінок, термінів і понять, не мають цифрових значень (наприклад, тяжкість патології) і заміряються у шкалі порядку.

3) Бінарні або дихотомічні. Набувають тільки двох значень ("0" або "1", "ТАК" або "НІ") і використовуються для фіксації у формалізованих документах наявності або відсутності якоїсь ознаки.

4) Класифікаційні або номінальні (наприклад, стать, професія, група крові). Це ознаки, заміряні в шкалі найменувань.

Для кожного з розглянутих типів ознак застосовуються свої методи дослідження, хоча багато алгоритмів, які розроблені для одного типу, адаптуються до інших типів [4]. При цьому змістовна частина інформації містить такі її складові:

- систематичну інформацію, котра викликана досліджуваними впливами, що і є зазвичай предметом дослідження;
- систематичну інформацію, пов'язану з умовами дослідження (методами дослідження), тобто постійними похибками (помилками);
- випадкову (залишкову) інформацію, викликану нерегулярними змінами у процесі дослідження.

Таким чином, при побудові моделей складних об'єктів (особливо в медичній діагностиці) використовується різнорідна, отримана різним дослідниками (і в різний час), недостатньо формалізована і така, що несе елементи суб'єктивної оцінки експерта, інформація. Тому обов'язковими етапами первинної обробки інформації є етапи формалізації опису і формування переліку вхідних ознак для даної задачі дослідження.

При формалізації опису кожному значенню ознаки ставиться у відповідність визначене кодове число (оцифровка шкал). При оцифровці шкал виконується зведення всіх типів ознак до однієї кількісної шкали.

Кодування не повинне змінювати семантичну силу і зміст вихідних даних, тобто предметний зміст інформації.

Кількісні ознаки кодуються у вихідному виді (значення) або в квантованому виді (належність кількісної ознаки до діагностично-значимих інтервалів). При неправильному визначенні кількості і границь інтервалів

можна утратити важливу для розпізнавання властивість даної ознаки.

У найпростішому випадку дихотомічної шкали, тобто коли ознака може набувати значень «так» або «ні», немає великої різниці, які числа будуть приписані позитивній або негативній відповіді. Найпоширеніші варіанти такі: відповіді «так» приписують число 1, відповіді «ні» – число –1, або 0.

Рангові ознаки підрозділяються на градації відповідно до зміни їхньої виразності або в зростаючому, або в спадному ряді значень. При цьому застосовується як 4-градаційна шкала, що відповідає прийнятому експертами ступеню виразності (відсутність ознаки або норма, слабкий ступінь, середній ступінь і сильний ступінь прояву ознак), так і 7-градаційна шкала, що дозволяє виявити дискретність прояву якісних ознак і яка відповідає психофізичним можливостям людини по переробці інформації (закон "сім плюс мінус два").

У випадку порядкових шкал, як правило, порядок проходження градацій ознаки відображає ступінь посилення або ослаблення тієї або іншої якості. Числові мітки ознаки в цьому випадку привласнюються таким чином, щоб відстані між двома оцінками інтуїтивно відповідали різниці між відповідними градаціями (наприклад, якщо ознака має шкалу «добре», то логічно приписати градаціям мітки –1; 0; 1, а от у випадку шкали «малий–середній–великий–дуже великий» більш доречним може виявитися використання логарифмічних міток, тобто 0.1; 1; 10; 100).

Номінальним ознакам кодові значення можуть бути привласнені довільно, відповідно до прийнятого порядку перерахування показників.

Наступним кроком у статистичній обробці даних, як правило, є знаходження точки середнього значення всіх ознак – геометричного центра багатовимірної хмари точок даних. Зручно зрушити всі точки даних на однаковий вектор таким чином, щоб центр хмари виявився на початку координат. Таке перетворення називається центруванням даних.

Далі виконується нормування даних – тобто ділення усіх значень ознак на визначене число таким чином, щоб значення ознак попадали в порівнянні за величиною інтервали. В якості такого числа вибирається один із характерних масштабів.

У багатовимірній хмарі даних існує кілька масштабів. Перший – це середньоквадратичне відхилення

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2},$$

де $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ (X_i – вектор даних).

У випадку, якщо вибірка може вважатися отриманою із нормального розподілу, то в колі з центром в $\bar{x}$ та радіусом σ знаходиться близько двох третин від числа точок даних. Існує масштаб, який характеризує максимальне розсіювання в множині даних

$$R = \max_{i=1,N} \left\| X_i - \bar{X} \right\| .$$

Нормування всіх ознак на R призводить до того, що вся множина даних поміщається в коло одиничного радіусу.

Якщо масштаб вибрано σ або R, то відповідні формули обробки (нормування на «одиничну дисперсію» і «на одиничне коло») мають вигляд:

$$\tilde{X}_i = \frac{X_i - \bar{X}}{\sigma} , \; \tilde{X}_i = \frac{X_i - \bar{X}}{R} ,$$

де $\tilde{X}_i$ – новий вектор ознак; X_i – старий вектор ознак.

Крім того, якщо діапазони значень для різних ознак сильно відрізняються один від одного, то краще для кожної з ознак застосувати власний масштаб. Тобто, для кожної з ознак можна ввести своє середньоквадратичне відхилення та розсіювання:

$$\sigma_j = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2} , \; R_j = \max_{i=1,N} \left\| x_{ij} - \bar{x}_j \right\| ,$$

де $\bar{x}_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij}$ (x_{ij} – значення i -ї ознаки на j -му векторі), .

Як результат отримаємо формули для нормування на «одиничну дисперсію для кожної ознаки» і «на одиничний куб»:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} , \; \tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{R_j} .$$

Наступним важливим етапом обробки даних є аналіз різких відхилень у вихідних даних і відновлення пропущеної інформації. Статистичні процедури виділення відзначених "сумнівних" даних засновані на припущенні про однорідність даних, у той час як “дані, що вискакують”, розглядаються як атипічно далеко віддалені від центра розподілу. Сумнівні спостереження або цілком виключаються з подальшого розгляду, або їхній внесок зменшується за допомогою вагової функції, що убуває в міру зростання ступеня аномальності спостережень. Якщо є пропуски інформації (типова ситуація для медичних БД), то застосовуються різні підходи – "обнуління", умовне

кодування, усереднення, відновлення [4].

Слід зазначити, що при синтезі вирішальних правил доводиться використовувати ознаки, вірогідність яких викликає сумнів у розробників моделей та відповідних систем підтримки прийняття рішень (СППР) на вказаних моделях. У цьому випадку вводиться система вагових коефіцієнтів, що відображає кількісну оцінку їхньої значимості.

Важливим етапом попереднього аналізу даних є оцінка особливостей розподілу, тому що значне число статистичних методів припускає нормальний характер розподілу ймовірностей. Однак саме біологічні і медичні дані часто характеризуються значним відхиленням від нормального закону. При відхиленні закону розподілу від нормального використовуються непараметричні критерії (метод квантильних шкал, "бутстреп" і ін.).

Оцінка інформативності і формування інформативного простору ознак. Серед комплексу проблем, розв'язуваних при розробці СППР у роботі [7] виділено дві актуальні задачі оптимізації: добір інформативних ознак і оптимізація вирішальних правил. При цьому відзначається таке: "Якщо оптимізація вирішальних правил у якомусь ступені знаходить задовільне рішення, що обґрунтовується наявністю добре розробленої теорії перевірки статистичних гіпотез, то проблема власне аналізу різних методик обстеження освітлена слабо". Розроблювальна система вихідних діагностичних ознак повинна задовольняти таким вимогам:

1) Повнота опису. Система вихідних ознак має охоплювати усі виділені аспекти вимірюваного поняття.

2) Ощадливість опису. Найбільш розповсюдженою помилкою багатьох дослідників є "спроба аналізу украй великого числа ознак, що, на думку дослідників, повинне сприяти підвищенню інформативності наведенної вибірки". Однак для розв’язання будь-якої класифікаційної задачі необхідно використовувати корисну для даної задачі інформацію, що несе не "шум" і неіррелевантну інформацію (не відноситься до мети дослідження), тому при розробці системи ознак варто уникати зайвого обсягу вихідної інформації.

3) Структурованість системи ознак. Ознаки повинні групуватися, відносно рівномірно описуючи всі сторони вимірюваного явища.

4) Кількісна визначеність діагностичних ознак. Ця визначеність забезпечується формалізацією опису ознак, що розглянута вище.

Наведені вимоги не є вичерпними. У випадку одержання даних за допомогою тестів-опитувальників велика увага має приділятися прийомам зниження можливості фальсифікації відповідей і зменшення систематичної помилки тестування. У деяких випадках добір корисної інформації зі змістовних розумінь виконується дослідником самостійно (можливо не зовсім удало), у противному разі боротьба з надмірністю здійснюється формальними методами шляхом оцінки інформативності простору

діагностичних ознак.

1.2.2. Методи аналізу структури даних

За результатами вимірювання характеристик об'єктів в заданій предметній області і формалізації опису формується двовимірна таблиця експериментальних даних (ТЕД), структура якої показана в табл. 1.1.

Таблиця 1.1 – Структура таблиці експериментальних даних

Об'єкти навч. вибірки	Вихідні ознаки					
	x_1	x_2	...	x_j	...	x_p
A_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
A_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
...	...	...	...	...	...	...
A_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
...	...	...	...	...	...	...
A_N	x_{N1}	x_{N2}	...	x_{Nj}	...	x_{Np}

У цій таблиці прийняті такі позначення:

N – загальна кількість об'єктів навчальної вибірки;

p – загальна кількість ознак;

x_j – j-а ознака;

x_{ij} – значення j-ї ознаки, виміряне в i-го об'єкта;

$X = (x_1, ..., x_p)'$ – вектор ознак;

$A = \{A_1, ..., A_N\}$ – множина об'єктів.

ТЕД служить для синтезу структури діагностичної моделі й для оцінки її параметрів. У зазначеній моделі повинен у визначеній формі виражатися зв'язок вимірюваного вектора ознак X з тестуємою властивістю Y. При цьому необхідно забезпечити економічність за формою і змістовність за змістом перетворення $Y = f(X)$ при дотриманні заданої точності моделі для відображення моделлю загальних закономірностей структури експериментальних даних з метою підвищення стійкості і надійності кількісної оцінки діагностуємих показників. Для забезпечення зазначених вимог до діагностичної моделі виконується оптимізація простору ознак X.

Структура експериментальних даних відображається за допомогою двох основних категорій взаємодії між елементами ТЕД – категорій подібності і розходження. Подібність і розходження ознак визначається мірами зв'язку, а об'єктів ТЕД – мірами близькості (відстаней). Матриця зв'язку задає відношення "ознака – ознака" і являє собою двовимірну симетричну квадратну матрицю розміром $p \times p$.

$$\boldsymbol{R} = \begin{vmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{vmatrix}, \qquad (1.29)$$

де r_{ij} – міра зв'язку між ознаками x_i і x_j.

У роботі [8] виділені дві представницькі групи зв'язку між ознаками, засновані на принципі коваріації і на принципі спряженості ознак.

Виходячи з першого принципу, висновок про наявність зв'язку між змінними робиться в тому випадку, коли збільшення значення однієї змінної супроводжується стійким збільшенням або зменшенням значень іншої. У математичному виразі задача зводиться до обчислення коваріації, тобто до супутньої зміни чисельних значень ознак.

На принципі коваріації заснований коефіцієнт парної кореляції Пірсона (r_{kj}^{p}), що є мірою лінійного зв'язку двох змінних: x_k і x_j. Він обчислюється за формулою:

$$r_{kj}^{p} = s_{kj} \Big/ \sqrt{s_{kk}\, s_{jj}}\,, \qquad (1.30)$$

де $s_{kj} = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - m_k)(x_{ij} - m_j)$, $m_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$ є елементами коваріаційної матриці генеральної сукупності, з якої витягнуті об'єкти x_i і x_j .

Критерієм значимості коефіцієнта кореляції служить критерій Стьюдента t для нормального розподілу значень випадкових величин.

Перевірка гіпотези про відсутність статистичного зв'язку H_0 між x_k та x_j виконується шляхом перевірки нерівностей

H_0: $r = 0$ при $\lvert t_0 \rvert < t_{1-\alpha/2\,(n\text{-}2)}$, H_1 : $r \neq 0$ при $\lvert t_0 \rvert > t_{1-\alpha/2\,(n\text{-}2)}$.	(1.31)

В нерівностях (1.31) розрахунковий критерій Стьюдента t_0 обчислюється за формулою

$$t_0 = \frac{\tilde{r}\sqrt{n-2}}{\sqrt{1-\tilde{r}^2}}$$

та порівнюється з табличним значенням критерію Стьюдента $t_{1-\alpha/2\,(n\text{-}2)}$ з рівнем значимості α (зазвичай 0,05 або 0,01) при n–2 ступенях свободи (n – обсяг навчальної вибірки).

Для інших типів ознак застосовуються інші міри зв'язку, які по суті, є

алгебраїчним перетворенням коефіцієнта кореляції Пірсона r_{kj}^{p}, і враховують тип ознак, що зіставляються. Для аналізу рангових ознак застосовується коефіцієнт рангової кореляції Спірмена, що визначається за формулою:

$$r_{kj}^{s} = 1 - 6 \frac{\sum_{i=1}^{N}(r_{ji} - r_{ki})}{N(N^2 - 1)},$$

де r_{ji} і r_{ki} – ранги ознак x_j і x_k відповідно.

Крім того, для рангових змінних використовується міра зв'язку, заснована на підрахунку числа розбіжностей у ранжируванні об'єктів ("тау" Кендалла):

$$\tau = \frac{P - Q}{N(N-1)/2},$$

де P – число збігів, а Q – число розбіжностей порядків з $N(N–1)/2$ пар ознак.

Незважаючи на розходження в підходах, між коефіцієнтами рангової кореляції Спірмена і Кендалла існує тісний логічний зв'язок, і в більшості випадків вони дають однакові результати.

Друга велика група мір зв'язку, заснована на принципі взаємної спряженості, спрямована на з'ясування наступного факту: чи з'являються деякі значення однієї ознаки одночасно з визначеними значеннями іншої частіше, ніж при випадковій вибірці?

Загальним для першої і другої груп є коефіцієнт спряженості φ, що призначений для виміру зв'язку двох дихотомічних ознак: $x_j = \{0,1\}$ і $x_i = \{0,1\}$. Для обчислення коефіцієнта φ будується чотириелементна таблиця спряженості, структура якої показана в табл. 1.2.

Таблиця 1.2 – Таблиця спряженості дихотомічних ознак

$x_i \setminus x_j$	0	1	S
0	a	b	$a + b$
1	c	d	$c + d$
S	$a + c$	$b + d$	

Елементами таблиці спряженості (a, b, c, d) є кількості об'єктів з відповідними комбінаціями значень дихотомічних ознак x_i і x_j (a – 0, 0; b – 0, 1; c – 1, 0; d – 1, 1). При цьому $a + b + c + d = N$. Тут же підраховуються відповідні суми (по рядках і по стовпцях).

Коефіцієнт φ являє собою алгебраїчне спрощення коефіцієнта кореляції Пірсона r_{kj}^{p}, з урахуванням специфіки дихотомічних ознак і обчислюється за

формулою:

$$\varphi = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \qquad (1.32)$$

Коефіцієнти спряженості будуються не тільки для дихотомічних ознак по чотириелементній таблиці, але й у більш складних випадках. У загальному випадку будується $(m \times m)$-елементна таблиця спряженості при розбивці діапазону зміни ознак на m частин і при порівнянні їхніх значень з $m - 1$ порогами. У такий спосіб коефіцієнти спряженості можна застосовувати для аналізу зв'язків між чисельними, ранговими і дихотомічними ознаками і їх комбінаціями.

При виборі тієї або іншої міри зв'язку для розв'язання конкретної задачі в роботі [7] відзначається, що застосування до тих самих даних різних мір зв'язку нерідко приводить до результатів, що відрізняються. Це обумовлено тим, що математики, які конструювали коефіцієнти зв'язків, як правило, досліджували їхні властивості в граничних ситуаціях – близько 0 або 1. Поводження ж різних мір зв'язку усередині інтервалу [0, 1] порівняно мало вивчене. Тому на практиці вибір якої-небудь міри зв'язку визначається особистими симпатіями дослідника.

Розглянуті вище міри зв'язків між ознаками вказують лише на наявність/відсутність лінійного зв'язку між двома окремими ознаками. При цьому слід зазначити, що коефіцієнт кореляції (і його алгебраїчні спрощення для інших типів ознак) має чіткий сенс як характеристика ступеня тісноти зв'язку тільки у випадку спільного нормального розподілу досліджуваних ознак. Наявність кореляційного зв'язку між ознаками не є однозначною вказівкою на їхню взаємозумовленість (можливі випадки "помилкової" кореляції). З іншого боку, некорельованість не може служити однозначною вказівкою на відсутність зв'язку між ознаками, вона лише вказує на відсутність лінійної залежності між ними.

Більшість методів синтезу діагностичних моделей припускають роботу з незалежною системою ознак (чого практично не буває при прийнятих методиках вимірювань), або в крайньому випадку дають прийнятні результати при слабозв'язаній системі ознак. Тому аналіз зв'язків між ознаками є необхідним етапом обробки експериментальних даних з метою ухвалення рішення про перетворення простору ознак.

Матриця близькостей (віддаленостей) задає відношення "об'єкт – об'єкт" і являє собою квадратну симетричну матрицю $N \times N$ з невід'ємними елементами (1.33). Елементи d_{ij} є значеннями деякої міри близькості (віддаленості або відстані) між об'єктами x_i і x_j у заданому просторі ознак (вихідному або перетвореному). Найчастіше при аналізі даних використовуються міри віддаленості

$$\boldsymbol{D} = \begin{vmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{vmatrix}. \qquad (1.33)$$

До цих мір пред'являються наступні вимоги [4, 5]: максимальна подібність об'єкта із самим собою $d_{ij} = \min d_{ij}$; вимога симетрії $d_{ij} = d_{ji}$; виконання нерівності трикутника $d_{ij} \le d_{ik} + d_{kj}$.

В задачі класифікації об'єктів матриця близькості, або відстаней $\boldsymbol{D}$, між об'єктами будується у просторі ознак, де для всіх об'єктів навчальної вибірки визначена належність до одного із заданої множини класів. Матриця близькості показує, наскільки в даному просторі ознак один клас віддаляється від іншого. Для синтезу діагностичних вирішальних правил (класифікації об'єктів) необхідне виконання гіпотези компактності: об'єкти, що належать одному класові, повинні розташовуватися в просторі ознак компактними групами.

Найбільш розповсюдженими мірами (відстанями) між об'єктами A_i і A_j є такі [8]: звичайна d_{ij}^{E} і зважена $d_{ij}^{\%E}$ евклідова відстань; відстань Махаланобіса d_{ij}^{M}; узагальнена відстань Мінковського d_{ij}^{MI}; відстань Хеммінга d_{ij}^{H}. Вказані міри визначаються таким чином:

$$d_{ij}^{E} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \quad d_{ij}^{\%E} = \sqrt{\sum_{k=1}^{p} w_k (x_{ik} - x_{jk})^2},$$

$$d_{ij}^{M} = \sqrt{(\bar{x}_i - \bar{x}_j)^T S^{-1} (\bar{x}_i - \bar{x}_j)}, \; d_{ij}^{MI} = \sqrt[q]{\sum_{k=1}^{p} (x_{ik} - x_{jk})^q}, \; d_{ij}^{H} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|,$$

де x_{ik}, x_{jk} – координати (ознаки) i і j об'єктів у просторі ознак X відповідно; w_k – ваги ознак; p – розмірність простору ознак X; $\bar{x}_i$, $\bar{x}_j$ – вектори ознак i та j об'єктів у просторі ознак X відповідно; S – коваріаційна матриця генеральної сукупності, з якої витягнуті об'єкти x_i і x_j.

Евклідова відстань застосовується для обчислення відстані між об'єктами, описаними кількісними, якісними і дихотомічними ознаками. Її використання доцільне, коли ознаки однорідні за семантичним навантаженням й однаково важливі для розв'язуваної задачі.

Зважена евклідова відстань використовується, коли необхідно кількісно

виразити важливість яких-небудь ознак або коли треба вирівняти масштаби неоднорідних ознак.

Відстань Махаланобіса застосовується при сильній залежності і неоднорідності досліджуваних ознак, тому що вона інваріантна до лінійних перетворень простору ознак (до зміни масштабу і повороту осей).

Узагальнена відстань Мінковского є універсальною мірою близькості. При $q = 1$ вона переходить у Хеммінгову відстань для дихотомічних ознак або в "міську метрику" для ординальных ознак; при $q = 2$ – у евклідову; при $q \to \infty$ – у так називану супремум-норму $d_{ij}^{(q)} = \max_k \left| x_{ik} - x_{jk} \right|$. Але в даний час "…не відомі приклади її використання при довільних $q \neq 1, 2, \infty$ …" [7].

Відстань Хеммінга найчастіше використовується для визначення розходжень між об'єктами, що задаються дихотомічними ознаками, й інтерпретується як число розбіжностей значень ознак x_i і x_j у розглянутих об'єктів A_i і A_j.

1.2.3. Методи зниження розмірності простору ознак

Пошук найбільш інформативної системи діагностичних ознак виконується безліччю формальних методів, об'єднаних загальним поняттям "аналіз даних". При цьому відбувається агрегування (стиск) вихідного простору ознак з метою зведення його до компактного і доступного для огляду вигляду при подальшому дослідженні. Існує ряд підходів до розв'язання даної задачі, серед яких можна умовно виділити дві групи методів:

– зниження розмірності простору ознак шляхом заміни значної кількості вихідних ознак невеликим числом інтегральних (узагальнених) ознак, що зберігають достатню інформацію про досліджувані об'єкти. Дана група методів заснована на дослідженні автоінформативності вихідного простору ознак. Інтегральні ознаки є лінійною комбінацією вихідних ознак. До зазначеної групи методів відносяться метод головних компонент, факторний аналіз, багатовимірне шкалювання та ін.;

– зниження розмірності простору ознак шляхом оцінки відносної значимості окремих ознак при розв'язанні класифікаційної задачі і виділення підпростору значимих ознак. У даній групі методів використовується "зовнішній" критерій інформативності – вплив окремої ознаки (або групи ознак) на властивість, що діагностується. Обчислення "зовнішнього" критерію може бути виконане методами дисперсійного аналізу, кореляційного аналізу, на основі теоретико-інформаційного підходу. При використанні будь-якого "зовнішнього" критерію можна застосовувати різні алгоритми зниження розмірності (повний перебір; *k* кращих; методи послідовного зменшення і збільшення простору ознак та ін.).

Перша група методів застосовується за наявності значного зв'язку (корельованності) між окремими ознаками вихідного простору, а друга група методів припускає роботу з незалежною або слабозв'язаною системою ознак.

Важливою задачею "аналізу даних" є кластеризація ознак, тобто формування з вихідної множини ознак m підмножин (кластерів). Кластеризація ознак знижує розмірність задачі і дозволяє застосовувати методи зниження розмірності усередині кожного кластера. Для кластеризації ознак застосовують методи кластерного аналізу, кореляційних плеяд та ін.

Розглянемо більш докладно відзначені вище методи й особливості їхнього застосування до експериментальних даних.

Кластерний аналіз [9] заснований на обчисленні відстаней між об'єктами в просторі ознак і формуванні m кластерів (значення m задане або визначається алгоритмами), усередині яких об'єкти мають максимальну подібність (мінімальна відстань), у той час як між кластерами спостерігається різнорідність (максимальна відстань). Відомі такі методи кластерного аналізу: одиночних і повних зв'язків, попарне середнє, центроїдний, Варда, k-середнього [9, 10]. Відмінність названих методів полягає в способі обчислення відстані між кластерами, а також – у виді цільової функції.

Нижче перераховані найбільш відомі методи кластерного аналізу.

Метод повних зв'язків (найбільш віддалених сусідів). Суть даного методу полягає в тім, що два об'єкти, які належать тому самому кластеру, мають коефіцієнт подібності, який менший від деякого граничного значення. У термінах евклідової відстані $D_{ij}^{(E)}$ це означає, що відстань між двома точками кластера не повинна перевищувати деяке граничне значення h. Таким чином, значення h визначає максимально припустимий діаметр підмножини, що утворить кластер. При цьому процедуру кластеризації можна розглядати як одержання графа, у якому ребра з'єднують усі вершини в групу, тобто кожна група утворить повний підграф. При цьому відстань між групами K_s і K_t визначається найбільш віддаленими вершинами в цих групах:

$$R_{st} = \max_{i \in K_s,\, j \in K_t} D_{ij} .$$

Метод одиночних зв'язків (метод найближчого сусіда). Кожний об'єкт розглядається як одноточковий кластер. Об'єкти групуються за таким правилом: два кластери поєднуються, якщо відстань між самими ближніми точками одного кластера K_s і точками іншого K_t мінімальна:

$$R_{st} = \min_{i \in K_s,\, j \in K_t} D_{ij} .$$

Таким чином, найближчі сусіди визначають найближчі підмножини. Процедура складається з (n-1) кроків і, виходячи з позиції теорії графів, коли

об'єкти розглядаються як вершини графа, полягає в знаходженні мінімального покриваючого дерева.

Метод Ворда. У цьому методі як цільову функцію застосовують внутрішньогрупову суму квадратів відхилень, яка є не чим іншим, як сумою квадратів відстаней між кожною точкою і центром кластера, який містить цю точку. На кожному кроці поєднуються такі два кластери, які приводять до мінімального збільшення цільової функції, тобто внутрішньогрупової суми квадратів. Цей метод спрямований на об'єднання близько розташованих кластерів.

Центроїдний метод (*k*-середніх). Відстань між двома кластерами K_s і K_t визначається як евклідова відстань між центрами (зваженими середніми за кожним показником) цих кластерів:

$$R_{st} = D^{(E)}(\overline{x_{K_s}}, \overline{x_{K_t}}),$$

де $\overline{x_{K_s}}$ – середнє арифметичне векторних спостережень x_i, при $i \in K_s$.

Кластеризація йде поетапно. На кожному з (n–1) кроків поєднують два кластери (K_s і K_t), що мають мінімальне значення R_{st}. При цьому на нульовому кроці за центри шуканих k кластерів приймають випадково обрані k спостережень — точки $x_1, x_2, ..., x_k$. Якщо при обчисленнях використовуються ваги для врахування різниці між розмірами кластерів (тобто кількістю точок у них), то цей метод називають ще методом зважених груп.

Метод попарного середнього. У цьому методі відстань між двома різними кластерами K_s і K_t, обсягом p_s і p_t об'єктів відповідно, обчислюється як середня відстань між усіма парами об'єктів у них:

$$R_{st} = \frac{1}{p_s p_t} \sum_{\substack{i \in K_s \\ j \in K_t}} D_{ij}.$$

У випадку нерівних розмірів кластерів враховується розмір відповідних кластерів як ваговий коефіцієнт, і тоді даний метод називають зваженим попарним середнім.

Усі розглянуті вище методи є методами, спрямованими на послідовне об'єднання одиночних груп.

Ієрархічне групування. Суть даного методу полягає у послідовному об'єднанні (алгомеративні процедури) або поділі (дивізимні процедури) кластерів. При послідовному об'єднанні на першому кроці кожен об'єкт вважається окремим кластером, і на кожному наступному кроці два найближчих кластери поєднуються в один доти, поки всі об'єкти не об'єднаються в один кластер. При послідовній розбивці, навпаки, всі об'єкти на першому кроці належать одному кластерові, а на кожному наступному

кроці з кластера відокремлюються об'єкти, найбільш віддалені від центра кластера. В обох процедурах як міра близькості між об'єктами використовується відстань у просторі ознак. При цьому оптимальну розбивку вибирає сам дослідник шляхом аналізу так званої дендограми, побудованої за результатами групування на всіх кроках алгоритму з урахуванням шкали подібності.

Зазначені методи кластерного аналізу призначені для кластеризації об'єктів, використовують координати об'єктів у просторі ознак і їхнє безпосереднє застосування до задачі кластеризації ознак неможливе (між ознаками можна визначити відстань, але немає координат), однак при використанні деяких прийомів (обчислення фіктивних координат ознак; транспонування ТЕД, у результаті чого об'єкти стають фіктивними координатами, а ознаки – фіктивними об'єктами відповідно) методи кластерного аналізу використовуються для кластеризації ознак.

Метод кореляційних плеяд. У даному методі [11] структуру зв'язків між діагностичними ознаками зображають у вигляді графа, де кожна вершина відповідає одній з ознак, а ребра вказують на кореляційний зв'язок між ознаками. Метод призначений для синтезу таких груп ознак – "плеяд", для яких зв'язок усередині плеяд досить великий, а між ними – малий. Задаючись деякими граничними значеннями коефіцієнта кореляції, виключають з вихідного графа всі ребра з коефіцієнтом кореляції, який по модулю менший від граничного. Поетапно збільшуючи граничне значення коефіцієнта кореляції, повторюють цю процедуру до розбивки графа на декілька підграфів. Вершини кожного підграфа й утворять плеяду. Для отриманих у такий спосіб плеяд, коефіцієнти кореляції всередині плеяд будуть більші від останнього граничного значення, а між ними – менші. Однак даний метод аналізує тільки окремі зв'язки між вершинами, а не їхню сукупність, що не завжди приводить до оптимальної розбивки, тому він використовується в основному для наочності зображення структури зв'язків між ознаками.

Метод головних компонентів здійснює перехід до нової системи координат y_1, ..., y_p у вихідному просторі ознак x_1, ..., x_p яка є системою ортонормованих лінійних комбінацій [4, 8].

$$\begin{cases} y_j(x) = w_{1j}(x_1 - m_1) + \cdots + w_{pj}(x_p - m_p); \\ \sum_{i=1}^{p} w_{ij}^2 = 1 \to (j = \overline{1,p}); \\ \sum_{i=1}^{p} w_{ij} w_{ik} = 0 \to (j,k = \overline{1,p}, j \neq k); \end{cases} \tag{1.34}$$

де m_i – математичне очікування ознаки x_i.

Метод головних компонентів заснований на виділенні лінійних комбінацій випадкових величин, що мають максимально можливу дисперсію, за допомогою матриці парних кореляцій [4, 8]. Лінійні комбінації вибираються таким чином, що серед всіх можливих лінійних нормованих комбінацій вихідних ознак перший головний компонент $y_1(x)$ має найбільшу дисперсію. Геометрично це виглядає як орієнтація нової координатної осі y_1 уздовж напрямку найбільшої витягнутості еліпсоїда розсіювання об'єктів досліджуваної вибірки в просторі ознак $x_1, ..., x_p$. Другий головний компонент має найбільшу дисперсію серед всіх лінійних перетворень, що залишилися, не корельованих з першим головним компонентом. Він інтерпретується як напрямок найбільшої витягнутості еліпсоїда розсіювання, перпендикулярний першому головному компоненту. Наступні головні компоненти визначаються за аналогічною схемою. У зв'язку з цим з'являється можливість виразити інформацію, що утримується у великому наборі вихідних ознак, за допомогою меншого числа незалежних головних компонентів. Метод застосовується для системи нормованих числових ознак.

Обчислення коефіцієнтів головних компонентів w_{ij} засновано на тому факті, що вектори $w_i = (w_{11}, ..., w_{p1})', ..., w_p = (w_{1p}, ..., w_{pp})'$ є власними (характеристичними) векторами кореляційної матриці S. У свою чергу, відповідні власні числа цієї матриці дорівнюють дисперсіям проекцій множини об'єктів на осі головних компонентів.

Алгоритми, що забезпечують виконання методу головних компонентів, входять практично в усі пакети статистичних програм.

Факторний аналіз [4, 8] заснований на припущенні, що вихідні ознаки є проявами невеликого числа об'єктивно існуючих, але таких, що не піддаються безпосередньому вимірові факторів, що детермінують розходження між об'єктами. В описаному вище методі головних компонентів під критерієм автоінформативності простору ознак розуміють те, що цінну для діагностики інформацію можна відобразити в лінійній моделі, яка відповідає новій координатній осі в даному просторі з максимальною дисперсією розподілу проекцій досліджуваних об'єктів. Такий підхід є продуктивним, коли явна більшість завдань «чорнового» варіанта тесту узгоджено «працює» на прояв властивості, що тестується й придушує вплив ірелевантних факторів на розподіл об'єктів. Також позитивний результат буде отриманий при порівняно невеликому обсязі групи зв'язаних інформативних ознак, але при неузгодженій взаємодії сторонніх факторів, під впливом яких не порушується однорідність еліпсоїда розсіювання, а лише зменшується витягнутість розподілу об'єктів уздовж напрямку тенденції, що діагностується. На відміну від методу головних компонентів, факторний аналіз заснований не на дисперсійному критерії автоінформативності системи ознак, а він орієнтований на пояснення наявних між ознаками

кореляцій. Тому факторний аналіз застосовується в більш складних випадках спільного прояву на структурі експериментальних даних тестуємої й іррелевантної властивостей об'єктів, порівнянних за ступенем внутрішньої погодженості, а також для виділення групи діагностичних показників із загальної вихідної множини ознак. Область застосування методу – система числових ознак. Основна модель факторного аналізу записується такою системою рівностей

$$x_i = \sum_{j=1}^{m} l_{ij} f_j + \varepsilon_i \,, \qquad (1.35)$$

де $i = 1, ..., p$ – кількість ознак; $m<p$ – число факторів; l_{ij} – навантаження i-ї ознаки на j-й фактор.

Таким чином припускається, що значення кожної ознаки x_i може бути виражене зваженою сумою латентних змінних (простих факторів) f_j, кількість яких менша від числа вихідних ознак, і залишковим членом ε_i (специфічним фактором) з дисперсією σ^2, що діє тільки на x_i.

Коефіцієнти l_{ij} називаються навантаженням i-ї змінної на j-й фактор або навантаженням j-го фактора на i-ту змінну. У найпростішої моделі факторного аналізу вважається, що фактори f_j взаємно незалежні і їхні дисперсії дорівнюють одиниці, а випадкові величини ε_i теж незалежні одне від одної та від будь-якого фактора f_j. Максимально можлива кількість факторів m при заданому числі ознак p визначається нерівністю

$$(p+m)<(p-m)^2,$$

яка повинна виконуватися, щоб задача не вироджувалася в тривіальну.

Дана нерівність одержана на підставі підрахунку ступенів свободи, що є у задачі. Суму квадратів навантажень у формулі основної моделі факторного аналізу називають спільністю відповідної ознаки x_i і чим більше це значення, тим краще описується ознака x_i виділеними факторами f_j. Спільність є частиною дисперсії ознаки, яку пояснюють фактори. У свою чергу, величина ε_i^2 показує, яка частина дисперсії вихідної ознаки залишається непоясненою при використовуваному наборі факторів, що використовується, і дану величину називають специфічністю ознаки. Таким чином,

$$Дисперсія\_ознаки = спільність(\sum_{j-1}^{m} l_{ij}^2) + специфічність(\varepsilon_i^2) \,.$$

Основне співвідношення факторного аналізу (1.35) показує, що коефіцієнт кореляції будь-яких двох ознак x_i і x_j можна виразити сумою добутку навантажень некорельованих факторів

$$r_{ij} = r(x_i, x_j) = l_{i1} l_{j1} + l_{i2} l_{j2} + \cdots + l_{im} l_{jm} \,.$$

Задачу факторного аналізу не можна розв'язати однозначно. Рівності основної моделі факторного аналізу (1.35) не піддаються безпосередній перевірці, тому що p вихідних ознак задається через $(p+m)$ інших змінних – простих і специфічних факторів. Тому представлення кореляційної матриці факторами (факторизацію) можна зробити нескінченно великим числом способів. Якщо вдалося зробити факторизацію кореляційної матриці за допомогою деякої матриці факторних навантажень F, то будь-яке лінійне ортогональне перетворення F (ортогональне обертання) приведе до такого ж результату.

Існуючі програми обчислення навантажень починають працювати з однофакторної моделі ($m = 1$). Потім перевіряється, наскільки кореляційна матриця, відновлена за однофакторною моделлю відповідно до виразу (1.35), відрізняється від кореляційної матриці вихідних даних. Якщо однофакторна модель визнається незадовільною, то випробовується модель із $m = 2$ і т. д. Це триває доти, поки при деякому m не буде досягнута адекватність або поки число факторів у моделі не перевищить максимально припустиме. В останньому випадку говорять, що адекватної моделі факторного аналізу не існує. Якщо факторна модель існує, то виконується обертання отриманої системи загальних факторів, тому що значення факторних навантажень і навантажень на фактори є лише одним з можливих розв'язків основної моделі. Обертання факторів може виконуватися різними способами. Найбільш часто воно здійснюється таким чином, щоб якомога більше число факторних навантажень стало нулями й щоб кожний фактор по можливості описував групу сильно корельованих ознак. Також можна обертати фактори доти, поки не будуть одержані результати, які піддаються змістовній інтерпретації. Можна, наприклад, забажати, щоб один фактор був навантажений переважно ознаками одного типу, а інший – ознаками іншого типу. Або, скажемо, можна забажати, щоб зникли якісь важко інтерпретуємі навантаження з негативними знаками. Нерідко дослідники йдуть далі й розглядають прямокутну систему факторів як окремий випадок косокутної, тобто заради змісту жертвують умовою некорельованості факторів.

На завершення всієї процедури факторного аналізу за допомогою математичних перетворень виражають фактори f_j через вихідні ознаки, тобто одержують у явному вигляді параметри лінійної діагностичної моделі.

Відома велика кількість методів факторного аналізу (ротацій, максимальної правдоподібності й ін.). Нерідко в тому самому пакеті програм аналізу даних реалізовано відразу кілька версій таких методів, і тому виникає правомірне питання про те, який з них кращий.

Метод контрастних груп [4, 8]. Вихідною інформацією при використанні методу контрастних груп, крім таблиці експериментальних даних з результатами обстеження «чорновим» варіантом діагностичного

тесту, є також «чорнова» версія лінійного правила обчислення показника, що тестується. Ця «чорнова» версія може бути складена експериментатором, виходячи з його теоретичних уявлень про те, які ознаки й з якими вагами повинні бути включені в лінійну діагностичну модель. Крім того, «чорнова» версія може бути почерпнута з літературних джерел, коли в експериментатора виникає потреба адаптувати опублікований діагностичний тест до нових умов. Метод контрастних груп застосовується також у складі процедури підвищення внутрішньої погодженості завдань раніше відпрацьованого тесту.

В основі методу контрастних груп лежить гіпотеза про те, що значна частина «чорнової» версії діагностичної моделі підібрана або вгадана правильно. Тобто в праву частину рівняння $y_ч = y_ч(x)$ увійшло досить багато ознак, що узгоджено відображають тестуєму властивість, що тестується. У той же час в «чорновій» версії $y_ч(x)$ певна частка ознак є непотрібною або навіть шкідливим баластом, від якого потрібно позбутися. Як і у всіх інших методах, що спираються на категорію внутрішньої погодженості, це означає, що в просторі ознак, включених у вихідну діагностичну модель, розподіл об'єктів уписується в еліпсоїд розсіювання, витягнутий уздовж напрямку діагностуємої тенденції. У свою чергу, вплив інформаційного баласту виражається в зменшенні такої витягнутості еліпсоїда розсіювання, тому що «шумові» ознаки збільшують розкид досліджуваних об'єктів по всіх інших напрямках. При цьому «зашумлення» основної тенденції буде тим сильніше, чим ближче до центра розподілу розташуються діагностуємі об'єкти, і тим слабкіше, чим ближче до полюсів головної осі еліпсоїда розсіювання перебуватимуть розглянуті об'єкти. Це пов'язане з тим, що попадання об'єктів у крайні області пояснюється, головним чином, кумулятивним ефектом погодженої взаємодії інформативних ознак. Описані уявлення про структуру експериментальних даних лежать в основі наступної процедури.

Спочатку призначаються вихідні шкальні ключі (ваги) w°_j для пунктів тесту (дихотомічних ознак) x_j. Для кожного i-го випробування підраховується сумарний тестовий бал

$$y_ч(x) = \sum_{j=1}^{p} w^{\circ}_j x_j \,. \qquad (1.36)$$

Зазвичай абсолютні значення ваг w_j визначають приблизно й часто беруть такими, що дорівнюють одиниці. Тому напрямок (1.36) буде трохи відрізнятися від напрямку головної діагоналі еліпсоїда розсіювання $y(x)$ (рис. 1.4).

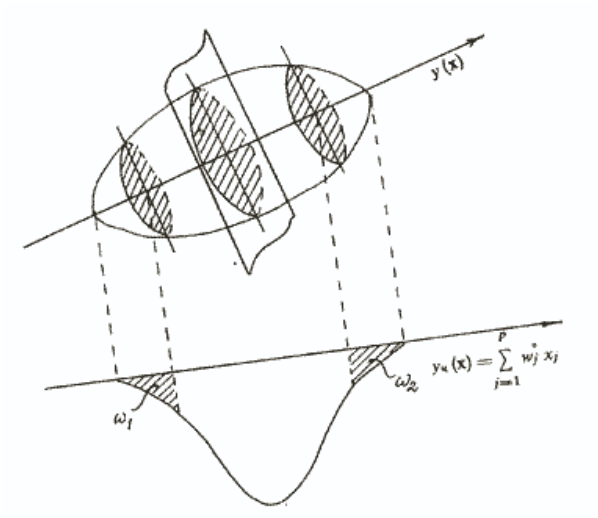

Рис. 1.4. Ілюстрація методу контрастних груп

Але якщо орієнтовно $y_ч(x)$ правильно відображає діагностуєму властивість, то на краях розподілу сумарного бала, побудованого по всіх об'єктах досліджуваної вибірки, можна виділити контрастні групи ω_1 і ω_2, у які ввійдуть об'єкти з мінімальними похибками, внесеними «шумовими» ознаками. Ці групи не повинні бути занадто малі. Для нормального розподілу, як правило, беруть контрастні групи обсягом 27 % від загального обсягу вибірки, для більше плоского – 33 %. У принципі вважається прийнятною будь-яка цифра від 25 до 33 %.

Наступний крок полягає у визначенні ступеня зв'язку кожного пункту з дихотомічної змінної – номером контрастної групи. Мірою цього зв'язку може служити так званий коефіцієнт розрізнення, що являє собою різницю відсотків тієї або іншої відповіді на аналізований пункт у полярних групах випробуваних. Найбільше часто використовується коефіцієнт зв'язку Пірсона φ, який потім порівнюється із граничним значенням

$$\left|\varphi_{гр}\right| = \sqrt{\chi^2{}_{гр} / N} ,$$

де $\chi^2{}_{гр}$ – стандартний квантиль розподілу χ_2 з одним ступенем свободи.
Зазвичай орієнтуються на рівні значимості 5 % і 1 % , для яких значення χ^2 дорівнює 3,84 і 6,63 відповідно. Якщо для *i*-го пункту $|\varphi_i|<|\varphi_{гр}|$, то ваговому коефіцієнту w_i привласнюється значення нуля, тобто ознака x_i виключається з лінійної діагностичної моделі $y_ч(x)$. У такий спосіб перевіряються всі пункти «чорнового» варіанта тесту. Потім для пунктів, що залишилися, вся

процедура знову повністю повторюється й т. д.

На практиці не зустрічаються випадки, коли остаточно відібрані за допомогою наведеної процедури інформативні ознаки абсолютно б збіглися з початково заданими. Збіжність цієї процедури залежить від вихідного співвідношення «гарних» і «поганих» завдань тесту. Очевидно, для діагностичних моделей, заснованих на принципі внутрішньої погодженості ознак, у кожному конкретному завданні існує певний поріг співвідношення інформативних і «шумових» ознак, починаючи з якого можливе виникнення ефекту самоорганізації або самовдосконалення діагностичної моделі за допомогою описаного вище алгоритму.

Багатовимірне шкалювання [8, 12] – сукупність методів, які дозволяють за заданою інформацією про міри розходження (близькості) між об'єктами приписувати кожному з цих об'єктів вектор його кількісних характеристичних показників. При цьому розмірність шуканого координатного простору задається заздалегідь, а "занурення" у нього аналізованих об'єктів виконується таким чином, щоб структура взаємних розходжень між ними, яка вимірюється за допомогою приписуваних їм допоміжних координат, у середньому найменш відрізнялася б від заданої в сенсі того або іншого функціоналу якості. Визначення координат об'єктів у просторі і розмірності простору засноване на перетворенні матриці відстаней у матрицю скалярних добутків центрованих векторів. Таким чином, на виході алгоритму виходять числові значення координат, що приписуються кожному об'єктові в деякій новій системі координат (у "допоміжних шкалах", зв'язаних з латентними змінними), причому розмірність нового простору ознак істотно менша від розмірності вихідного. При функціональному шкалюванні [13] будується єдиний (інтегральний) показник. Методи багатовимірного шкалювання застосовуються в ряді випадків, коли більшість методів факторизації є непридатною.

Дисперсійний аналіз [14, 15] був розроблений у 20-х роках XX сторіччя англійським математиком і генетиком Рональдом Фішером. На дисперсійному аналізі заснований широкий клас критеріїв значущості. Дисперсійний аналіз дозволяє визначити вплив різних факторів (умов) x_i на досліджувану ознаку (явище) y, що досягається шляхом розкладання сукупної дисперсії D_y на окремі компоненти, викликані впливом різних джерел мінливості. При цьому перевіряється тільки наявність або відсутність статистично значущого зв'язку x_i з y, тому метод застосовується для попередньої оцінки і добору значущих факторів, і для відсівання всіх інших.

Вхідні змінні x_i представляються номінальними, ординальними або дихотомічними шкалами, вихідна ознака y зазвичай є числовою. При аналізі числових факторів x_i, їх необхідно звести до ординальної шкали, виконуючи квантування динамічного діапазону Δx на p рівнів.

Для незалежної (або слабозалежної) системи факторів застосовується однофакторний дисперсійний аналіз, а для врахування сумісної дії факторів – двофакторний дисперсійний аналіз.

Однофакторний дисперсійний аналіз. Ідея методу заснована на оцінці вибіркових середніх m_y і дисперсій D_y по вибірці. Якщо вибірки отримані при різних значеннях рівнів факторів (дві групи при бінарних факторах, p груп при номінальних або ординарних ознаках, що мають p градацій), то виникає таке питання: чи розходження в оцінках m_y і D_y викликані впливом факторів, або вони випадкові? Використовуються методи оцінки статистичної значущості розходжень (їх називають критеріями значущості, або просто критеріями). Методів цих існує безліч, але усі вони побудовані на одному принципі. Спочатку формулюється нульова гіпотеза H_0, тобто припускається, що фактори x_i, які досліджуються, не роблять ніякого впливу на досліджувану величину y, й отримані розходження оцінок випадкові і обумовлені обмеженістю вибірки. Потім ми визначаємо, яка імовірність одержати розходження, що спостерігаються (або більш сильні) за умови справедливості нульової гіпотези. Якщо ця імовірність менша від заданого порогового значення, то нульова гіпотеза відкидається і робиться висновок, що результати експерименту статистично значущі (приймається гіпотеза H_1). Це ще не означає, що ми довели дію саме досліджуваних факторів (це питання насамперед планування експерименту), але, у всякому разі, малоймовірно, що результат обумовлений випадковістю.

Інтуїтивно зрозуміло, що вибірки «не розрізняються», якщо розкид вибіркових середніх значно менший від розкиду значень у кожній з вибірок, і навпаки – вибірки «розрізняються», якщо розкид вибіркових середніх перевищує розкид у кожній з вибірок. Залишилося тільки формалізувати це судження та оформити його кількісно.

Дослідник не може спостерігати генеральну сукупність, з якої взяті вибірки. Усе, що він може аналізувати – це його експериментальні групи. Дисперсію сукупності D_y можна оцінити двома способами.

По-перше, дисперсія, обчислена для кожної групи $D_{yi}\ i=\overline{1,p}$, – це оцінка дисперсії сукупності D_y. Тому дисперсію сукупності D_y можна оцінити на підставі групових дисперсій D_{yi}. Така оцінка не буде залежати від розходжень групових середніх.

По-друге, розкид вибіркових середніх теж дозволяє оцінити дисперсію сукупності D_y. Зрозуміло, що така оцінка дисперсії залежить від розходжень вибіркових середніх.

Якщо експериментальні групи – це p випадкові вибірки з тієї самої нормально розподіленої сукупності (це означає, що фактор не впливає на y), то обидві оцінки дисперсії сукупності D_y дали б приблизно однакові результати. Тому, якщо ці оцінки виявляються близькими, то ми не можемо

відкинути нульову гіпотезу. У противному разі ми відкидаємо нульову гіпотезу H_0, тобто вважаємо малоймовірним те, що ми одержали би такі розходження між групами, якби вони були просто випадковими вибірками з однієї нормально розподіленої сукупності.

Перейдемо до обчислень. Як оцінити дисперсію сукупності D_y за p вибірковими дисперсіями D_{yi}? Якщо вірна гіпотеза H_0 про те, що x не впливає на величину y, то кожна з них дає однаково гарну оцінку. Тому за оцінку дисперсії сукупності D_y візьмемо середнє вибіркових дисперсій D_{yi}. Ця оцінка називається внутрішньогруповою дисперсією, що позначається як D_R

$$D_R = \frac{\sum_i D_{yi}}{p}. \tag{1.37}$$

Оцінимо тепер дисперсію сукупності за вибірковими середніми m_y. Оскільки ми припустили, що всі p вибірки витягнуті з однієї сукупності, стандартне відхилення p вибіркових середніх служить оцінкою помилки середнього. Стандартна помилка середнього σ_{m_y} зв'язана зі стандартним відхиленням сукупності σ_y і з обсягом вибірки n наступним співвідношенням:

$$\sigma_{m_y} = \frac{\sigma_y}{\sqrt{n}}. \tag{1.38}$$

Таким чином, дисперсію сукупності σ_y^2 можна розрахувати в такий спосіб:

$$D_A = \sigma_y^2 = n\sigma_{m_y} \tag{1.39}$$

Ця оцінка називається міжгруповою дисперсією, що позначається як D_A. Якщо вірна нульова гіпотеза, то тоді як внутрішньогрупова, так і міжгрупова дисперсії служать оцінками тієї ж самої дисперсії і повинні бути приблизно рівні. Виходячи з цього, обчислюється критерій значущості Фішера F:

$$F = D_A/D_R. \tag{1.40}$$

І чисельник, і знаменник цього відношення – це оцінки тієї ж самої величини – дисперсії сукупності D_y, тому значення F повинне були близьке до 1. Отже, якщо F значно перевищує 1, нульову гіпотезу H_0 варто відкинути, у противному разі H_0 варто прийняти. Залишилося зрозуміти, починаючи з якої саме величини F варто відкидати H_0.

Якщо витягати випадкові вибірки обсягом n з нормально розподіленої сукупності N ($N >> n$), то значення F буде мінятися з кожним експериментом,

тому F є випадковою величиною, яка має розподіл Фішера $f(F)$. При цьому вид кривої розподілу залежить від параметрів n та p.

Значення критерію, починаючи з якого ми відкидаємо нульову гіпотезу, називається критичним значенням F_k. Імовірність помилково відкинути вірну нульову гіпотезу, тобто знайти розходження там, де їх немає, позначається як P та визначається як площа під кривою $f(F)$ на інтервалі $F>F_k$. Імовірність помилки P називається рівнем значущості α (зазвичай $\alpha = 0{,}05$ або $0{,}01$).

Таким чином, критичне значення F_k однозначно визначається рівнем значущості α і ще двома параметрами, що називаються внутрішньогруповим $\nu_{вну}$ і міжгруповим числом ступенів свободи $\nu_{між}$, значення яких визначається на основі формул (1.37, 1.39):

$$\nu_{між} = p - 1, \ \nu_{вну} = p\,(n - 1).$$

Обчислити критичне значення F_k досить складно, тому користуються таблицями критичних значень F_k для різних α, $\nu_{між}$ і $\nu_{вну}$. Після обчислення F та F_k приймається таке рішення:

H_0: якщо $F < F_k$;
H_1: якщо $F > F_k$.

При плануванні експериментів для проведення дисперсійного аналізу кількість експериментів n на кожному рівні p вибирається однаковою, хоча в роботі [15] наведено вирази, аналогічні формулам (1.38, 1.40) для різних обсягів груп, які застосовуються при пасивному експерименті.

При виконанні розрахунку за допомогою програмних пакетів (наприклад, StatGraphics) результати розрахунків видаються у вигляді табл. 1.3. Наведені у цій таблиці середні значення на кожному i-му рівні фактора y_{i0} та по всій вибірці y_{00} розраховуються за формулами:

$$y_{i0} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}; \qquad y_{00} = \frac{1}{q \cdot n}\sum_{j=1}^{q}\sum_{i=1}^{n} y_{ij}\ .$$

Якщо взяти рівень значущості $\alpha = 0{,}05$, то при $\alpha > 0{,}05$ приймається гіпотеза H_0, а при $\alpha < 0{,}05$ – H_1.

Якщо справедлива гіпотеза H_1, то для визначення відхилень у рівнях фактора x_i звичайно розраховуються граничні (95 % довірчий інтервал) відхилення вихідної величини y при заданому рівні фактора p.

Таблиця 1.3 – Результати однофакторного дисперсійного аналізу

Компоненти дисперсії	Сума квадратів	Число ступенів свободи	Середньо-квадратичне відхилення	Відхилення Фішера	Рівень значу-щості
Між рівнями фактора	$S_A = \sum_{i=1}^{p} n(y_{i0} - y_{00})^2$	$p - 1$	$d_A = \frac{S_A}{p-1}$	$\frac{d_A}{d_R}$	α
Усередині факторів	$S_R = \sum_i \sum_j (y_{ij} - y_{i0})^2$	$p(n-1)$	$d_R = \frac{S_R}{p(n-1)}$		
Повна дисперсія	$S_G = \sum_{i=1}^{p} \sum_{j=1}^{n} (y_{ij} - y_{00})^2$	$pn - 1$			

Двофакторний дисперсійний аналіз. Задача двосторонньої класифікації виникає при проведенні спостережень в експерименті, у якому одночасно діють два фактори (x_i та x_j), які мають p та q рівнів варіацій відповідно. Тому результати вимірів заносяться в двовимірну таблицю Y розміром $p*q$, причому кожна її комірка містить результати n вимірів вихідної величини y_{ijk}, $i = \overline{1,p}$, $j = \overline{1,q}$, $k = \overline{1,n}$.

Середні значення по рядках, по стовпцях і повні середні розраховуються за формулами

$$y_{ij0} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}; \qquad y_{i00} = \frac{1}{q \cdot n}\sum_{j=1}^{q}\sum_{k=1}^{n} y_{ijk};$$

$$y_{0j0} = \frac{1}{p \cdot n}\sum_{i=1}^{p}\sum_{k=1}^{n} y_{ijk}; \qquad y_{000} = \frac{1}{p \cdot q \cdot n}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} y_{ijk}.$$

Дисперсії розраховуються за формулами

$$S_G = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - y_{000})^2; \qquad S_A = q \cdot r\sum_{i=1}^{p}(y_{i00} - y_{000})^2;$$

$$S_B = p \cdot r\sum_{r}^{q}(y_{0j0} - y_{000})^2; \qquad S_{AB} = r\sum_{i=1}^{p}\sum_{j=1}^{a}(y_{ij0} - y_{000})^2;$$

$$S_R = r\sum_{i=1}^{p}\sum_{j=1}^{a}\sum_{k=1}^{r}(y_{ijk} - y_{ij0})^2;$$

Результати розрахунків двофакторного дисперсійного аналізу видаються у вигляді табл. 1.4.

Таблиця 1.4 – Результати двофакторного дисперсійного аналізу

Компоненти дисперсії	Сума квадратів	Число ступенів свободи	Середньо-квадратичне відхилення	Відношення Фішера	Рівень значущості
Між факторами А	S_A	p-1	$\frac{S_A}{p-1}$	$\frac{S_A/(p-1)}{S_R/pq(n-1)}$	α_A
Між факторами В	S_B	q-1	$\frac{S_B}{q-1}$	$\frac{S_B/(q-1)}{S_R/pq(n-1)}$	α_B
Між взаємодією АВ	S_{AB}	$(p-1)(q-1)$	$\frac{S_{AB}}{(p-1)(q-1)}$	$\frac{S_{AB}/(p-1)(q-1)}{S_R/pq(n-1)}$	α_{AB}
Помилка (усередині фактора)	S_R	$pq(n-1)$	$\frac{S_R}{pq(n-1)}$		
Повна	S_G				

Таким чином: $S_G = S_A + S_B + S_{AB} + S_R$.

де S_G визначає повну дисперсію; S_A – дисперсію за фактором A; S_B – дисперсію за фактором B; S_{AB} – дисперсію за взаємодією A та B; S_R – дисперсію вимірів (помилку).

Перевіряється гіпотеза H_0.

Якщо $\left.\begin{array}{l}\alpha_A > 0{,}05\\ \alpha_B > 0{,}05\\ \alpha_{AB} > 0{,}05\end{array}\right\}$, то має місце гіпотеза H_0.

Якщо справедлива гіпотеза H_0, то вплив факторів A, B та їх взаємодії AB статистично не значущі. Якщо справедлива H_1, то можна визначити, які з факторів впливають на результати виміру вихідної величини y.

Регресійний аналіз. [8, 16]. З позиції регресійного аналізу розглядається зв'язок між вектором вхідних змінних $X = (x_i, ..., x_p)$ і вихідною (залежною) змінною y:

$$y = f(X) + \varepsilon. \qquad (1.41)$$

Для оцінки ефективності регресійної діагностичної моделі вводиться вектор залишків $\varepsilon=(\varepsilon_1, ..., \varepsilon_n)'$, що відображує вплив на y сукупності неврахованих випадкових факторів або міру досяжної апроксимації значень y

функціями типу $y = f(x)$. Зазвичай $f(x)$ шукається в вигляді полінома з цілочисельними значеннями степенів x. За кількістю вхідних змінних (факторів) розрізняють одно- та багатофакторні (множинні) регресії, за степенем полінома – лінійні та нелінійні. У найбільш загальному випадку (нелінійна множинна регресія) рівняння регресії задається у вигляді полінома Колмогорова – Габора

$$\hat{y} = P(x_1,\ldots,x_n) = a_0 + \sum_i a_i x + \sum_i \sum_j a_{ij} x_i x_j + \sum_i \sum_j \sum_k a_{ijk} x_i x_j x_k \ldots \quad (1.42)$$

Задачею регресійного аналізу є визначення параметрів моделі – коефіцієнтів полінома (1.42) на основі аналізу синхронно виміряних значень вектора вхідних даних $X = (x_i, \ldots, x_p)$ та вихідних значень y. Розрізняють два підходи залежно від походження матриці даних. У першому вважається, що вектор ознак X є детермінованим, а випадковою величиною є тільки залежна змінна y. Ця модель використовується найбільш часто й називається моделлю з фіксованою матрицею даних. У другому підході вважається, що вектор ознак X і y – випадкові величини, що мають спільний розподіл. У такій ситуації оцінка рівняння регресії є оцінкою умовного математичного очікування випадкової величини y залежно від випадкових величин $x_i, \ldots, x_p$. Дана модель називається моделлю з випадковою матрицею даних. Кожний з наведених підходів має свої особливості. У той же час показано, що моделі з фіксованою матрицею даних і з випадковою матрицею даних відрізняються тільки статистичними властивостями оцінок параметрів рівняння регресії, тоді як обчислювальні аспекти цих моделей збігаються. Слід зазначити, що число вимірювань (точок) N повинне бути не менше від числа коефіцієнтів n у виразі (1.42), а з врахуванням того, що вхідні дані є випадковими величинами, число вимірювань N вибирається за умови $N>>n$.

У рівнянні функції регресії (1.41) зазвичай вважається, що величини ε_i (i=1, N) незалежні й випадково розподілені з нульовим середнім і дисперсією σ^2_ε, а оцінка параметрів a_i виконується за допомогою методу найменших квадратів (МНК).

Всі зазначені раніше види регресійних залежностей можна звести до лінійної багатофакторної регресії шляхом заміни всіх складових у виразі (1.42) новими вхідними змінними $z_i,\ i = \overline{1,n}$:

$$z_1 = x_1, \ldots, z_p = x_p, z_{p+1} = x_1 x_2, \ldots, z_n = x_p x_p \ldots x_p.$$

Якщо ввести фіктивну змінну $z_0 \equiv 1$ та підставити вираз (1.42) в (1.41), то одержимо

$$y_i = \sum_{j=1}^{n} a_j z_{ji} + \varepsilon_i . \quad (1.43)$$

Оскільки величина y обчислюється завжди з деякою помилкою ε, то можна обчислити суму квадратів відхилень між виміряним та розрахунковим значенням y_i:

$$S = \sum_{i=1}^{N} (y_j - \sum_{j=1}^{n} a_j z_{ji})^2 . \qquad (1.44)$$

При використанні МНК сума квадратів відхилення повинна прямувати до мінімальної. Так як функція S залежить від коефіцієнтів регресії a_j, їх потрібно вибрати таким чином, щоб $S \to \min$, тобто

$$\frac{\partial S}{\partial a_j} = 2\sum_{i=1}^{N} (y_j - \sum_{j=1}^{n} a_j z_{ji}) z_j = 0;\ j = \overline{1,n} . \qquad (1.45)$$

Це приводить до нормальної системи лінійних рівнянь відносно невідомих коефіцієнтів регресії:

$$\boldsymbol{S^*A=C_{yz}}, \qquad (1.46)$$

де $\boldsymbol{S}$ – матриця коваріації ознак z_0, ..., z_n, елементи якої обчислюються за формулою $s_{ij} = \sum_{k=1}^{N} z_{ik} z_{jk}$; $\boldsymbol{C_{yz}}$ – вектор оцінок коваріації між вихідною величиною y та ознаками z_0, ..., z_n, елементи якого обчислюються за формулою $c_i = \sum_{k=1}^{N} y_k z_{jk}$; $\boldsymbol{A}$ – вектор коефіцієнтів, які потрібно обчислити.

Розв'язання системи рівнянь (1.46) виконується будь-яким відомим методом, доступним для дослідника (метод Гаусса, квадратного кореня та ін.). Можна визначити довірчі інтервали та гіпотези щодо значущості коефіцієнтів a_i за виразом (1.31), тому що значення t_a, як і t_r в кореляційному аналізі розподілені за законом Стьюдента з $(N-2)$ ступенями свободи.

Перевірка гіпотези про адекватність регресійної моделі виконується з використанням критерію Фішера. Якщо перевірити гіпотезу H_0 про те, що розглянута модель адекватна об'єктові, то для перевірки цієї гіпотези необхідно зіставити досягнуту точність моделі з величиною, що характеризує точність спостережень. Якщо точність моделі перевершує точність спостережень, то гіпотеза H_0 відхиляється.

Порівняння дисперсій виконується за формулою

$$W_0 = \frac{\frac{S_d}{N-n-1}}{\frac{S_y}{N-1}}, \qquad (1.47)$$

де S_d – сума квадратів відхилення експериментальних точок і точок моделі, яка розраховується за виразом (1.44); S_y – сума квадратів відхилення вихідної величини y від свого середнього m_y, яка розраховується за формулою

$$S_y = \sum_{i=1}^{N} (y_i - m_y)^2 \ .$$

Якщо $W_0 < W_{0,95}(N - n - 1, N - 1)$, то приймається гіпотеза H_0. Інакше вона відкидається. Крім того, використовуються показники якості регресійної діагностичної моделі, а саме:

- залишкова сума квадратів, яка розраховується за виразом (1.44);
- незміщена оцінка дисперсії помилки $s_\varepsilon^2 = S_d / (N - n)$;
- оцінка дисперсії прогнозованої змінної $\sigma_y^2 = \frac{1}{N} S_y$;
- коефіцієнт детермінації $R^2 = \frac{N\sigma_y^2 - S_d}{N\sigma_y^2}$, який змінюється від 0 (відсутній зв'язок X з y) до 1 (повністю детермінований зв'язок);
- оцінка дисперсії коефіцієнтів регресії $D_{ai} \approx \frac{s_\varepsilon^2}{N} s_{ii}$, де s_{ii} – відповідний елемент коваріаційної матриці ознак.

Складна регресійна залежність. Якщо регресійна залежність має складний характер, то підібрати відповідний поліном практично не вдається. Тому використовується кусково-параметрична апроксимація, при цьому вісь X розбивається на кілька ділянок, у яких вибирається лінійна або квадратична апроксимація, яка щонайкраще описує цю ділянку. Часто для цих цілей використовується сплайнова апроксимація.

Методи самоорганізації регресійних моделей. Як було зазначено раніше, визначення параметрів моделі (коефіцієнтів полінома) за МНК зводиться до розв'язання нормальної системи лінійних алгебраїчних рівнянь щодо невідомих коефіцієнтів полінома. При значному числі вихідних ознак і при збільшенні ступеня полінома число коефіцієнтів у виразі (1.42) зростає лавиноподібно, що висуває підвищені вимоги до обсягу навчальної вибірки (для реалізації МНК число точок має бути істотно більше від сумарного числа коефіцієнтів). Крім того, у реальних даних, як правило, зустрічаються групи сильнозв'язаних ознак. У цих умовах виникає явище мультиколінеарності, що приводить до поганої обумовленості і, у граничному випадку, до виродження коваріаційної матриці. При цьому розв'язання нормальної системи лінійних рівнянь є нестійким, або розв'язок одержати не можна. Тому на практиці зазвичай обмежуються лінійними регресійними моделями, хоча вони неточні і

використовуються для "грубої" оцінки з метою добору множини впливаючих факторів моделі *X*.

Для синтезу регресійної моделі за невеликим числом експериментальних даних використовуються методи самоорганізації [17], які одночасно визначають структуру полінома і його коефіцієнти. У методах самоорганізації використовується ітераційна процедура послідовного ускладнення полінома з вибором найкращих рішень на кожному кроці ітерації. Особливістю методів самоорганізації є застосування принципів неостаточності рішення і зовнішнього критерію.

Принцип неостаточності рішення запозичений від еволюційних і генетичних алгоритмів і полягає в тому, що на кожному кроці залишається не один кінцевий результат, а група найкращих рішень, тобто відтинаються безперспективні рішення, і тільки на останньому кроці з усіх найкращих рішень вибирається єдине найкраще (оптимальне) серед усіх рівних.

Принцип зовнішнього критерію полягає в тому, що оцінка якості прогнозуючої моделі (точності наближення експериментальних даних до рівнянь регресії) виконується за допомогою критерію, що є зовнішнім стосовно критерію, за допомогою якого визначаються коефіцієнти. Одним з варіантів зовнішнього критерію може бути розбиття всіх експериментальних точок на дві частини, де перша з них (навчальна вибірка) служить для визначення коефіцієнтів по МНК, а друга (перевірочна (зовнішня) вибірка) – для оцінки точності рівняння регресії. Застосування зовнішнього критерію дозволяє одержувати модель оптимальної складності.

Суть алгоритмів самоорганізації полягає в тому, що повний опис об'єкта виду (1.42) заміняється множиною часткових описів. Як часткові описи беруться поліноми ступеня, що не вищий від другого, причому від числа змінних не більше двох (лінійні, з коваріаційними частковими описами, із квадратичними частковими описами). Ускладнення моделі (збільшення числа змінних і ступеня полінома) виконується при переході до наступного кроку ітерації, причому результати найкращих моделей попереднього кроку (відносно зовнішнього критерію) є вихідними даними наступного кроку ітерації. Безліч алгоритмів самоорганізації відрізняються структурою часткових описів, зовнішнім критерієм (точність моделі на точках перевірочної послідовності, баланс коефіцієнтів), способом одержання результуючої моделі й ін.

1.2.4. Методи оцінки інформативності та формування інформативного простору ознак

Теоретико-інформаційний підхід. Оцінка інформативності різнорідних ознак. Теоретико-інформаційний підхід [18] є найбільш строгим і

формалізованим методом оцінки інформативності окремих ознак (або комплексу ознак) щодо заданої системи станів $\{D\}_n$. У даному методі для оцінки інформативності використовуються такі фундаментальні поняття теорії інформації, як ентропія та кількість внесеної інформації [19].

Якщо система станів утворює повну групу несумісних подій (кожен об'єкт в навчальній вибірці належить тільки одному стану D_i, немає об'єктів з декількома станами), то сумарна ймовірність всіх станів дорівнює 1:

$$\sum_{i=1}^{n} P(D_i) = 1 .$$

Тоді невизначеність системи можливих станів оцінюється за допомогою ентропії

$$H(D) = -\sum_{i=1}^{n} P(D_i) \cdot \log_2 P(D_i) . \qquad (1.48)$$

При цьому для n можливих рівноймовірних складових її значення буде максимальним:

$$H(D) = \log_2 n,$$

а у випадку диференціальної діагностики ($n = 2$) $H(D)$ не перевищить 1 біта.

Оскільки ентропія відображає міру невизначеності системи, її величина буде змінюватися при надходженні в систему нової інформації. Такою інформацією для станів є дані, отримані в результаті вимірювання ознак об'єкта, що діагностується. Зменшення ентропії відбувається на величину, що дорівнює кількості внесеної інформації. Крайнє значення, якого може набувати ентропія, дорівнює нулеві і має місце для достовірної події. У цьому випадку нуль показує відсутність невизначеності в системі.

Відповідно кількість інформації, що надійшла в систему, визначається як різниця між величиною ентропії до і після вимірювання:

$$I_D(x_j) = H(D) - H(D / x_j) , \qquad (1.49)$$

де $I_D(x_j)$ – кількість інформації, внесеної в систему після проведення вимірювання ознаки x_j; $H(D)$ – початкова ентропія системи діагнозів; $H(D/x_j)$ – ентропія системи після проведення вимірювання ознаки x_j.

Таким чином, величина $I_D(x_j)$ характеризує діагностичну цінність ознаки x_j стосовно системи діагнозів D і ґрунтується на кількості інформації, яка надійшла.

Діагностична цінність простої дихотомічної ознаки, яка набуває одного з двох можливих значень, визначається за формулою:

$$I_{D_i}(x_j) = \log_2 \frac{P(x_j / D_i)}{P(x_j)}, \qquad (1.50)$$

де $I_{Di}(x_j)$ – діагностична вага ознаки x_j для стану D_i; $P(x_j/D_i)$ – апріорна імовірність наявності ознаки при стані D_i; $P(x_j)$ – апріорна імовірність наявності ознаки у всій системі можливих станів D.

Величина $P(x_j/D_i)$ розраховується як відношення кількості об'єктів, у яких є присутньою ознака x_j при стані D_i, до загального числа об'єктів з розглянутим станом. На підставі формули (1.50) можна зробити висновок, що при однаковому значенні ймовірностей наявності ознаки для конкретного стану і для всієї системи станів діагностична вага ознаки дорівнює нулю й ознака не несе ніякої інформативності.

Діагностична вага відсутності простої ознаки визначається за допомогою виразу, що виходить з формули (1.50) шляхом внесення обернених величин ймовірностей:

$$I_{D_i}(\overline{x_j}) = \log_2 \frac{1 - P(x_j / D_i)}{1 - P(x_j)}, \qquad (1.51)$$

Варто враховувати, що діагностична вага ознаки може бути як позитивною, так і негативною величиною, тобто вона може як зменшувати, так і збільшувати імовірність того або іншого стану.

Повна діагностична вага простої ознаки для стану D_i враховує як наявність, так і відсутність ознаки і може бути розрахована за формулою:

$$I_{D_i}(x_j) = P(x_j / D_i) \cdot \log_2 \frac{P(x_j / D_i)}{P(x_j)} + [1 - P(x_j / D_i)] \cdot \log_2 \frac{1 - P(x_j / D_i)}{1 - P(x_j)}. \qquad (1.52)$$

Діагностична цінність простої ознаки для системи станів визначається як

$$I_D(x_j) = \sum_{i=1}^{n} P(D_i) \cdot I_{D_i}(x_j). \qquad (1.53)$$

Оцінка інформативності складних ознак. До складних ознак належать рангові, значення яких можна виразити скінченним числом інтервалів, і числові, котрі теж можна виразити скінченним числом інтервалів, тому що будь-які виміри виконуються з скінченною точністю. Якщо розглядати кожен інтервал (діагностичний розряд) складної ознаки як просту ознаку, то діагностична цінність s-го інтервалу складної ознаки за формулою (1.50) запишеться у вигляді:

$$I_{D_i}(x_{js}) = \log_2 \frac{P(x_{js} / D_i)}{P(x_{js})}, \qquad (1.54)$$

де $P(x_{js} / D_i)$ – апріорна ймовірність *s*-го діагностичного інтервалу складної ознаки для стану D_i.

Величина

$$I_{D_i}(x_j) = \sum_{s=1}^{m} P(x_{js} / D_i) \cdot \log_2 \frac{P(x_{js} / D_i)}{P(x_{js})} \qquad (1.55)$$

визначає діагностичну цінність складної ознаки для діагнозу D_i.

Для визначення повної діагностичної цінності складної ознаки відносно системи станів застосовується формула

$$I_D(x_j) = \sum_{i=1}^{n} \sum_{s=1}^{m} P(D_i) \cdot P(x_{js} / D_i) \cdot \log_2 \frac{P(x_{js} / D_i)}{P(x_{js})}. \qquad (1.56)$$

При визначення інформативності за формулами (1.54–1.56), особливо при великому *m,* крім збільшення складності обчислень, ставляться підвищені вимоги до обсягу і репрезентативності навчальної вибірки, що не завжди може бути досягнуто в реальних базах даних.

Для оцінки інформативності числових ознак за виразами (1.55, 1.56) необхідно виконати розбиття динамічного діапазону ознаки на діагностично-значущі інтервали. Дана задача вирішується на інтуїтивному рівні, причому для зручності обчислень, розбиття виконується на *m* рівномірних інтервалів [18]. У роботі [7] пропонується метод розбиття динамічного діапазону числової ознаки на задане число *m* нерівномірних діагностично-значущих інтервалів на основі аналізу інтегральної помилки аппроксимації теоретичного закону розподілу гістограмою та обмеженістю навчальної вибірки.

Алгоритми визначення групи інформативних ознак. Для вирішення задачі визначення групи інформативних ознак використовуються такі підходи й алгоритми [8]:

- перебір усіх можливих комбінацій ознак;
- метод "*k*" кращих ознак;
- методи послідовного зменшення (DEL або ПЗМП) і збільшення (ADD або ПЗБП) простору ознак;
- узагальнений алгоритм (ADD - DEL або "плюс *l* мінус *r*");
- методи, засновані на критерії максиміну;
- еволюційні алгоритми, зокрема алгоритми випадкового пошуку з

адаптацією;

- метод гілок і границь та інші.

Повний перебір усіх можливих комбінацій ознак. Задача пошуку групи інформативних ознак при побудові алгоритмів розпізнавання формулюється в такий спосіб. Нехай задані вихідна множина ознак $X = \{x_i\}, i = \overline{1, p}$ і деякий критерій якості розпізнавання *J*.

Позначимо через $X_n \in X$ групу з *n* ознак, що має найкраще значення критерію якості рішення задачі розпізнавання $J(X_n)$ у порівнянні з будь-якою іншою групою $X_n^l \in X$. Тобто $J(X_n) = \max_l J(X_n^l)$.

Самий надійний метод пошуку X_n полягає в повному переборі всіх можливих груп ознак. Але кількість таких груп складає C_p^n, і для високих размірностей повний перебір є нереальним. Тому всі розглянуті нижче алгоритми визначення X_n пов'язані зі спробами уникнути повного перебору.

Метод "*k*" кращих ознак. У даному алгоритмі використовується припущення про статистичну незалежність ознак. Вихідні ознаки ранжируються за обраним критерієм якості:

$$J(x_{i_1}) \geq J(x_{i_2}) \geq \ldots \geq J(x_{i_j}) \geq \ldots \geq J(x_{i_p}), \qquad (1.57)$$

і з побудованого ряду відбирається "*k*" перших, найбільш цінних ознак.

Чим суворіше дотримується умова незалежності ознак, що відбираються, тим кращий виходить кінцевий результат. Але умова незалежності рідко виконується в реальних БД, тому для добору і наближеного визначення ваг коррельованих ознак використовують більш складні методи.

Методи послідовного збільшення і зменшення простору ознак (ПЗБП і ПЗМП). Зазначені евристичні алгоритми враховують коррельованість ознак, виконують спрямований перебір і одержують рішення, близьке до оптимального.

У ПЗБП спочатку визначається одна ознака, що має максимальне значення критерію *J*. Потім на кожному кроці пошуку нова група утворюється шляхом додавання однієї ознаки, яка, будучи включеною у розширену групу, дає максимальне значення критерію *J*. Процедура повторюється, поки не буде побудована група з *n* ознак.

Незважаючи на більш витончені операції з експериментальною інформацією у порівнянні з методом "*k*" кращих ознак, метод ПЗБП не гарантує одержання оптимального результату, який може бути досягнутий за допомогою повного перебору всіх можливих комбінацій вихідних ознак.

ПЗМП заснований на послідовному зменшенні групи на одну ознаку.

Спочатку з поточної групи X_k по черзі вилучаються одиночні ознаки. Групи X_{k-1}, що утворилися, перевіряються за критерієм J. Вилученню підлягає ознака, при відсутності якої зменшена група X_{k-1} має максимальне значення критерію J. Процедура повторюється доти, поки не буде побудована група з n ознак. За допомогою зазначеного алгоритму можуть бути отримані більш ефективні результати, ніж для ПЗБП, у випадку порівняно невеликого обсягу групи вихідних ознак. Для високих розмірностей простору вихідних ознак виникають серйозні проблеми оцінки показника якості, тому що вплив окремо узятої ознаки на сумарний ефект стає порівняним з похибкою його виміру.

Алгоритм "плюс l мінус r". Узагальненням ПЗБП і ПЗМП служить метод "плюс l мінус r", що по черзі працює то на додавання ознак, то на вилучення цих ознак. У цьому методі на k-му ступені пошуку група спочатку розширюється на l ознак, з використанням методу ПЗБП, а потім із одержаної групи вилучається r ознак за допомогою методу ПЗМП. Якщо $l < r$, то маємо метод зменшення групи, а якщо $l > r$, – то метод збільшення групи. Хоча алгоритм "плюс l мінус r" дозволяє вилучати або додавати ознаки в поточну групу, врахування їхнього взаємного відношення не виконується. Цей недолік деякою мірою може бути компенсований шляхом додавання і виключення з групи одночасно декількох ознак ("узагальнений алгоритм плюс l мінус r").

Методи, засновані на стратегії максиміну. Особливістю цих методів є використання дуже обмеженого обсягу інформації, наприклад, тільки про індивідуальну $J(x_i)$ і парну ефективність ознак $J(x_i, x_j)$. У даному випадку нова ознака x_j включається в групу, якщо її приєднання до однієї з уже наявних ознак забезпечує максимальне додаткове збільшення критеріїв якості, тобто

$$\Delta J(x_i, x_j) = J(x_i, x_j) - J(x_i) = \max . \qquad (1.58)$$

У той же час включення x_j у групу виправдане, якщо x_j не коррельована з іншими ознаками.

Алгоритм випадкового пошуку з адаптацією. Суть алгоритму полягає в такому. З множини вихідних ознак X випадковим чином вибираються n ознак і генерується серія l груп ознак $X_n^{(l)}$. Потім визначаються величини критерію J для всіх отриманих $X_n^{(l)}$. Група з максимальним значенням критерію заохочується збільшенням імовірності вибору її ознак у наступних шагах алгоритму, а група з найменшою величиною критерію карається відповідним чином. Ця процедура повторюється доти, поки ймовірність вибору інших груп не наблизиться до нуля. Група з максимальною ймовірністю вибору приймається за найбільш цінну групу з n ознак.

Метод гілок і границь. В основі даного методу лежить припущення про монотонність функції критерію, що виражається такою умовою:

$$J(X_p) > J(X_{p-1}) > ... > J(X_n). \quad (1.59)$$

Метод полягає в процедурі зменшення групи, але з можливістю повернення, і дозволяє вилучити з розгляду деякі групи ознак без розрахунків оцінки їхньої ефективності. На кожному k-му кроці аналізу дерева рішень утворюється група ознак k-го рівня, що містить ознаки, обрані з групи $(k+1)$-го рівня. Потім одна з гілок дерева рішень обстежується до останнього n-го рівня. Максимальна величина критерію груп X_n останнього рівня береться за поріг. Якщо при пошуку на k-му рівні недослідженої частини дерева величина критерію $J(X_k)$ виявилася меншою від порога, то подальший пошук серед груп, що утворяться з X_k, не виконується, тому що внаслідок монотонності критерію всі ці групи будуть мати величину критерію, що нижча від порога. Як правило, для отримання оптимальної групи ознак велика кількість гілок дерева обстежується не до останнього рівня, що забезпечує значне скорочення обчислювальних витрат.

У цілому можна відзначити, що багато які зі згаданих методів визначення складу ознак містять евристичну складову. У кожному конкретному випадку важко заздалегідь угадати, який з цих методів приведе до результату, більш близькому до оптимального. Тому на практиці спроби наблизитися до бажаного оптимуму завжди пов'язані з комбінованим застосуванням різних алгоритмів пошуку групи інформативних ознак.

Контрольні питання

1. Математичні методи перетворення простору ознак. Їх призначення й область застосування.
2. Типи вихідних даних і методи їхньої обробки.
3. Як виконується центрування та нормування даних? Поняття масштабу.
4. Які вимоги повинна задовольняти система вихідних діагностичних ознак.
5. Як і для чого формується таблиця експериментальних даних (ТЕД).
6. Міри зв'язку між ознаками.
7. Коефіцієнт кореляції Пірсона, його призначення та метод обчислення.
8. Як визначається значущість коефіцієнта кореляції?
9. Коефіцієнти рангової кореляції Спірмена та Кендалла, Їх призначення та методи обчислення.
10. Коефіцієнт спряженості, його призначення та метод обчислення.
11. Міри близькості (віддаленості) між об'єктами. Вимоги до них.
12. Як обчислюються і де застосовуються евклідова та зважена евклідова відстані, відстані Махаланобіса, Мінковского, Хеммінга?
13. Які вам відомі методи зниження розмірності простору ознак?

14. Які задачі вирішуються за допомогою кластерного аналізу та області його застосування?

15. Які вам відомі методи кластерного аналізу та особливості їх реалізації?

16. Як реалізується метод кореляційних плеяд?

17. Сутність та реалізація методу головних компонент.

18. Факторний аналіз. Його сутність та реалізація.

19. Сутність та реалізація методу контрастних груп.

20. Що таке і як реалізується багатовимірне шкалювання?

21. Перевірка статистичних гіпотез. Основи дисперсійного аналізу.

22. Реалізація однофакторного дисперсійного аналізу.

23. Реалізація двофакторного дисперсійного аналізу.

24. Основи регресійного аналізу. Класифікація регресійних моделей.

25. Синтез регресійних моделей. Метод найменших квадратів.

26. Показники якості регресійних моделей.

27. Методи самоорганізації регресійних моделей.

28. Міра інформативності дихотомічних ознак.

29. Міра інформативності складних ознак.

30. Методи зменшення розміру простору ознак. Повний перебір, *к*-найкращих, ПЗМГ (DEL), ПЗБГ (ADD), (*к*-плюс *l*-мінус).

31. Методи зменшення розміру простору ознак, засновані на стратегії максиміну. Алгоритм випадкового пошуку з адаптацією. Метод гілок і границь.

1.3. Методи синтезу та області застосування вирішальних правил

При формуванні вирішальних правил класифікації моделлю об’єкта діагностики (ОД) є "чорна шухляда", і шукається залежність між формалізованими станами Y і вектором вхідних ознак X, тобто $Y = f(X)$. Існує безліч методів синтезу вирішальних правил класифікації об'єктів, серед яких виділяють такі класи [8, 18]:

- детерміновані методи;
- імовірнісні методи;
- метод послідовного аналізу (метод Вальда);
- методи, засновані на теорії розпізнавання образів;
- логіко-лінгвістичні методи;
- методи, засновані на нечіткій логіці;
- методи на основі штучних нейронних мереж та ін.

1.3.1. Типи вирішальних правил

Детерміновані методи. Застосовуються у випадках наявності детермінованих зв'язків між ознаками і формалізованими станами об'єктів, як правило на етапі попередньої класифікації. Наприклад, в медичній діагностиці існує група діагностичних ознак (патогномонічні синдроми), які однозначно визначальні при деяких захворюваннях, і, навпаки, існує група синдромів, що ніколи не зустрічаються при деяких захворюваннях. Тому використання детермінованих зв'язків виконується для вирішення двох задач:

- однозначна постановка діагнозу за наявності в пацієнта патогномонічного синдрому;
- виключення з подальшого розгляду визначеної групи захворювань за відсутності в пацієнта патогномонічних синдромів зазначеної групи.

Алгоритмічно детерміновані методи реалізуються у вигляді обчислення хеммінгової відстані (на код детермінованих ознак об'єкта діагностики послідовно накладаються коди синдромів діагностуємих станів) або у виді розгалуженого дерева можливих рішень з перевіркою визначених умов у вузлах розгалуження – система детермінованих правил.

Імовірнісні (статистичні) методи. Ці методи засновані на використанні апарата математичної статистики [5, 21, 22]. Вони найчастіше застосовуються у випадках, коли відомі ймовірнісні характеристики класів або коли вони можуть бути визначені за наявною навчальною вибіркою, що звужує область їхнього застосування. Найчастіше імовірнісні методи розрізняються за критерієм розпізнавання. Нижче перераховані основні види критеріїв.

Критерій Байєса. Байєсівський підхід полягає в обчисленні умовних апостеріорних імовірностей. При цьому рішення приймається на підставі порівняння значень цих імовірностей. Якщо об'єкт ω характеризується N ознаками x_i, що набувають значення $x_1 = x_1^0$, $x_2 = x_2^0$, …, $x_N = x_N^0$, то апостеріорна ймовірність віднесення об'єкта до класу Ω_m, $m = \overline{1, M}$ при здійсненні події $a_N = (x_1^0, x_2^0, \ldots, x_N^0)$ обчислюється таким чином:

$$P(\Omega_m / a_N) = \frac{P(\Omega_m) f_m(x_1^0, x_2^0, \ldots, x_N^0)}{\sum_{m=1}^{M} P(\Omega_m) f_m(x_1^0, x_2^0, \ldots, x_N^0)}, \qquad (1.60)$$

де $P(\Omega_m)$ – апріорна ймовірність появи об'єктів класу Ω_m;

$f_m(x_1^0, x_2^0, \ldots, x_N^0)$ – умовна щільність розподілу ймовірностей значень ознак x_i об'єктів класу Ω_m.

Більш детально байєсівський підхід розглядається в 1.3.3.

Мінімаксний критерій. Мінімаксний критерій мінімізує максимально можливе значення середнього ризику. Алгоритм класифікації формулюється в такий спосіб: якщо виміряне значення ознаки x в об'єкті ω дорівнює x^0, то $\omega \in \Omega_1$, якщо $x = x^0 < x_0$, і $\omega \in \Omega_2$, якщо $x = x^0 > x_0$. Граничне значення x_0 визначається зі співвідношення

$$c_1 Q_1(x_0) = c_2 Q_2(x_0), \tag{1.61}$$

де c_1, c_2 – втрати, зв'язані з помилками 1-го і 2-го роду відповідно;

Q_1, Q_2 – умовні імовірності помилок 1-го і 2-го роду відповідно.

Таким чином, «мінімаксна стратегія є байєсівська стратегія для найгірших значень апріорних імовірностей, що дає хоча й обережне, але гарантоване значення середнього ризику».

Критерій Неймана-Пірсона. Суть критерію Неймана-Пірсона полягає в тім, щоб домогтися мінімуму умовної ймовірності помилки 2-го роду Q_2 при заданому значенні умовної імовірності помилки 1-го роду Q_1. Тобто у випадку класифікації об'єктів на два класи (Ω_1 і Ω_2) при $Q_1 \le A$, де A – деяка постійна величина, потрібно визначити рішення x_0, що задовольняє рівнянню

$$\int_{x_0}^{\infty} f_1(x)dx = A, \tag{1.62}$$

де $f_1(x)$ – умовна щільність розподілу ймовірностей значень ознаки x об'єктів класу Ω_1.

Тоді якщо в об'єкта ω виміряне значення ознаки $x = x^0$, то $\omega \in \Omega_1$, якщо $x = x^0 < x_0$, і $\omega \in \Omega_2$, якщо $x = x^0 > x_0$.

Метод послідовного аналізу (метод Вальда) [23]. Використовується для диференціальної діагностики станів D_1 і D_2 і являє собою послідовну процедуру обстежень за допомогою системи простих незалежних ознак (бінарний, якісний або діагностичний інтервал кількісної ознаки), за умовою якої досягається заданий рівень вірогідності стану. Спочатку проводиться обстеження за ознакою x_1, за результатами якого визначається відношення правдоподібності

$$\Theta = \frac{P(x_1 / D_2)}{P(x_1 / D_1)}, \tag{1.63}$$

яке порівнюється з порогами A та B – відповідно верхньою і нижньою границею "області невизначеності", необхідної для ухвалення рішення.

Якщо $\Theta > A$, то робиться висновок на користь стану D_2, у противному разі, якщо $\Theta < B$, – на користь стану D_1. Якщо не досягнутий жоден поріг, тобто $B < \Theta < A$, то для ухвалення рішення проводиться додаткове обстеження за ознакою x_2 і т. д.

Для реалізації методу необхідно ранжувати ознаки за критерієм їхньої діагностичної цінності, а крім того, необхідно використовувати систему незалежних ознак.

1.3.2. Методи, засновані на теорії розпізнавання образів

У даній групі методів результати вимірювання характеристик об'єктів представляються точками в просторі діагностичних ознак. При цьому різні класи повинні утворювати компактні множини в просторі ознак. Діагностика нового об'єкта зводиться до обчислення міри близькості до кожного класу. Розрізняють такі групи методів [24]:

1) Методи, засновані на принципі поділу (*R-моделі*). Основою зазначених методів є формування поділяючої поверхні або групи поверхонь, що щонайкраще розділяють елементи різних класів. До *R-моделей* належать:

- дискримінантний аналіз;
- метод порівняння з прототипом (еталоном);
- метод К-найближчих сусідів;
- метод січних площин;
- метод узагальненого портрета.

2) Методи, побудовані на основі "потенційних функцій" (*П-моделі*) Ці методи базуються на запозиченій з фізики ідеї електричного потенціалу, що визначений у будь-якій точці простору і змінюється в міру віддалення від заряду.

3) Методи обчислення оцінок (голосування) (*Г-моделі*) [28]. В основу цих методів покладено принцип часткової прецедентності, тобто прийняття рішень за аналогією.

4) Логіко-лінгвістичні методи (*Л-моделі*) – засновані на обчисленні висловлень, зокрема, на апараті алгебри логіки.

5) Методи, засновані на нечіткій логіці, на основі штучних нейронних мереж та інші.

В наступних підпунктах більш детально розглядаються відмічені методи.

Дискримінантний аналіз [29]. Якщо критеріальний показник y вимірюється у номінальній шкалі або якщо зв'язок цього показника з вихідними ознаками X є нелінійним й носить невідомий характер, то для визначення параметрів діагностичної моделі використовуються методи дискримінантного аналізу. У цьому випадку результати обстеження

розбиваються на групи (класи), а ефективність діагностичної моделі розглядається під кутом зору її здатності розділяти (дискримінувати) класи, що діагностуються.

Показником якості дискримінантного аналізу є ймовірність помилкової класифікації досліджуваних об'єктів P_ε, яка мінімізується. У свою чергу, для розкриття взаємозв'язку P_ε зі структурою експериментальних даних у дискримінантному аналізі широко використовуються геометричні знання про поділ діагностуємих класів у просторі ознак.

При цьому вважається, що сукупність об'єктів, які належать до одного класу ω_i, утворює «хмару» у p-вимірному просторі R_p, що задається вихідними ознаками. Для успішної класифікації необхідно:

а) щоб хмара з ω_i в основному була сконцентрована в деякій області D_i простору R_p;

б) щоб в область D_i потрапила незначна частина «хмар» об'єктів, що відповідають іншим класам.

Тоді побудова вирішального правила розглядається як задача пошуку K непересічних областей D_i ($i = \overline{1,K}$), що задовольняють умови а і б. Дискримінантні функції (ДФ) дають визначення цих областей шляхом завдання їхніх границь у багатовимірному просторі R_p. Якщо об'єкт x попадає в область D_i, то приймається рішення про належність об'єкта до ω_i. Тоді критерієм правильного визначення областей буде

$$Q = \sum_{i=1}^{K-1}\sum_{j>i}^{K} P(\omega_i)P(\omega_i / \omega_j), \qquad (1.64)$$

де $P(\omega_i)$ – апріорна ймовірність появи об'єкта з класу ω_i в області D_i;

$P(\omega_i/\omega_j)$ – ймовірність того, що об'єкт із класу ω_j помилково попадає в область D_i, що відповідає класу ω_i.

Критерій Q називається критерієм середньої ймовірності помилкової класифікації. Мінімум Q досягається при використанні, зокрема, розглянутого вище байєсівського підходу, що, однак, може бути практично реалізований тільки при справедливості дуже сильного припущення про незалежність вихідних ознак, і в цьому випадку дає оптимальну лінійну діагностичну модель. Велика кількість інших підходів також використовує лінійні дискримінантні функції, але при цьому на структуру даних накладаються менш тверді обмеження.

Для випадку двох класів (ω_1 і ω_2) методи побудови лінійної дискримінантної функції (ЛДФ) спираються на два припущення. Перше полягає в тому, що області D_1 і D_2, у яких концентруються об'єкти з діагностуємих класів ω_1 і ω_2, можуть бути розділені (p-1)-вимірною

гіперплощиною – дискримінантною функцією (ДФ):

$$y(X) = \sum_{i=0}^{p} a_i x_i \,, \tag{1.65}$$

де p – кількість дискримінантних змінних; x_i – значення незалежних змінних; a_i – коефіцієнти (константи), що оцінюються за допомогою ДА.

Коефіцієнти a_i у цьому випадку інтерпретуються як параметри, що характеризують нахил гіперплощини до координатних осей, а a_0 називається порогом і відповідає відстані від гіперплощини до початку координат. Переважне розташування об'єктів одного класу, наприклад ω_1, по одну сторону гіперплощини виражається в тому, що для них, здебільшого, буде виконуватися умова $y(X)<0$, а для об'єктів іншого класу ω_2 – обернена умова $y(X)>0$. Друге припущення стосується критерію якості поділу областей D_1 і D_2 гіперплощиною (1.65). Найчастіше вважають, що поділ буде тим кращий, чим далі відстоятимуть одне від одного середні значення випадкових величин:

$$m_1 = E\{y(X)\}, \quad y \in \omega_1 \text{ і } m_2 = E\{y(X)\}, \; y \in \omega_2,$$

де $E\{\ \}$ – оператор усереднення.

У найпростішому випадку вважають, що класи ω_1 і ω_2 мають однакові коваріаційні матриці $S1=S2=S$.

Лінійний дискримінантний аналіз не тільки виявляє лінійні комбінації змінних для найкращого поділу заданих груп об'єктів, але і може використовуватися для зниження розмірності вхідних даних.

Як було зазначено раніше, результатом ДА є побудова дискримінантної функції виду (1.65).

Задачею дискриминантного аналізу є визначення таких коефіцієнтів a_i, щоб за значеннями ДФ можна було з мінімальною помилкою провести поділ усієї сукупності на класи. Якщо значення x_i нормовані, то чим більше значення коефіцієнтів a_i, тим більший внесок відповідної змінної в дискримінацію сукупності. Якщо класів більш двох, то будується декілька ДФ. Процедура ДА може проводитися двома методами:

– методом одночасного врахування змінних;

– покроковим методом.

Покроковий метод є одним із способів виключення зайвих змінних і полягає у використанні процедури послідовного добору найбільш корисних дискримінантних змінних, хоча отримана множина дискримінантних змінних може і не бути найкращою їх комбінацією.

У результаті застосування методу одночасного врахування змінних за наявності великої кількості змінних частина з них може виявитися зайвими,

тобто такими, що не роблять ніякого внеску в аналіз, а лише ускладнюють його, деякі з них можуть нести однакову інформацію, будучи при цьому хорошими дискримінаторами.

Щоб вибрати найбільш оптимальний за критерієм якості (Λ-статистика Уілкса) класифікації вид ДФ, дискримінантний аналіз необхідно проводити кожним з перерахованих вище методів.

Як міра розходження між дискримінантними змінними застосовується Λ-статистика Уілкса. Її значення обчислюється за формулою

$$\Lambda = \prod_{i=k+1}^{k_0} \frac{1}{1+\lambda_i}, \tag{1.66}$$

де k_0 – кількість станів (класів); k – кількість вже обчислених ДФ; λ_i – власне число i-ї ДФ, кожне з яких визначається при розв'язанні системи рівнянь

$$\begin{cases} \sum_{i=1}^{r} b_{1i}\nu_i = \lambda \sum_{i=1}^{r} w_{1i}\nu_i \\ \quad \ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{r} b_{ni}\nu_i = \lambda \sum_{i=1}^{r} w_{ni}\nu_i \end{cases}, \tag{1.67}$$

де r – кількість дискримінантних змінних; b_{ij} ($i,j = \overline{1,r}$) – елементи матриці $\boldsymbol{B}$ міжгрупової суми квадратів відхилень і попарних добутків; w_{ij} ($i,j = \overline{1,r}$) – елементи матриці $\boldsymbol{W}$ розкиду усередині класу; ν_i ($i = \overline{1,r}$) – послідовність коефіцієнтів, яка використовується для побудови ДФ, причому $\sum_{i=1}^{r} \nu_i = 1$.

Побудова ДФ зводиться до розв'язання системи рівнянь відносно λ і послідовності ν_i (кількість отриманих розв'язків відповідає кількості побудованих ДФ). Коефіцієнти кожної ДФ визначаються за формулою

$$a_i = \nu_i \sqrt{m - k_0}, \tag{1.68}$$

де m – загальне число спостережень по всіх класах.

Величина Λ є оберненою мірою, тобто її значення, близьке до нуля, говорить про високе розрізнення класів – центри класів добре розділені і сильно відрізняються один від одного стосовно ступеня розкиду усередині класів. Збільшення величини Λ до її максимального значення, що дорівнює 1, приводить до поступового погіршення розрізнення, тому що центри класів при цьому збігаються.

На основі Λ-статистики Уілкса одержують тест значущості, апроксимуючи розподіл деякої функції від неї розподілом χ^2. Для того щоб

визначити рівень значущості цього розподілу, користуються спеціальними таблицями.

Метод порівняння з прототипом (еталоном). Даний метод найчастіше використовується, коли класи Ω_m ($m = \overline{1, M}$) утворюють компактні множини об'єктів, що мають сферичну форму в просторі ознак. У цьому випадку кожний із класів Ω_m описується прототипом або еталоном $\omega^{m\text{э}}$, у якості якого вибирається геометричний центр угруповання класу.

Невідомий об'єкт ω належить до класу Ω_t, відстань до прототипу якого $R(\omega, \omega^{t\text{э}})$ буде мінімальною:

$$R(\omega, \omega^{t\text{э}}) = \min_{m=\overline{1,M}} R(\omega, \omega^{m\text{э}}), \qquad (1.69)$$

де $R(\omega, \omega^{m\text{э}})$ – відстань між об'єктом ω та еталоном $\omega^{m\text{э}}$ класу Ω_m;
M – кількість класів.

Метод Фікса–Ходжеса (метод "найближчих сусідів") [25]. Даний метод використовується в тому випадку, коли структура класів досить складна, далека від сферичної, або взагалі невідома, але підтверджується гіпотеза про безперервність багатовимірної щільності розподілу в кожній локальній області простору ознак. Суть даного методу полягає у визначенні деякого заданого числа k найближчих до невідомого об'єкта (в обраній метриці) об'єктів навчальної вибірки ("найближчих сусідів"). Невідомий об'єкт ω належить до того класу, число представників якого переважає серед обраних k "найближчих сусідів".

Реалізація методу. Визначають відстані $R(\omega, \omega_i)$ ($\text{i} = \overline{1, N}$), де N – об'єм навчальної вибірки, між невідомим об'єктом ω та усіма об'єктами навчальної вибірки, які являють собою сукупність всіх класів Ω_m ($m = \overline{1, M}$). Після цього знайдені відстані ранжуються за зростанням, вибираються перші k елементів – "найближчі сусіди", серед яких визначається клас Ω_t, число представників якого переважає серед обраних k "найближчих сусідів". До класу Ω_t і належить невідомий об'єкт ω.

Метод січних площин. Даний алгоритм полягає в апроксимації поділяючої поверхні «шматками» гіперплощин. Для формування поділяючої гіперповерхні необхідно: провести січні гіперплощини; виключити зайві гіперплощини; виключити зайві шматки гіперплощин.

Метод узагальненого портрета [26]. Даний метод дозволяє побудувати оптимальну нормально орієнтовану гіперплощину, що розділяє множини векторів Ω_1 і Ω_2. Таке розбиття має місце, якщо для $k<1$ існує вектор ψ, для якого виконуються нерівності

$$\begin{aligned}(\omega,\psi)\geq 1,\ \forall\omega\in\Omega_1,\\ (\omega,\psi)\leq k,\ \forall\omega\in\Omega_2.\end{aligned} \tag{1.70}$$

Кожному значенню ψ, що задовольняє вираз (1.70) ставиться у відповідність гіперплощина $(\omega,\psi)=\dfrac{1+k}{2}$. Мінімальний по модулі вектор ψ, що задовольняє нерівностям (1.70), називається узагальненим портретом множини Ω_1 відносно Ω_2.

Метод потенційних функцій [27]. В якості потенційної функції, що характеризує належність об'єкта відповідному класові, використовується усюди позитивна і монотонно спадна функція відстані, аналогічна за формою електричному потенціалові φ. Прикладами таких функцій можуть бути

$$\varphi(R)=\left|\frac{\sin(\alpha R^2)}{\alpha R^2}\right|,\ \ \varphi(R)=e^{-\alpha R^2}\ \text{ або }\ \varphi(R)=\frac{1}{1+\alpha R^2}, \tag{1.71}$$

де R – визначена будь-яким чином відстань між точкою-джерелом і точкою-приймачем, у якій обчислюється потенціал; $\alpha>0$ – ваговий коефіцієнт, який характеризує швидкість убування потенціалу φ.

Точками-джерелами потенціалу виступають об’єкти класу Ω_m, а точкою-приймачем – об’єкт ω, який підлягає класифікації. Об'єкт ω належить до класу Ω_t, сумарний потенціал якого буде максимальним.

Методи обчислення оцінок (голосування) [28]. Як було відмічено раніше, в основу цих методів покладено принцип часткової прецедентності, тобто прийняття рішень за аналогією. Між частинами описів розпізнаваного й еталонного об'єктів проводиться аналіз "близькості", наявність якої служить частковим прецедентом і оцінюється за деяким заданим правилом (за допомогою числової оцінки). Загальна оцінка розпізнаваного класу, отримана за набором оцінок близькості, і є значенням функції належності об'єкта класові. При цьому в алгоритмах розпізнавання, заснованих на обчисленні оцінок, ступінь подібності об'єктів обчислюється шляхом зіставлення всіх

можливих (або визначених) сполучень ознак, що входять в опис об'єктів.

Логіко-лінгвістичні методи. Засновані на обчисленні висловлень, зокрема на апараті алгебри логіки. Тут як класи та ознаки об'єктів виступають логічні змінні. Цей клас методів можна розбити на два підкласи: логічні і лінгвістичні (структурні) [30, 31]. Логічні методи використовуються в тих випадках, коли є дані лише про детерміновані логічні зв'язки між об'єктами і їх ознаками, при цьому апріорна інформація про кількісний розподіл об'єктів по просторовим, часовим, ваговим, енергетичним або будь-яким іншим інтервалам у просторі ознак невідома. Ці відомості представляються у виді булевих співвідношень, що відображають причинно-наслідкові зв'язки між розглянутими класами об'єктів і їхніми ознаками. Належність розпізнаваного об'єкта до одного з класів визначається на підставі розв'язку булевих рівнянь.

Лінгвістичні або структурні методи побудовані на теорії формальних граматик. При цьому складні об'єкти описуються множиною незвідних (атомарних) елементів і набором граматичних правил. Рішення про належність розпізнаваного об'єкта приймається на підставі синтаксичного аналізу або граматичного розбору.

Методи, засновані на нечіткій логіці. Бурхливо розвивається в останнє десятиліття група методів, у яких для класифікації використовується теорія нечітких множин [32], і на її основі – нечіткі правила. Належність елемента x до нечіткої множини M задається безперервною функцією належності $\mu_M(x)$ ($0 \le \mu_M(x) \le 1$). Для кількісної оцінки правил, що близькі до речень природної мови, використовуються лінгвістичні змінні L, можливі значення яких є множиною термів $T(L)$, де кожен терм являє собою мітку нечіткої множини і задається своєю функцією належності $\mu_{Ti}(x)$.

У такий спосіб виконується перехід від числової змінної x до лінгвістичної L. Визначено логічні операції з нечіткими множинами і формалізовані правила нечіткого висновку, що використовуються для формування вирішальних правил. Вид функції належності $\mu_M(x)$ задається дослідником (**S**, **П**, трикутна, трапецеїдальна та ін.), а її параметри визначаються на навчальній вибірці за критерієм мінімуму помилки класифікації, для чого широке застосування одержали генетичні алгоритми [33]. Оскільки в основі методу лежать експертні оцінки, то він є серйозною альтернативою ймовірнісних методів при недостатньому обсязі навчальної вибірки або при її відсутності.

Методи розпізнавання на основі штучних нейронних мереж (ШНМ). Утворюють велику групу перспективних методів розпізнавання в умовах зашумлених і часто суперечливих даних [34]. Однак використання більшості

ШНМ у системах підтримки прийняття рішень, які, як правило, повинні уточнювати свої знання в процесі експлуатації, утруднене існуючими методами навчання ШНМ. Наприклад, багатошарові ШНМ, що добре зарекомендували себе при вирішенні різноманітних задач розпізнавання і прогнозування, навчаються трудомісткими алгоритмами методу зворотного поширення помилки, застосування яких припускає наявність і використання всієї інформації про всі класи. Поява навіть одного нового класу в загальному випадку припускає повне перенавчання мережі. Подібна ситуація характерна і при використанні інших ШНМ, наприклад мережі Хопфілда, двохнаправленої асоціативної пам'яті, мережі Кохонена і т. д. Деякі з ШНМ не вимагають повного і трудомісткого перенавчання, наприклад мережа Хеммінга, однак ця мережа не може самостійно виділяти нову інформацію і самонавчатися.

Неможливість за допомогою відомих ШНМ вирішити проблему стабільності, тобто збереження раніше отриманих знань при запам'ятовуванні нової інформації, і проблему пластичності до нової інформації, тобто сприйняття нової інформації, що дозволяє як модифікувати й уточнювати збережені в пам'яті образи, так і створювати нові, зумовила необхідність розробки принципово нових ШНМ – мереж адаптивної резонансної теорії АРТ (ART – adaptive resonance theory). Ці мережі якоюсь мірою дозволяють вирішувати суперечливі задачі чутливості до нових даних і збереження раніше отриманих знань.

Нейронна мережа АРТ відносить вхідний об'єкт до одного з відомих класів, якщо він подібний або резонує з прототипом цього класу. Якщо знайдений прототип із визначеною точністю, що задається спеціальним параметром подібності, відповідає вхідному об'єктові, то він модифікується, щоб стати більш схожим на пред'явлений об'єкт. Коли вхідний об'єкт недостатньо подібний до жодного з наявних прототипів, то на його основі створюється новий клас, не спотворюючи характеристик існуючих класів.

Слід зазначити, що типологія методів, розглянута вище, є в деякому сенсі умовною. Існує ряд алгоритмів розпізнавання образів, які не можна однозначно віднести до тієї або іншої групи методів, описаних вище. Алгоритми персептронного типу, що реалізуються в різних архітектурах ШНМ можна віднести до *R*-моделей, тому що розбиття об'єктів на класи виконується шляхом формування поділяючої гіперповерхні зі шматків гіперплощин. З іншого боку, в ряді джерел показаний зв'язок алгоритмів персептронного типу з потенційними функціями.

На підставі вище викладеного можна зробити висновок, що для тих або інших методів і алгоритмів синтезу вирішальних правил є цілком конкретна область застосування, визначена апріорними обмеженнями. Як такі обмеження можуть розглядатися: характеристики ознак, у просторі яких потрібно відрізняти об'єкти один від одного; клас вирішальних функцій;

ймовірнісні характеристики класів і т. д. До того ж багато методів досить критичні до обсягу та репрезентативності навчальної вибірки, тому в системах підтримки прийняття рішень використовується розробка нових, модифікація і комбінація відомих методів.

1.3.3. Вирішальні правила в умовах суттєвої апріорної невизначеності

Байєсівський метод. Байєсівський підхід базується на припущенні, що завдання сформульоване в термінах теорії ймовірностей і відомі всі необхідні величини: апріорні ймовірності $P(\omega_i)$ для класів ω_i ($i=\overline{1,K}$) і умовні щільності розподілу значень вектора ознак $P(X/\omega_i)$ в кожному з класів. Правило Байєса полягає в знаходженні апостеріорної ймовірності $P(\omega_i/X)$, що обчислюється в такий спосіб:

$$P(\omega_i / X) = \frac{P(X / \omega_i)P(\omega_i)}{P(X)}, \tag{1.72}$$

де $P(X)$ – умовна щільність розподілу значень вектора ознак по всій вибірці, яка обчислюється за формулою

$$P(X) = \sum_{j=1}^{K} P(X / \omega_j)P(\omega_j). \tag{1.73}$$

Рішення про належність об'єкта A_0 з вектором ознак X_0 до класу ω_{ji} приймається при виконанні умови, яка забезпечує максимум апостеріорної ймовірності $P(\omega_i/X)$ і відповідно – мінімум середньої ймовірності помилки класифікації:

$$P(\omega_j / X_0) = \max_{i=1,K} P(\omega_i / X_0). \tag{1.74}$$

Якщо розглядаються два діагностичних класи (ω_1 і ω_2), то відповідно до цього правила приймається рішення ω_1 при $P(\omega_1/X)>P(\omega_2/X)$ і ω_2 при $P(\omega_2/X)>P(\omega_1/X)$. Величину $P(\omega_i/X)$ у правилі Байєса часто називають правдоподібністю ω_i при даному X і прийняття рішення здійснюється через відношення правдоподібності або через його логарифм

$$L(X) = \log_2 \frac{P(\omega_1 / X)}{P(\omega_2 / X)}. \tag{1.75}$$

Для дихотомічних ознак, з якими в багатьох випадках доводиться мати справу, p-вимірний вектор ознак X може набувати одного з $n=2^p$ дискретних

значень v_1, ..., v_n. Функція щільності $P(X/\omega_i)$ стає сингулярною й заміняється на $P(v_k/\omega_i)$ – умовну ймовірність того, що $X = v_k$ за умови належності об'єктів до класу ω_i. На практиці в дискретному випадку, як і в безперервному, коли число вихідних ознак x_i велике, визначення умовних імовірностей зустрічає значні труднощі й найчастіше не може бути здійснене. З одного боку, це пов'язане, з нереальністю навіть простого перегляду всіх точок дискретного простору дихотомічних ознак. З іншого боку, навіть при набагато меншій кількості ознак для достовірної оцінки умовних імовірностей необхідно мати результати вимірювання досить великої кількості об'єктів.

Розповсюдженим способом подолання зазначених труднощів служить модель, в основі якої лежить припущення про незалежність вихідних дихотомічних ознак. Нехай для визначеності компонента вектора X набувають значення 1 або 0. Позначимо: $p_i = P(x_i = 1/\omega_1)$ – ймовірність того, що ознака x_i дорівнює 1 за умови належності об'єктів до діагностичного класу ω_1, і $q_i = P(x_i = 1/\omega_2)$ – ймовірність рівності 1 ознаки x_i у класі ω_2. У випадку $p_i > q_i$ варто очікувати, що i-та ознака буде частіше набувати значення 1 у класі ω_1, ніж в ω_2. У припущенні про незалежність ознак можна представити $P(X/\omega_i)$ у вигляді добутку ймовірностей:

$$P(X/\omega_1) = \prod_{i=1}^{p} p_i^{x_i}(1-p_i)^{1-x_i};$$
$$P(X/\omega_2) = \prod_{i=1}^{p} q_i^{x_i}(1-q_i)^{1-x_i}. \qquad (1.76)$$

Логарифм відношення правдоподібності в цьому випадку визначається в такий спосіб:

$$L(X) = \sum_{i=1}^{p}\left[x_i \log_2 \frac{p_i}{q_i} + (1-x_i)\log_2 \frac{1-p_i}{1-q_i}\right] + \log_2 \frac{P(\omega_1)}{P(\omega_2)}. \qquad (1.77)$$

З виразу (1.77) видно, що дане рівняння лінійно щодо ознак x_i. Тому можна записати:

$$L(X) = \sum_{i=1}^{p} w_i x_i + w_0, \qquad (1.78)$$

де вагові коефіцієнти w_i обчислюються за формулою

$$w_i = \log_2 \frac{p_1(1-q_i)}{q_1(1-p_i)},$$

а величина порога w_0 – за формулою

$$w_0 = \sum_{i=1}^{p} \log_2 \frac{1-p_i}{1-q_i} + \log_2 \frac{P(\omega_1)}{P(\omega_2)}.$$

Якщо $L(X)>0$, то приймається рішення про належність об'єкта до діагностичного класу ω_1, а якщо $L(X)<0$, то до класу ω_2.

Теорія Демпстера–Шефера та реалізація вирішальних правил на її основі [35]. У даному розділі розглядається один з методів формування нестрогих обчислень, який називається теорією Демпстера–Шефера, або теорією Шефера–Демпстера. Цей метод заснований на роботі Демпстера, який спробував змоделювати невизначеність, задаючи ряд ймовірностей, а не окреме ймовірнісне значення. Шефер доповнив і уточнив результати, отримані Демпстером, у своїй книзі *A Mathematical Theory of Evidence,* опублікованій в 1976 році. Теорія Демпстера–Шефера має добрий теоретичний фундамент. До того ж можна показати, що теорія коефіцієнтів достовірності (див. далі) є частковим випадком теорії Демпстера–Шефера, що дозволяє перевести методи, засновані на використанні коефіцієнтів достовірності на теоретичну основу.

Теорія Демпстера–Шефера заснована на припущенні про те, що задано фіксовану множину взаємовиключних і вичерпних елементів, яка називається *середовищем* і позначається грецькою буквою Θ:

$$\Theta = \{\Theta_1, \Theta_2, \ldots, \Theta_n\}.$$

Термін "середовище" аналогічний терміну "універсум", що застосовується в теорії множин. Таким чином, середовище – це множина об'єктів, які становлять для нас інтерес. Нижче наведено деякі приклади визначення варіантів середовища:

Θ = {авіалайнер, бомбардувальник, винищувач}

Припустимо, що всі можливі елементи універсуму знаходяться в заданій множині, а отже, множина є вичерпною. Крім того, щоб спростити наведений тут опис, припустимо, що множина Θ є скінченою, хоча є варіанти середовища Демпстера–Шефера, елементами яких є безперервні змінні, такі як час, відстань, швидкість і т. д.

Один зі способів міркування про структуру множини полягає в тому, що розглядаються питання і відповіді, які відносяться до цієї множини. Припустимо, що дана наведена нижче множина, а питання, що стосується елементів цієї множини, сформульоване так: "Які з цих літаків мають військове призначення?"

Θ = {авіалайнер, бомбардувальник, винищувач}.

Відповіддю стає наступна підмножина множини Θ:

$$\{\Theta_2, \Theta_3\} = \{\text{бомбардувальник, винищувач}\}.$$

Таким чином, кожна підмножина множини Θ може інтерпретуватися як можлива відповідь на деяке питання. А оскільки всі елементи є взаємовиключними і середовище – вичерпним, то правильною відповіддю на будь-яке питання може служити тільки одна підмножина.

Усі можливі підмножини в середовищі представлення типів літаків наведено на рис. 1.5.

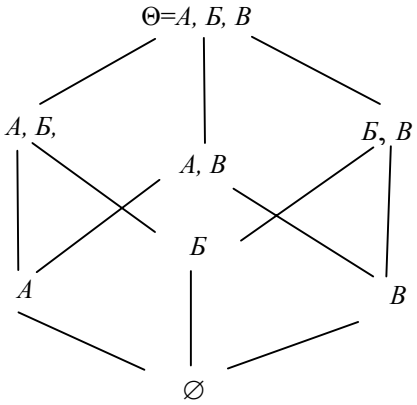

Рис. 1.5. Усі підмножини середовища, у якому розглядаються типи літаків

Лінії на цьому рисунку показують відношення між підмножинами. Для позначення елементів (авіалайнер, бомбардувальник і винищувач) використовуються букви *А, Б, В*. Ця діаграма зображена у вигляді ієрархічної решітки, у якій множина Θ знаходиться вгорі, а порожня множина $\varnothing = \{\ \}$ – унизу.

Один зі шляхів від вузла Θ до вузла $\varnothing$ виражає ієрархічне відношення підмножин, що з'єднує батьківські підмножини з дочірніми:

$$\varnothing \subset \{А\} \subset \{АБ\} \subset \{АБВ\}\ .$$

Якщо всі елементи середовища можуть інтерпретуватися як можливі відповіді і якщо тільки одна відповідь є правильною, то середовище називається *рамками розрізнення* (frame of discernment). У даному випадку термін "розрізнення" показує, що існує можливість відрізнити одну правильну відповідь від всіх інших можливих відповідей на питання. Множина з кількістю елементів N має точно 2^N підмножини, включаючи саму себе. Усі ці підмножини визначають степеневу множину $P(\Theta)$. Таким

чином, для середовища, що представляє типи літаків, маємо:

$$P(\Theta) = \{\varnothing, \{A\}, \{Б\}, \{B\}, \{A, Б\}, \{A, B\}, \{Б, B\}, \{A, Б, B\}\}.$$

Степенева множина середовища включає як свої елементи усі відповіді на всі можливі питання з рамок розрізнення. Це означає, що існує взаємно однозначна відповідність між елементами $P(\Theta)$ і підмножинами множини Θ.

Масові функції і незнання. Згідно з байєсівською теорєю, апостеріорна імовірність при одержанні нових свідчень змінюється. Аналогічним чином у теорії Демпстера–Шефера може змінюватися ступінь довіри до свідчення. Крім того, у теорії Демпстера–Шефера прийнято розглядати ступінь довіри до свідчення аналогічно *масі* фізичного об'єкта, що позначається буквою *m,* яку можна переміщати, дробити і комбінувати.

Фундаментальне розходження між теорією Демпстера–Шефера і теорією ймовірностей полягає в тому, що в цих теоріях поняття *незнання* трактується по-різному. Відповідно до теорії ймовірностей за відсутності апріорних знань, ймовірність P кожного можливого випадку визначається формулою $P=1/N$, де N – кількість можливих випадків. Таке присвоювання значення P застосовується "у безвихідній ситуації" на підставі принципу байдужості. Крайній випадок застосування принципу байдужості виникає, якщо є тільки два можливих випадки, таких, як наявність або відсутність випадкової дії (наприклад, наявність нафти), що можна позначити символічно як H і H'. У випадках, подібних до цього, P = 50 %, навіть якщо взагалі немає ніяких знань про те, чи є нафта чи ні, оскільки теорія ймовірностей говорить, що справедлива наступна формула:

$$P(H) + P(H') = 1. \qquad (1.78)$$

Таким чином, усе, що не обґрунтовує гіпотезу, повинне її спростовувати, оскільки можливість незнання не припускається.

Якщо подібний підхід застосовується без міркувань, то можуть виявлятися деякі безглузді наслідки. Наприклад, припустимо, що людина міркує, чи є родовище нафти під її будинком чи його нема. Відповідно до принципу байдужості, якщо цілком відсутні які-небудь інші знання, то ймовірність наявності родовища нафти під будинком дорівнює 50 %. Варто тільки про це подумати, можна дійти висновку, що такі шанси наявності нафти є досить вражаючими і дають набагато кращу можливість швидко розбагатіти, чим за допомогою будь-яких інших законних капіталовкладень. А оскільки є 50 % шансів знайти нафту, то чи не варто негайно зняти усі свої заощадження, найняти бурову установку і приступити до буріння свердловини на кухні?!

Але навіть якщо принцип байдужості не використовується, наступне обмеження примусово диктує необхідність присвоювання ймовірності

заперечення гіпотези і при відсутності свідчення, що відноситься до заперечення на основі формули (1.78).

Таким чином, теорія ймовірностей вимагає, щоб свідчення, що не обґрунтовує гіпотезу, спростовувало її. З іншого боку, теорія Демпстера–Шефера не змушує призначати ступінь довіри незнанню або спростуванню гіпотези. Замість цього маса привласнюється тільки тим підмножинам середовища, яким бажано призначити деякий ступінь довіри. Весь ступінь довіри, не привласнений конкретній підмножині, розглядається як ступінь відсутності довіри (no belief), або як ступінь недостачі довіри (nonbelief), і зв'язується із середовищем Θ. А ступінь довіри, що спростовує гіпотезу, являє собою *ступінь недовіри* (disbelief), який не слід плутати зі ступенем відсутності довіри.

Наприклад, припустимо, що деякий датчик, такий, як датчик системи впізнання "свій-чужий", не одержує відповіді від радіомаяка-відповідача літака. Робота системи впізнання "свій-чужий" заснована на використанні радіопередавача/приймача, що передає радіограму на літак і приймає відповідь. Якщо літак належить до власного повітряного флоту (є "своїм"), його радіомаяк-відповідач повинен відреагувати на радіограму, відправивши у відповідь ідентифікаційний код. Літаки, що не відповідають, за замовчуванням розглядаються як "чужі". Але літак може не відповісти на сигнали системи впізнання "свій-чужий" з багатьох причин, зокрема, з таких:

- несправність у системі впізнання "свій-чужий";
- несправність у радіомаяку-відповідачі літака;
- відсутність на літаку системи впізнання "свій-чужий";
- зникнення сигналу впізнання "свій-чужий" у перешкодах;
- одержання наказу дотримуватися режиму радіомовчання.

Припустимо, що невдала спроба системи впізнання "свій-чужий" одержати відповідь указує на наявність ступеня довіри 0,7 до свідчення, що розглянутий літак є "чужим", причому як "чужі" літаки розглядаються тільки бомбардувальники і винищувачі. Таким чином, присвоювання маси підмножині $\{Б, В\}$ здійснюється за наступною формулою, у якій m_1 позначає перше свідчення датчика впізнання "свій-чужий":

$$m_1(\{Б, В\})=0,7.$$

Інша частина ступеня довіри привласнюється середовищу Θ, як ступінь відсутності довіри:

$$m_1(\Theta) = 1 - 0,7 = 0,3.$$

Ми довіряємо гіпотезі, що розглянутий літак є "чужим", у ступені 0,7 і резервуємо судження, що відповідає ступені 0,3, за недовірою і додатковою довірою до гіпотези, що літак "чужий".

Кожна підмножина в степеневій множині середовища, що має масу більше 0, розглядається як *фокальний елемент*, у якому фокусується (або концентрується) доступне свідчення.

У цьому теорія Демпстера–Шефера істотно відрізняється від теорії ймовірностей, у якій було б прийняте таке припущення:

$$P(\text{hostile}) = 0{,}7;\ P(\text{non-hostile}) = 1 - 0{,}7 = 0{,}3.$$

Як показано в табл. 1.5, застосування поняття маси забезпечує набагато більшу свободу вибору в порівнянні з поняттям ймовірності.

Таблиця 1.5 –Порівняння можливостей теорії Демпстера–Шефера і теорії ймовірностей

Теорія Демпстера–Шефера	Теорія ймовірностей
Значення $m(\Theta)$ не обов'язково повинне дорівнювати 1	$\sum_i P_i = 1$
Якщо $X \subseteq Y$, то вимога про дотримання рівності $m(X) = m(Y)$ не є обов'язковою	$P(X) \le P(Y)$
Не потрібна наявність зв'язку між $m(X)$ і $m(X')$	$P(X) + P(X') = 1$

Кожну масу можна формально представити за допомогою функції, що відображає кожен елемент степеневої множини в дійсне число, що знаходиться в інтервалі від 0 до 1. Зазначене відображення формально представляється за допомогою наступної формули:

$$m : P(\Theta) \to [0,1].$$

Відповідно до прийнятої угоди маса пустої множини зазвичай визначається як така, що дорівнює нулю:

$$m(\varnothing) = 0,$$

а сума мас усіх підмножин X степеневої множини дорівнює одиниці:

$$\sum_{X \in P(\theta)} m(X) = 1.$$

Комбінування свідчень. Тепер розглянемо випадок, у якому стають доступними додаткові свідчення. При цьому хотілося б мати можливість комбінувати усі свідчення, щоб зробити кращу оцінку ступеня довіри до свідчення. Для ознайомлення з тим, як це робиться, спочатку розглянемо приклад, у якому застосовується окремий випадок загальної формули комбінування свідчень.

Припустимо, що для розпізнавання літаків застосований датчик другого типу, який визначає розглянутий літак як бомбардувальник, зі ступенем довіри до свідчення, що дорівнює 0,9. Тепер маси свідчень, отриманих від обох датчиків, набувають такого вигляду:

$$m_1(\{Б, В\})=0,7; \qquad m_1(\Theta)=0,3;$$

$$m_2(Б)=0,9; \qquad m_2(\Theta)=0,1.$$

У цих формулах змінні m_1 і m_2 відносяться до датчиків першого і другого типів. Отримані свідчення можна скомбінувати за допомогою наступної спеціальної форми правила комбінування Демпстера для одержання такої *комбінованої маси:*

$$m_1 \oplus m_2(Z) = \sum_{X \cap Y = Z} m_1(X) m_2(Y). \qquad (1.79)$$

У формулі (1.79) операція підсумування поширюється на всі елементи, для яких перетинання $X \cap Y = Z$. Знак операції $\oplus$ відповідає операції ортогональної суми, або прямій сумі. Результат цієї операції визначається шляхом підсумовування перетинань добутків мас правої частини правила. Правило комбінування Демпстера дозволяє комбінувати маси для одержання нової маси, що являє собою *консенсус* стосовно оригінальних, можливо конфліктуючих свідчень. Важливо відзначити, що це правило повинне застосовуватися для комбінування свідчень, що мають взаємно незалежні помилки, а це – не те ж саме, що незалежно зібрані свідчення.

У табл. 1.6 у наведено маси і перетинання добутків для середовища з літаками різних типів. Записи в цій таблиці були обчислені шляхом перехресного множення добутків мас по рядках і стовпцях, як показано нижче, де T_{ij} – *i*-й рядок та *j*-й стовпець таблиці.

Таблиця 1.6 – Підтвердження свідчень

	$m_2(Б) = 0,9$	$m_2(\Theta) = 0,1$
$m_1(\{Б, В\}) = 0,7$	$\{Б\} = 0,63$	$\{Б, В\} = 0,07$
$m_1(\Theta) = 0,3$	$\{Б\} = 0,27$	$\Theta = 0,03$

Слідом за обчисленням окремих добутків мас, відповідно до описаних формул, виконується додавання добутків за загальними правилами перетинання множин, відповідно до правила Демпстера:

$$m_3(\{Б\}) = m_1 \oplus m_2(\{Б\}) = 0,63 + 0,27 = 0,90 – \text{бомбардувальник};$$
$$m_3(\{Б, В\}) = m_1 \oplus m_2(\{Б, В\}) = 0,07 – \text{бомбардувальник або винищувач};$$
$$m_3(\Theta) = m_1 \oplus m_2(\Theta) = 0,03 – \text{відсутність довіри}.$$

Значення $m_3(\{Б\})$ виражає довіра до того, що розглянутий літак являє собою бомбардувальник і тільки бомбардувальник. Але під значеннями $m_3(\{Б, В\})$ і $m_3(\Theta)$ мається на увазі додаткова інформація. Відповідні множини включають бомбардувальник, тому правдоподібним є припущення, що їхні ортогональні суми можуть зробити свій внесок у визначення ступеня довіри до того, що розглянутий літак є бомбардувальником. Тому значення

суми 0,07 + 0,03 = 0,1 для цих множин може бути додане до ступеня довіри, що стосується підмножини бомбардувальника, для одержання максимального ступеня довіри до гіпотези, що літак може бути бомбардувальником, який дорівнює 0,90, тобто є правдоподібним ступенем довіри. Таким чином, ступінь довіри не обмежується одним значенням, а виражається у виді *ряду ступенів довіри* до свідчення. У даному випадку ряд ступенів довіри починається з мінімального значення 0,9, відповідно до якого відомо, що розглянутий літак — бомбардувальник, і закінчується максимальним правдоподібним значенням ступеня довіри 0,90 + 0,1 = 1, відповідно до якого цей літак може являти собою бомбардувальник. При цьому передбачається, що істинний ступінь довіри знаходиться десь у діапазоні від 0,9 до 1.

У таких міркуваннях на основі свідчень вважається, що свідчення задають *інтервал прояву свідчення* (evidential interval). При цьому в міркуваннях на основі свідчень нижня границя інтервалу називається *обґрунтуванням* (support – Spt); в теорії Демпстера–Шефера вона позначається як Bel (belief). З іншого боку, верхню границю прийнято називати правдоподібністю (plausibility – Pis). Для даного прикладу інтервал свідчень дорівнює [0,90, 1], тобто нижня границя дорівнює 0,90, а верхня границя – 1. Обґрунтування являє собою мінімальний ступінь довіри, заснований на свідченні, а правдоподібність — максимальний ступінь довіри, якого бажано досягти. Узагалі говорячи, діапазони, у яких змінюються Bel і Pis, виражаються співвідношенням

$$0 \le \text{Bel} \le \text{Pis} \le 1. \qquad (1.80)$$

У теорії Демпстера–Шефера нижню і верхню границі іноді називають нижньою і верхньою ймовірностями, відповідно до оригінальної статті Демпстера. У табл. 1.7 показані деякі широко застосовувані інтервали прояву свідчень.

Таблиця 1.7 – Деякі широко застосовувані інтервали прояву свідчень

Інтервал прояву свідчення	Область визначення змінної	Значення
[1, 1]		Цілком істинний.
[0, 0]		Цілком помилковий.
[0, 1]		Цілком невідомий.
[Bel, 1]	$0 < \text{Bel} < 1$	Як правило такий, що обґрунтовує.
[0, Pls]	$0 < \text{Pis} < 1$	Як правило такий, що спростовує.
[Bel, Pis]	$0 < \text{Bel} \le \text{Pis} < 1$	Такий, що і обґрунтовує, і спростовує.

Обґрунтування, або довірча функція, Bel, являє собою загальний ступінь довіри до множини і до всіх її підмножин. Таким чином, Bel – це вся маса,

що обґрунтовує множину і визначається в термінах маси:

$$\mathrm{Bel}(X) = \sum_{Y \subseteq X} m(Y) \,. \qquad (1.81)$$

Маса – це ступінь довіри до множини, а не до якої-небудь з її підмножин, а довірча функція застосовується до множини і до всіх її підмножин. Значення Bel являє собою сумарний ступінь довіри і тому є більш глобальним, ніж локальний ступінь довіри, що виражається масою. Маса і довірча функція зв'язані наступним співвідношенням:

$$m(X) = \sum_{Y \subseteq X} (-1)^{|X-Y|} \mathrm{Bel}(Y) \,, \qquad (1.82)$$

де $|X - Y|$ – *кардинальність* множини $X - Y = \{x | x \in X \text{ і } x \notin Y\}$.

Таким чином, $|X - Y|$ – це кількість елементів у множині $X - Y$.

Отже, довірчі функції визначаються в термінах мас, тому комбінація двох довірчих функцій також може бути виражена в термінах ортогональних сум мас множини і всіх її підмножин, наприклад, так:

$$\mathrm{Bel}_1 \oplus \mathrm{Bel}_2(\{Б\}) = m_1 \oplus m_2(\{Б\}) + m_1 \oplus m_2(\varnothing) = 0{,}90 + 0 = 0{,}90.$$

У звичайному випадку маса порожньої множини не записується, оскільки вона дорівнює нулю. Сумарний ступінь довіри до підмножини {*Б, В*}, що складається з бомбардувальника і винищувача, включає більше підмножин, ніж наведена вище множина:

$$\mathrm{Bel}_1 \oplus \mathrm{Bel}_2(\{Б, В\}) = m_1 \oplus m_2(\{Б, В\}) + m_1 \oplus m_2(\{Б\}) + \\ + m_1 \oplus m_2(В) = 0{,}07 + 0{,}90 + 0 = 0{,}97.$$

У цей вираз включені терми для множин {*Б*} і {*В*}, оскільки вони являють собою підмножини множини {*Б, В*}.

Підмножині {*В*} маса не привласнена, тому $m(\{В\}) = 0$, і ця підмножина не робить ніякого внеску в суму. У дійсності $m(\{В\})$ і інші маси, що дорівнюють нулю, узагалі не вводилися в табл. 1.7, оскільки результат будь-якого перехресного добутку між ними дорівнює нулю. Якби маси були привласнені кожній підмножині множини {*А, Б, В*}, крім порожньої множини, то табл. 1.7 являла б собою таблицю з 49 комірок, відповідно до розрахунку $(2^3 - 1)(2^3 - 1) = 7 \cdot 7 = 49$.

Комбінована довірча функція для Θ, заснована на усіх свідченнях, має такий вигляд:

$$\mathrm{Bel}_1 \oplus \mathrm{Bel}_2(\Theta) = m_1 \oplus m_2(\Theta) + m_1 \oplus m_2(\{Б, В\}) + m_1 \oplus m_2(\{Б\}) = \\ = 0{,}03 + 0{,}07 + 0{,}90 = 1.$$

Фактично Bel(Θ) = 1 у всіх випадках, оскільки сума мас завжди повинна дорівнювати 1. При комбінуванні свідчень просто відбувається перерозподіл мас по різних підмножинах.

Інтервал прояву свідчення множини *S, EI*(*S*) може бути визначений у термінах ступеня довіри в такий спосіб:

$$EI(S) = [\text{Bel}(S),\ 1–\text{Bel}(S')]. \qquad (1.83)$$

Якщо $S = \{Б\}$, то $S' = \{A, B\}$ і має місце наступна формула, оскільки є елементи, відмінні від фокальних, то маса нефокальних елементів дорівнює нулю:

$$\text{Bel}(\{A, B\}) = m_1 \oplus m_2(\{A, B\}) + m_1 \oplus m_2(\{A\}) + m_1 \oplus m_2(\{B\}) = 0 + 0 + 0 = 0.$$

Таким чином, інтервал прояву свідчень для $\{Б\}$ визначається як:

$$EI(\{Б\}) = [0{,}90,\ 1 – 0] = [0{,}90,\ 1].$$

Аналогічним чином, якщо $S = \{Б, B\}$, то $S' = \{A\}$ і має місце наступна формула, оскільки $\{A\}$ – нефокальний елемент:

$$Bel(\{A\}) = 0.$$

Крім того, справедливі наведені нижче співвідношення, у яких інтервал прояву свідчень [0,1] відображає сумарний ступінь незнання стосовно підмножини $\{A\}$:

$$\text{Be1}(\{Б, B\}) = \text{Bel}_1 \oplus Bel_2(\{Б, B\}) = 0{,}97;$$

$$EI(\{Б, B\}) = [0.97,\ 1 – 0] = [0{,}97,\ 1];$$

$$EI(\{A\}) = [0,\ 1].$$

Метод обчислення коефіцієнтів достовірності. Даний метод був вперше застосований в комп'ютерній медичній системі MYCIN [35]. Ступінь підтвердження гіпотези *H* був спочатку визначений як коефіцієнт вірогідності, що визначається як різниця між *ступенем довіри* (belief) і *ступенем недовіри* (disbelief):

$$CF(H, E) = MB(H, E) – MD(H, E), \qquad (1.84)$$

де *CF* (Certainty Factor) – коефіцієнт вірогідності гіпотези *H,* обумовлений наявністю свідчення *E;*

MB (measure of belief) – міра підвищення ступеня довіри до гіпотези *H* з огляду на наявність свідчення *E;*

MD (measure of disbelief) – міра підвищення ступеня недовіри до

гіпотези H з огляду на наявність свідчення E.

Коефіцієнт достовірності – це спосіб об'єднання двох значень, ступеня довіри і ступеня недовіри, в одне число. Об'єднання мір довіри і недовіри в одне число здійснюється для досягнення двох цілей:

- насамперед, коефіцієнт достовірності може використовуватися для ранжирування гіпотез у порядку їхньої важливості;
- якщо в пацієнта є деякі симптоми, що свідчать про наявність декількох можливих захворювань, то необхідно, щоб на підставі медичних аналізів у першу чергу було проведене обстеження для діагностування саме того захворювання, якому відповідає найбільше значення CF.

Міри довіри і недовіри були визначені в термінах ймовірностей за такими формулами:

$$MB(H,E) = \begin{cases} 1, & \text{якщо } P(H) = 1; \\ \dfrac{\max[P(H/E), P(H)] - P(H)}{\max[1,0] - P(H)} & \text{якщо } P(H) < 1; \end{cases} \quad (1.85)$$

$$MD(H,E) = \begin{cases} 1, & \text{якщо } P(H) = 0; \\ \dfrac{\min[P(H/E), P(H)] - P(H)}{\min[1,0] - P(H)} & \text{якщо } P(H) > 0; \end{cases} \quad (1.86)$$

Відзначимо, що значення max[1,0] завжди дорівнює 1, а значення min[l, 0] завжди дорівнює 0. Але значення 1 і 0 записані в термінах max і min, оскільки це дозволяє показати формальну симетрію між виразами MB і MD. У табл. 5.1 показані деякі характерні випадки застосування значень MB, MD і CF, що визначені на підставі наведених вище формул.

Таблиця 1.8 – Деякі характерні випадки застосування значень MB, MD і CF

Характеристики	Значення
Інтервали значень	$0 \le MB \le 1$; $0 \le MD \le 1$; $-1 \le CF \le 1$
Деяка істинна гіпотеза $P(H/E) = 1$	$MB = 1$; $MD = 0$; $CF = 1$
Деяка помилкова гіпотеза $P(H'/E) = 1$	$MB = 0$; $MD = 1$; $CF = -1$
Відсутність свідчення $P(H/E) = P(H)$	$MB = 0$; $MD = 0$; $CF = 0$

Коефіцієнт достовірності CF показує, який чистий ступінь довіри до гіпотези, що заснований на деякому свідченні. Позитивне значення CF говорить про те, що свідчення обґрунтовує гіпотезу, оскільки $MB > MD$. Той випадок, у якому $CF = 1$, означає, що свідчення повністю доводить гіпотезу. З іншого боку, випадок $CF = 0$ відповідає одній з двох можливостей. По-перше, $CF = MB - MD = 0$ може означати, що є нульові рівні і MB, і MD,

тобто, що відсутнє яке-небудь свідчення. По-друге, можливо, що $MB = MD$ і обидва ці значення відмінні від нуля, а це означає, що довіра до гіпотези спростовується таким самим ступенем недовіри. На жаль, спростування сильного ступеня довіри такою самою недовірою веде не просто до незнання, але до деякого стану плутанини (а це набагато гірше). Наприклад, що може бути неприємнішим для водія, який під'їхав до перехрестя і не знає, куди повернути, та почув від пасажира, що указує вліво, пораду повернути праворуч!

У системі MYCIN значення *CF* антецедента правила повинне бути більше 0,2, для того щоб антецедент розглядався як істинний і активізував правило. Таке значення 0,2 розглядається як граничне значення, але воно не визначене як фундаментальна аксіома теорії коефіцієнтів вірогідності. Замість цього граничне значення розглядається як довільний спосіб зведення до мінімуму можливості активізації правил, що лише незначною мірою підтверджують гіпотезу. Без використання граничного значення можуть активізуватися численні правила, що є несуттєвими або які взагалі не мають значення, тому ефективність системи істотно зменшується.

У 1977 році приведене вище визначення *CF у* системі MYCIN було змінено і прийняло такий вид:

$$CF = \frac{MB - MD}{1 - \min(MB, MD)} \tag{1.87}$$

Це було зроблено з метою ослаблення впливу єдиних частин свідчень, що спростовують гіпотезу, на численні підтверджуючі частини свідчень. Якщо при використанні зазначеного визначення застосовуються наступні значення $MB = 0{,}999$; $MD = 0{,}799$ то значення *CF* приймає вид:

$$CF = \frac{0{,}999 - 0{,}799}{1 - \min(0{,}999,\ 0{,}799)} = \frac{0{,}200}{1 - 0{,}799} = 0{,}995$$

Це значення досить істотно відрізняється від значення, отриманого відповідно до попереднього визначення (1.84), при якому результат був $CF = 0{,}999 - 0{,}799 = 0{,}200$, і тому не відбувалася активізація правила, оскільки значення не було більше від граничного значення 0,2.

У табл. 1.9 показані правила, які застосовуються в системі MYCIN для комбінування свідчень в антецедентах правил. Слід зауважити, що ці правила збігаються з правилами системи PROSPECTOR, заснованими на нечіткій логіці.

Таблиця 1.9 – Правила комбінування свідчень

Свідчення *E*	Вірогідність антецедента
*E*1 AND *E*2	min[*CF*(*H, E1*), *CF*(*H, E2*)]
*E*1 *OR E*2	max[*CF*(*H, E1*), *CF*(*H, E2*)]
NOT *E*	-*CF*(*H, E*)

Наприклад, якщо дано наступний логічний вираз, що застосовується для комбінування свідчень

$$E = (E1 \text{ AND } E2 \text{ AND } E3) \text{ OR } (E4 \text{ AND NOT } E5),$$

то значення свідчення *E* можна обчислити в такий спосіб:

$$E = \max[\min(E1, E2, E3), \min(E4, -E5)].$$

При використанні значень $E1 = 0{,}9$; $E2 = 0{,}8$; $E3 = 0{,}3$; $E4 = -0{,}5$; $E5 = -0{,}4$ одержуємо наступний результат:

$$E = \max[\min(0{,}9; 0{,}8; 0{,}3), \min(-0{,}5; -(-0{,}4))] = \max[0{,}3; -0{,}5] = 0{,}3.$$

Більш складна ситуація відбувається у разі, коли самі свідчення є випадковими подіями з відомою невизначеністю. Тоді фундаментальна формула визначення коефіцієнта достовірності *CF* для правила

IF *E* THEN *H*

задається наступною формулою:

$$CF(H, e) = CF(E, e)CF(H, E), \qquad (1.88)$$

де *CF*(*E*, *e)* — коефіцієнт достовірності свідчення *E*, що формує антецедент правила на основі невизначеного свідчення *e*;

CF(*H*, *E*) — коефіцієнт достовірності гіпотези, у якій передбачається, що свідчення відоме з усією вірогідністю, коли $CF(E, e) = 1$;

CF(*H*, *e)* — коефіцієнт достовірності гіпотези, заснованої на невизначеному свідченні *e*.

Таким чином, якщо усі свідчення в антецеденті відомі з усією вірогідністю, то формула для коефіцієнта достовірності гіпотези набуває наступного вигляду, оскільки $CF(E, e) = 1$:

$$CF(H, e) = CF(H, E).$$

Як приклад застосування таких коефіцієнтів достовірності розглянемо значення *CF* для правила визначення наявності стрептококової інфекції, описаного таким чином:

IF 1) The stain of the organism is gram positive, and

2) The morphology of the organism is coccus, and

3) The growth conformation of the organism is chains THEN There is suggestive evidence (0,7) that the identity of the organism is streptococcus.

У цьому правилі коефіцієнт достовірності гіпотези при наявності вірогідного свідчення визначається наступною формулою й іменується також коефіцієнтом ослаблення (attenuation factor):

$$CF(H, E) = CF(H, E1 \cap E2 \cap E3) = 0{,}7.$$

Це визначення коефіцієнта ослаблення засноване на припущенні, що усі свідчення ($E1$, $E2$ і $E3$) відомі з повною достовірністю. Таким чином, справедлива наступна формула, у якій e_i являють собою свідчення, які спостерігається:

$$CF(E1, e1) = CF(E2, e2) = CF(E3, e3) = 1.$$

Коефіцієнт ослаблення виражає ступінь вірогідності, що відноситься до гіпотези, якщо є деякі достовірні свідчення.

Якщо не відомі усі свідчення з повною достовірністю, то для визначення результуючого значення CF застосовується фундаментальна формула (1.88).

Наприклад, якщо в результаті статистичної обробки чи експертних оцінок визначені ймовірністі свідчень

$$CF(E1, e1) = 0{,}5;\ CF(E2, e2) = 0{,}6;\ CF(E3, e3) = 0{,}3,$$

то у цьому випадку одержуємо такий результат:

$$CF(E, e) = CF(E1 \cap E2 \cap E3) = \min[CF(E1, e1), CF(E2, e2), CF(E3, e3)] = \\ = \min[0{,}5;\ 0{,}6;\ 0{,}3] = 0{,}3.$$

Тоді

$$CF(H, e) = CF(E, e)CF(H, E) = 0{,}3 \times 0{,}7 = 0{,}21.$$

1.3.4. Оцінка якості діагностичних моделей. Основи ROC-аналізу

Оцінка якості діагностичних кореляційних та регресійних моделей була розглянута у відповідних розділах.

Розглянемо випадок, коли модель диференціальної діагностики двох станів (D_0 – позитивний результат; D_1 – негативний результат) заснована на порогових вирішальних правилах виду

$$\text{IF } (x > r) \text{ THEN } D_0 \text{ ELSE } D_1,$$

де x – числове значення деякого діагностичного показника, а r – його порогове значення, необхідне для прийняття рішення щодо належності об’єкта діагностики до одного з класів (D_0 або D_1).

Тоді оцінка якості діагностичних моделей і відповідних до них порогових вирішальних правил, а також діагностична вага показника x виконується на основі чотириелементної таблиці спряженості (табл. 1.10), яка будується на результатах класифікації моделлю й на об'єктивній належності об'єктів до класів.

Таблиця 1.10 – Таблиця спряженості результатів класифікації

	Дійсно	
Модель	D_0	D_1
D_0	*TP*	*FP*
D_1	*FN*	*TN*

Що є позитивним результатом D_0, а що – негативним D_1, залежить від конкретної задачі. Якщо прогнозується захворювання, то D_0 – клас "Хворі", а D_1 – "Здорові". І навпаки, якщо визначається ймовірність того, що людина здорова, то D_0 – "Здорові", а D_1 – "Хворі". Елементами табл. 1.10 є:

- *TP* (True Positives) – кількість вірно класифікованих позитивних результатів;
- *TN* (True Negatives) – кількість вірно класифікованих негативних результатів;
- *FN* (False Negatives) – кількість позитивних результатів, що класифіковані як негативні (помилка I роду або "помилковий пропуск");
- *FP* (False Positives) – кількість негативних результатів, що класифіковані як позитивні (помилка II роду або "помилкове виявлення").

Об'єктивна цінність моделі визначається такими показниками:

- чутливістю (Sensitivity) $Se = TP/(TP + FN)\cdot 100$ %;
- специфічністю (Specificity) $Sp = TN/(TN + FP)\cdot 100$ %.

У медичній діагностиці, при класифікації пацієнтів на хворих і здорових, чутлива діагностична модель характеризується гіпердіагностикою – максимальному запобіганні пропуску хворих, а специфічна модель діагностує тільки істинно хворих. Це важливо у випадку, коли, наприклад, лікування хворого пов'язане із серйозними побічними ефектами і коли гіпердіагностика пацієнтів не бажана.

Інтегральним критерієм якості діагностичної моделі є ROC-аналіз [36, 37]. ROC-крива є графіком залежності

$$Se = f(100 - Sp)$$

при зміні параметра моделі, що керує точністю моделі – граничного елемента r. Площа під кривою ROC - AUC ($0{,}5 \le AUC \le 1$) використовується як скалярна міра оцінки прогностичної здатності моделі. У роботі [37] розглядається посилений ROC-аналіз, у роботі [36] проаналізовано поводження ROC-кривої при нерівновеликих обсягах вибірок класів D_0 і D_1,

отримано залежність імовірностей помилок I і II роду від асиметрії обсягів вибірок D_0 і D_1 і визначено умови оптимальності порога.

Контрольні питання

1. Класифікація математичних методів синтезу вирішальних правил.
2. Детерміновані методи синтезу вирішальних правил. Їх реалізація.
3. Класифікація ймовірнісних методів синтезу вирішальних правил.
4. Метод послідовного аналізу (метод Вальда). Його реалізація.
5. Класифікація методів синтезу вирішальних правил на основі теорії розпізнавання образів.
6. Принципи дискримінантного аналізу.
7. Метод порівняння із прототипом (еталоном). Його реалізація.
8. Метод K-найближчих сусідів. Його реалізація.
9. Метод потенційних функцій. Його реалізація.
10. Методи січних площин, узагальненого портрета та голосування.
11. Методи розпізнавання, засновані на нечіткій логіці.
12. Методи розпізнавання на основі штучних нейронних мереж.
13. Застосування ймовірнісної логіки при синтезі вирішального правила. Метод Байєса.
13. Реалізація методу Байєса для незалежних дихотомічних ознак.
14. Теорія Демпстера–Шефера. Її основні визначення та аксіоми.
15. Синтез вирішальних правил на основі логіки Демпстера–Шефера.
16. Синтез вирішальних правил на основі коефіцієнтів достовірності.
17. Оцінка якості діагностичних моделей. Основи ROC-аналізу.