МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

С. В. Розсоха, І. Б. Туркін, І. В. Шостак

ФОРМАЛЬНІ МЕТОДИ ІДЕНТИФІКАЦІЇ ТА ПРОГНОЗУВАННЯ ОБ'ЄКТІВ І ПРОЦЕСІВ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

Навчальний посібник

Харків «ХАІ» 2014

УДК 004.942:519.876.5 (075.8)

ББК 32.97я73

Р65

Рецензенти: д-р техн. наук, проф. В. М. Левикін,
д-р техн. наук, проф. Г. Г. Асєєв

Розсоха, С. В.

Р65 Формальні методи ідентифікації та прогнозування об’єктів і процесів програмної інженерії [Текст] : навч. посіб. / С. В. Розсоха, І. Б. Туркін, І. В. Шостак. – Х. : Нац. аерокосм. ун-т ім. М. Є. Жуковського «Харк. авіац. ін-т», 2014. – 133 с.

ISBN 978-966-662-330-3

Наведено мету і принципи побудови математичних моделей, об’єктів і процесів програмної інженерії, поняття структури математичної моделі й узагальнений алгоритм її побудови. Розглянуто вимоги до експериментальних даних, оцінок параметрів і безпосередньо самих моделей об’єктів і процесів програмної інженерії. Викладено найбільш поширені у програмній інженерії типи моделей і відповідні їм методи ідентифікації та прогнозування.

Для студентів спеціальностей 8.05010301 «Програмне забезпечення систем», 8.05010302 «Інженерія програмного забезпечення».

Іл. 25. Табл. 7. Бібліогр. : 42 назви

УДК 004.942:519.876.5 (075.8)
ББК 32.97я73

ISBN 978-966-662-330-3

ВСТУП

Ідентифікація і прогнозування є основними етапами прийняття рішень (ПР) у людській практиці. Комп'ютеризація ПР як об'єкт програмної інженерії на сьогодні накопичила представницький набір апробованих засобів синтезу формальних структур, що забезпечують адекватне подання процесів людської діяльності у вигляді машинних програм.

Загалом процедура формалізації будь-якої задачі ПР передбачає послідовну реалізацію таких етапів: аналіз процесу, попереднє оброблення даних, аналіз наявності нелінійностей, формування структури моделі, введення припущень і спрощень, вибір методу реалізації моделі, планування машинного експерименту, оброблення результатів моделювання.

Матеріал цього навчального посібника побудовано на основі навчальних посібників В. Є Снитюка «Прогнозування. Моделі. Методи. Алгоритми» і П. І. Бідюка «Методи прогнозування». При цьому матеріал переглянуто і доповнено з позицій застосування формальних методів при комп'ютеризації задач ПР у будь-якій галузі людської практики.

У першому і другому розділах розглянуто типи математичних моделей, поняття їхньої структури, наведено методику машинного моделювання. Третій розділ присвячено застосуванню різницевих рівнянь до описання статистичних даних, у четвертому розділі висвітлено комплекс питань щодо комп'ютерного прогнозування динаміки розвитку процесів, які подано у вигляді системи різницевих рівнянь. П'ятий розділ посібника містить відомості про метод групового врахування аргументів і особливості побудови на цій основі алгоритмів самоорганізації.

У посібник поряд із класичними підходами до формалізації задач ПР включено матеріали, які ілюструють застосування методів і засобів штучного інтелекту. Так, у шостому розділі розглянуто основи байєсівського методу аналізу ймовірнісних процесів, сьомий розділ присвячено опису найбільш вживаних методів Soft Computing, а саме: методам оброблення нечіткої інформації (побудова нечітких відношень і організація нечіткого логічного виведення); аналізу нечітких експертних заключень; мурашиним алгоритмам; програмуванню генетичних виразів.

Кожен розділ посібника містить перелік контрольних запитань.

Виробництво й використання комп'ютерних програм у цей час є масовою діяльністю, розробленням програм займаються майже сім мільйонів чоловік, а використовують їх у своїй професійній діяльності за фахом десятки мільйонів. У зв'язку з постійно зростаючими обсягами програмних розробок є потреба у підготовці фахівців, здатних вирішувати проблеми створення нових програмних продуктів на інженерній основі, із використанням накопиченого запасу знань в області програмування й керування системами.

Сформовану структуру й зміст підготовки фахівців треба розширити методами керування, планування і регулювання робіт, адаптуючи їх до умов колективного розроблення програмних систем з гарантованою якістю. Передумовами цього є становлення нової спеціальності, що одержала назву програмної інженерії, або інженерії програмного забезпечення (Software Engineering), що увібрала в себе накопичений запас знань у практиці й теорії програмування за останні десятиліття.

У зв'язку з цим предметом навчання студентів, а саме майбутніх розроблювачів програмного забезпечення, менеджерів програмних проектів, тестерів, верифікаторів, контролерів якості тощо мають стати не тільки теоретичні й прикладні методи проектування, а й інженерні методи керування колективом, планування й оцінювання якості виконуваних робіт і укладення їх у заданий термін і вартість проекту.

Мета даного посібника – подати класичні методи й засоби програмної інженерії (Software engineering) у систематизованому вигляді для їх застосування у процесах проектування, тестування й оцінювання якості програмних систем.

Програмна інженерія – це інтегрування принципів інформатики й комп'ютерних наук з інженерними підходами, розробленими для матеріального виробництва.

Дисципліна програмної інженерії може розглядатися як інженерна галузь, що має більш тісні зв'язки з комп'ютерними науками, ніж інші інженерні галузі. Серед інженерних дисциплін вона виділяється нематеріальною природою програмного забезпечення і його принциповою незвідністю до фізичних процесів навколишнього світу. Використовуючи досягнення інформатики, програмна інженерія займається вирішення завдань здешевлення програмних продуктів за рахунок розробленням

методів масового виробництва високоякісного програмного забезпечення.

Термін «Інженерія програмного забезпечення» був розглянутий у 1968 році на Конференції НАТО «Інженерія програмного забезпечення» і призначався подолати існуючу на той час «кризу програмного забезпечення».

Програмні інструменти призначені для забезпечення підтримки процесів життєвого циклу програмного забезпечення, дозволяють автоматизувати окремі повторювані дії й зменшити завантаження фахівців із програмної інженерії одноманітними операціями. Ці інструменти становлять основу програмної інженерії. За своїми можливостями вони можуть варіюватися від підтримки окремих індивідуальних завдань до охоплення усього життєвого циклу (платформа розроблення).

У свою чергу, методи програмної інженерії формують певну процедуру дій, спрямованих на досягнення успіху в деякій конкретній сфері. Методи зазвичай надають відповідні нотації, словники термінів і процедури виконання певного набору завдань, а також рекомендації з оцінювання і перевірки виконуваного процесу й одержуваного в його результаті продукту. Методи, як і інструменти, можуть мати різний масштаб. Конкретні методи описані у відповідних розділах.

Розрізняють евристичні й формальні методи програмної інженерії.

Евристичні методи – послідовність приписань або процедур оброблення інформації, що виконується з метою пошуку більш раціональних і нових конструктивних рішень.

Евристичні методи звичайно протиставляють формальним методам вирішення, що спираються на точні математичні моделі. У психологічній і кібернетичній літературі під евристичними методами розуміють будь-які методи, спрямовані на скорочення перебору, або індуктивні методи вирішення завдань. До них відносять:

– структурні методи, що припускають побудову системи з функціональної точки зору, починаючи з високорівневого розуміння поводження системи з поступовим уточненням деталей на більш низьких рівнях;

– методи, орієнтовані на дані, а також на розроблення структур даних, якими маніпулює створюване програмне забезпечення, а функції при цьому розглядаються як вторинні;

– об’єктно-орієнтовані методи, що являють собою програмну систему у вигляді сукупності об'єктів;

– методи, орієнтовані на конкретну область застосування, і спеціалізовані методи, що розробляються з урахуванням конкретних розв'язуваних завдань, наприклад усілякі системи захисту інформації.

Під терміном «формальні методи» розуміють ряд операцій, до складу яких входять створення формальної специфікації системи, аналіз і доказ специфікацій, реалізація системи на основі перетворення формальної специфікації у програми й верифікація програм. Усі ці дії залежать від формальної специфікації програмного забезпечення. Формальна специфікація – це системна специфікація, записана мовою, словник, синтаксис і семантика якої визначені формально. Необхідність формального визначення мови припускає, що ця мова ґрунтується на математичних концепціях. Тут використовується область математики, що називається дискретною математикою й ґрунтується на алгебрі, теорії множин і алгебрі логіки.

Формальні методи поділяють на такі категорії:

– мови й нотації специфікацій, які можуть бути орієнтовані на модель, властивості й поводження, наприклад формальні методи опису вимог;

– методи трансформації, основані на уточненні (трансформації), перетворенні специфікацій у кінцевий результат, максимально близький бажаному, тобто це – здійсненний програмний продукт;

– методи підтвердження, що ґрунтуються на строгому математичному доказі точності будь-яких характеристик вихідних передумов і одержуваного продукту з використанням теорем і перевірки точності моделей.

Слід зазначити, що строгі формальні методи зазвичай не ефективні в області розроблення прикладних систем, оскільки вони потребують виконання складних і трудомістких доказів на основі звичайно досить неповних даних. Більше того, успіх проекту з розроблення ПЗ залежить від такої кількості непередбачених факторів, що формальні методи не завжди ґарантують його досягнення.

Таким чином, цей навчальний посібник дасть можливість у процесі вивчення дисципліни повною мірою засвоїти знання, вміння і навички в сфері використання формальних методів у програмній інженерії.

1 МЕТА І ПРИНЦИПИ ПОБУДОВИ МАТЕМАТИЧНИХ МОДЕЛЕЙ

Розділ присвячено аналізу принципів побудови математичних моделей процесів різної природи, поданих експериментальними або статистичними даними. Розглядаються мета побудови моделі та елементи структури моделі, які необхідно визначити в результаті аналізу процесу або об'єкта, що описується математичною моделлю. Наведено аналіз структурного і функціонального методів побудови моделей; перший являє собою теоретичний метод на основі законів і закономірностей функціонування процесу, а другий орієнтований на збір і використання експериментальних даних. Узагальнений алгоритм побудови моделі за експериментальними даними дає уявлення про загальні етапи моделювання, які можна застосовувати до процесів різної природи. Також сформульовано вимоги до експериментальних даних, оцінок параметрів і моделі в цілому. Дотримання цих вимог дає можливість будувати моделі високого ступеня адекватності та скорочувати затрати часу на побудову й аналіз якості моделі. Розділ закінчується переліком деяких типів математичних моделей, що описують стаціонарні і нестаціонарні процеси, а також процеси з нелінійностями відносно змінних і параметрів.

1.1 Мета побудови математичних моделей

Можна виділити три мети побудови математичних моделей процесів і об'єктів:

- поглиблене вивчення процесів;
- прогнозування значень змінних за допомогою моделі;
- використання моделі для створення (синтезу) систем керування.

Поглиблене вивчення процесів за допомогою математичних моделей дозволяє дослідити кількісні зв'язки між вхідними та вихідними змінними, дослідити, яким чином змінюються вихідні змінні при варіації вхідних у широкому діапазоні, і розглянути поведінку процесів на будь-яких часових інтервалах у прийнятному масштабі часу. Математична модель, що будується для цього, може бути дуже складною і трудомісткою, оскільки вона повинна враховувати тонкощі взаємодії кількісних і якісних змінних з можливим урахуванням реального часу, тобто в даному разі часто використовують імітаційне моделювання. За допомогою математичних моделей можна виявити ефекти і явища, які недоступні при безпосередньому спостереженні за допомогою приладів. Крім того, при проектуванні нових систем у різних галузях можна швидко змінювати варіанти реалізації системи завдяки можливості її швидкого дослідження на моделі, виявити вплив початкових умов і обмежень на ключові змінні.

Прогнозування значень змінних виконується, як правило, на основі набагато простіших моделей, ніж поглиблене вивчення процесів. Для цього

досить зручними є дискретні моделі у вигляді авторегресійних рівнянь (АР) і авторегресії з ковзним середнім (АРКС). Часто використовують також стандартизоване подання у просторі станів неперервних чи дискретних моделей. Тип моделі залежить ще й від того, який прогноз потрібен – короткостроковий, середньостроковий або довгостроковий. Задача прогнозування може бути вельми складною, якщо стохастичний процес нелінійний і нестаціонарний.

Для розв'язання задачі синтезу систем керування процесами використовують різні класи моделей у неперервному та дискретному часі, які подають, як правило, в уніфікованій формі простору станів. У цьому випадку моделі не повинні мати дуже високу точність (адекватність), оскільки вони працюють у замкненому контурі з від'ємним зворотним зв'язком. Це сприяє асимптотичному зменшенню похибок керування до мінімально можливих значень навіть при використанні грубих моделей.

Моделювання системи керування на комп'ютері дозволяє за мінімальний час дослідити можливість застосування різних варіантів законів керування, комбінацій керуючих дій і визначити ефективність системи керування при зміні вхідних керуючих сигналів у широкому діапазоні їхніх значень. Комп'ютерна система керування може працювати як в автоматичному режимі (при керуванні технічними або технологічними процесами в реальному часі), так і в режимі порадника при керуванні економічними, екологічними, соціальними чи іншими складними процесами. При використанні комп'ютерної системи як порадника керуючі дії, зґенеровані комп'ютером, порівнюють з експертними оцінками входів і на підставі певного критерію якості вибирають найбільш прийнятні варіанти керуючих впливів.

1.2 Поняття структури математичної моделі

Уведемо поняття структури математичної моделі, яке використовуватимемо надалі. Структура моделі містить у собі такі елементи (параметри):

1. **Порядок моделі**, тобто порядок диференціального, різницевого чи іншого рівняння, що використовується для опису динаміки процесу чи об'єкта. Наприклад, стохастичне різницеве авторегресійне (АР) рівняння другого порядку має вигляд

$$y(k) = a_0 + a_x y(k-1) + a_2 y(k-2) + \varepsilon(k).$$

Тобто порядок цього різницевого рівняння визначається кількістю задіяних у часі значень змінної, що використовуються у правій частині рівняння для опису змінної в лівій частині. Стохастичним воно називається тому, що в правій частині є випадкова змінна $\varepsilon(k)$, призначення якої ми пояснемо трохи нижче.

2. **Вимірність моделі** визначається кількістю рівнянь, що використовуються для схематичного описання об'єкта чи процесу. Процес,

котрий описують одним рівнянням, називається одновимірним, або скалярним. Процес, що описується двома і більше рівняннями, називається багатовимірним. Хоча більшість процесів у природі є багатовимірними, часто обмежуються одновимірними моделями виходячи з їхньої простоти та зручності застосування.

3. **Наявність нелінійностей та їхній характер**. Виявити нелінійності – не завжди просте завдання. Так, для механічних і деяких інших систем наявність нелінійностей можна визначити шляхом вивчення законів, закономірностей і особливостей їхнього функціонування. Наприклад, відомо, що для механічних систем характерною є наявність нелінійностей типу «люфт», «тертя», гістерезис.

При побудові регресійних моделей найчастіше трапляються нелінійності відносно змінних і нелінійності відносно параметрів. Прикладом нелінійності відносно змінних може бути поліноміальна регресія вигляду

$$y(k) = a_0 + a_1 x(k) + a_2 x^2(k) + a_3 x^3(k) + \varepsilon(k).$$

Коефіцієнти цього рівняння можна оцінювати звичайним методом найменших квадратів (МНК) [10, 11] при належній побудові матриці вимірювань (вона розглядатиметься нижче). Нелінійність відносно параметрів зумовлена наявністю у моделі добутків коефіцієнтів, наприклад, у вигляді

$$y(k) = a_0 + a_1 a_2 x(k) + a_2 \exp(-bx(k)) + \varepsilon(k).$$

Коефіцієнти (параметри) такої моделі неможливо оцінити за допомогою звичайного МНК, тому для розв'язання цієї задачі використовують нелінійний МНК, метод максимальної правдоподібності або інші методи нелінійного оцінювання.

4. **Час запізнення** реакції на виході об'єкта відносно вхідного сигналу. Запізнювання на вхід (лаг) досить легко враховується як у неперервних, так і в дискретних моделях. Для моделі з дискретними змінними у вигляді різницевого рівняння

$$y(k) = a_0 + a_1 y(k-1) + a_2 x\ \ (k-d) + \varepsilon(k)$$

час запізнення d є ціле число, що дорівнює кількості періодів дискретизації вимірювань, на які запізнюється вихідний сигнал відносно вхідного. Тривалість періоду дискретизації вимірювань залежить від динаміки конкретного процесу і може змінюватися у межах від декількох мікросекунд у фізико-технічних системах до одного року в макроекономіці.

Розглянемо модель неперервного процесу у вигляді передавальної функції із запізненням

$$W(p) = \frac{Ke^{-p\tau}}{1 + Tp},$$

де K – статичний коефіцієнт передачі об'єкта; p – змінна Лапласа [9]; T – стала часу; τ – час запізнення на вхід. Запізнення у дискретній формі

d і запізнення в неперервній формі τ пов'язані між собою так:

$$\hat{d} = int(\tau / T_S).$$

де $\hat{d}$ – оцінка часу запізнення у дискретній формі $(\hat{d} = 0, 1, 2, ...)$.

5. **Тип збурень**, що діють на процес, і спосіб їх урахування. Під збуреннями розуміють вхідні впливи процесу, котрі створюють, як правило, негативні умови для його перебігу і не використовуються з тих чи інших причин як керуючі. Збурення поділяють на детерміновані і стохастичні, а враховуються вони в адитивній чи мультиплікативній формі. Вище ми навели різницеві рівняння, у яких збурення $\varepsilon(k)$ входить в процес у адитивній формі. Приклад мультиплікативної форми

$$h(k) = \upsilon(k)[\alpha_0 + \alpha_1 h(k-1)],$$

де $\upsilon(k)$ – мультиплікативне збурення.

Найчастіше збурення описують розподілами випадкових величин (статистичні моделі), але в окремих випадках його можна виміряти й описати функціонально (математичні моделі збурень). Наприклад, можна виміряти температуру навколишнього середовища, яка впливає на хід реакції у хімічному реакторі, і побудувати відповідну функціональну залежність температури від часу.

Вибір структури моделі, адекватної процесу, – задача не проста і вирішується, як правило, ітераційно. Спочатку структуру моделі оцінюють наближено на підставі дослідження закономірностей перебігу процесу, аналізу кореляційних функцій, візуального аналізу даних. При цьому вибирають декілька найбільш імовірних структур (кандидатів). Потім обчислюють оцінки параметрів моделей-кандидатів і вибирають найкращу з них, використовуючи відповідні статистичні характеристики якості моделей.

Якщо жодна з моделей-кандидатів не може вважатися адекватною для конкретного застосування, то необхідно дослідити на інформативність експериментальні дані, які можуть виявитися недостатньо інформативними для оцінювання моделі. У такому разі знадобляться повторний або додатковий збір експериментальних даних і коригування структури моделі.

1.3 Два основних методи побудови математичних моделей

Основними методами побудови математичних моделей є:

- структурний;
- функціональний.

Структурний метод передбачає моделювання внутрішнього механізму взаємодії змінних, відображує їхні фактичні взаємозв'язки.

Критерієм правильності структурної моделі є однаковий характер поведінки основних змінних реального процесу і моделі.

Розглянемо, наприклад, зростання інфляції унаслідок випуску додаткової грошової маси. Оскільки логіка цього процесу досить проста й існують експериментальні (статистичні) дані, які ілюструють зростання інфляції, то можна постулювати, що інфляція описується диференціальним або різницевим рівнянням першого/другого порядку (рисунок 1.1).

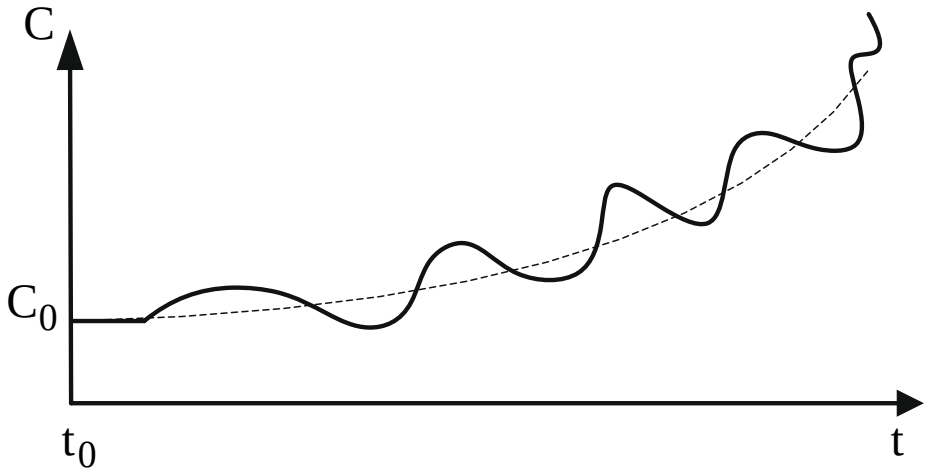


Рисунок 1.1 – Можлива крива зростання інфляції

Структурним підходом можна скористатися, наприклад, для побудови математичної моделі процесу трансформування власності або макроекономіки в цілому. Для цього необхідно визначити вхідні керуючі змінні, збурення та вихідні змінні, а також визначити, якого типу зв'язки існують між ними (рисунок 1.2).

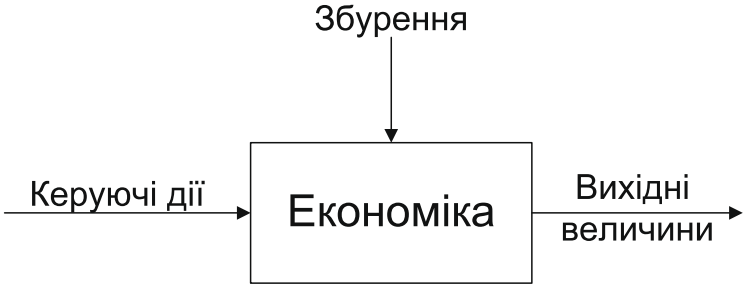


Рисунок 1.2 – Спрощене зображення макроекономічного процесу

Серед керуючих змінних макроекономічного процесу можна виділити внутрішні та зовнішні інвестиційні потоки, потоки сировини, робочої сили, нові технології та структурні зміни в промисловості в цілому, а також в окремих галузях. Метою використання керуючих змінних є досягнення заданих рівнів макроекономічних показників – рівень виробництва валового внутрішнього продукту (ВВП), індекси інфляції, індекс людського розвитку, середня заробітна плата тощо.

Як правило, у такі моделі вводять у явному вигляді *збурення* – випадкові змінні, що негативно впливають на перебіг процесу. Наприклад, при створенні моделі макроекономіки збуреннями можуть бути:

- помилкові рішення уряду;
- затримка платежів між підприємствами та державами;
- значні коливання цін на енергоносії;

- хронічна технологічна відсталість;
- швидкі зміни податкового законодавства;
- відплив капіталу за кордон;
- використання недостовірної статистичної інформації.

Очевидно, що врахувати подібну інформацію у моделі надзвичайно складно, а тому випадкові змінні агрегують (об'єднують) і подають у моделі однією або двома випадковими змінними, які охоплюють усі збурення.

На основі знання логіки взаємодії змінних процесу та використання відомих макроекономічних законів (наприклад, зрівноваженого розвитку процесів) будується система рівнянь, які описують розвиток окремої галузі або макроекономіки в цілому.

Функціональний підхід використовують для формального описання процесу, не заглиблюючись у фактичну структуру цього процесу і взаємодії його змінних. У цьому разі процес = «чорний ящик».

Наприклад, для побудови моделі ціноутворення можна скористатись такими вхідними змінними: I – обсяг імпорту; D – обсяг грошової маси в обороті; P – місячний обсяг виробництва. За вихідну змінну можна взяти індекс споживчих цін C, тобто

$$C = f(I, D, P).$$

Для побудови цієї моделі потрібно мати чотири часових ряди, які необхідні для обчислення оцінок коефіцієнтів моделі.

Очевидно, що функціональний підхід є простішим від структурного і саме він найчастіше використовується на практиці. Гнучкість цього підходу дає можливість відносно швидко побудувати високоякісні моделі для прогнозування та синтезу систем керування.

1.4 Узагальнений алгоритм побудови моделі

Розглянемо узагальнений алгоритм побудови математичної моделі на основі експериментальних даних. Він є узагальненим щодо застосування до систем або процесів практично будь-якого його типу і може бути поданий у вигляді таких кроків:

1. *Визначення мети побудови моделі*, попереднє вивчення процесу (об'єкта). На цьому етапі визначається мета побудови моделі, тобто буде вона використовуватися для поглибленого вивчення процесу, прогнозування його стану чи керування. Виконується аналіз функціонування процесу на основі літературних джерел і (можливо) експериментальних даних за їх наявності з метою встановити кількість входів і виходів, логіку взаємодії складових частин процесу, визначити можливі зовнішні збурення і їхній тип (детерміновані чи випадкові). Якщо можливо, слід установити розподіл імовірностей для випадкових збурень або функціональний опис для детермінованих збурень. У разі наявності моделей аналогічних процесів їх також необхідно досконально вивчити та врахувати можливі недоліки.

2. *Попередня оцінка структури моделі*. На основі вивчення процесу, виконаного на першому етапі, необхідно встановити типи структур моделей-кандидатів. Їх може бути декілька залежно від того, наскільки невизначеною є інформація стосовно процесу. Чим більшою є невизначеність, тим більше структур моделей необхідно досліджувати в процесі побудови адекватної моделі.

3. *Планування експерименту* та підготовка до його виконання. На цьому етапі виконують такі дії щодо планування виконання експерименту з метою отримання експериментальних даних:

- визначають діапазони зміни вхідних і вихідних величин, збурень;
- установлюють дискретність зміни вхідних величин, період дискретизації вимірів (якщо змінні неперервні), визначають типи вимірювальних приладів;
- планують режими роботи процесу, для яких потрібно зібрати експерементальні дані;
- якщо дані мають статистичний характер, то визначають періодичність їх збору та занесення у базу даних;
- задають тип, обсяг і якість продукції, що буде вироблена під час експерименту, а також визначають необхідні обсяги сировини і енергії.

4. *Виконання експерименту* та формування бази даних. На цьому етапі ревізують розроблений на третьому етапі план експерименту і формуються часові ряди з вимірів (або статистичних даних), які використовуватимуться для обчислення оцінок параметрів математичних моделей.

5. *Обчислення оцінок параметрів* (коефіцієнтів) математичних моделей на основі експериментальних даних. Оцінюють параметри для усіх моделей-кандидатів, вибраних на другому етапі. Для цього необхідно:

- вибрати метод оцінювання параметрів моделі залежно від її структури;
- попередньо обробити дані; залежно від конкретної задачі це можуть бути масштабування, логарифмування, цифрова фільтрація, вилучення недостовірних даних тощо;
- обчислити оцінки (векторів) параметрів моделей.

6. *Визначення ступеня адекватності кожної моделі-кандидата* процесу за допомогою статистичних критеріїв. Визначити найкращу модель з множини кандидатів.

7. Якщо побудована модель відповідає висунутим вимогам (за точністю прогнозу чи якістю керування), то завершити процедуру; інакше перейти на восьмий крок.

8. *Уточнити структуру моделі*, зібрати, якщо це необхідно, додаткові експериментальні дані й перейти на п’ятий крок.

Хоча планування і виконання експерименту для соціально-економічних і фінансових систем є досить складною проблемою, в окремих випадках це цілком можливо, особливо якщо підприємство має наміри впровадити нові

інформаційні технології оброблення даних і методи прогнозування розвитку процесів на виробництві. Наприклад, цілком можливо спланувати та провести інвестиційний експеримент, експерименти з новими технологіями, новими типами продукції.

1.5 Вимоги до експериментальних даних, оцінювання параметрів моделі

1.5.1 Вимоги до експериментальних даних

1. Вимога *неперервності та синхронності даних*. Експериментальні дані повинні вимірюватися і реєструватися через однакові проміжки часу (період дискретизації вимірювань T_S). Цю вимогу необхідно виконувати для процесів будь-якого типу – технічних, економічних, екологічних тощо. Порушення цієї вимоги призводить до зміни спектрального складу вимірюваного сигналу, що неприпустимо, оскільки при цьому змінюється інформативність сигналу. Крім того, вимірювання вхідних і вихідних сигналів потрібно робити синхронно, тобто в одні й ті ж моменти часу. Інакше вони будуть непридатні для побудови передавальних функцій, оскільки порушуються причинно-наслідкові зв'язки між входами та виходами. Як правило, у системах управління реального часу задача збору вимірюваних даних має найвищий пріоритет.

2. *Вибірка даних має бути представницькою*. Це означає, що вона повинна охоплювати тривалий проміжок часу, щоб включити в розгляд усі режими роботи, які передбачається описати моделлю. Розрізняють два основних режими роботи процесів: перехідний і усталений. У перехідному режимі система керування переводить процес з початкового стану в заданий. Перебування процесу в заданому (номінальному) стані протягом певного відносно тривалого проміжку часу називають усталеним режимом роботи.

Прикладом широко відомого перехідного процесу є процес нагрівання до кипіння вмісту каструлі на кухонній плиті. На початку цього процесу ми задаємо режим максимальних витрат енергії, щоб скоротити тривалість процесу – нагрівання. Після досягнення режиму кипіння затрати енергії можна суттєво скоротити, при цьому процес переходить в усталений режим «повільного» кипіння. Подачу великої кількості енергії у початковий момент часу можна порівняти з подачею на вхід об'єкта сигналу у вигляді сходинки, а зміну температури вмісту каструлі можна вважати за перехідну характеристику цього процесу. Очевидно, що безліч прикладів такого типу можна знайти в промисловості. Зараз у нашому суспільстві спостерігається перехідний процес від ідеалістичного ладу, який ґрунтувався на суспільній власності на засоби виробництва, до капіталістичного з приватною власністю.

3. Вибірка вимірюваних даних має бути інформативною. Найчастіше інформативність пов'язують з кількістю похідних, що їх містить

вимірюваний сигнал. Чим більшу кількість похідних можна отримати з вимірів, тим інформативнішим є сигнал. Наприклад, припустимо, що процес описується диференціальним рівнянням другого порядку

$$a_2 \frac{d^2 y}{dt^2} + a_1 \frac{dy}{dt} + a_0 = b\ u(t),$$

де $y(t)$ – вихідний сигнал процесу; $u(t)$ – вхідний сигнал процесу; $\theta^T = [a_0, a_1, a_2]$ – вектор коефіцієнтів рівняння, які необхідно оцінити за допомогою експериментальних даних. Очевидно, що оцінки коефіцієнтів a_2 і a_1 можна обчислити тільки в тому разі, якщо виміри $y(t)$ містять другу і першу похідні за часом

Іноді інформативність визначають величиною дисперсії сигналу, тобто чим більшою є дисперсія, тим вища інформативність сигналу. Так, константа має нульову дисперсію і відповідно мінімальну інформативність.

Вимога інформативності виконується тоді, коли вхідний сигнал задовольняє умову достатнього збудження процесу. Основна ідея достатнього рівня збудження полягає у тому, щоб смуга частот вхідного сигналу перекривала амплатудно-частотну характеристику процесу. Тобто вихідний сигнал $y(t)$ буде інформативним тоді, коли достатньо інформативним буде вхідний сигнал $u(t)$. Умову інформативності (достатнього збудження) задовольняють такі основні типи вхідних сигналів: білий шум, псевдовипадковий двійковий сигнал і одиничний імпульс. Білий шум (гауссів процес) теоретично має нескінченний частотний спектр; досить широкі спектри мають і два інші типи сигналів.

З одного боку, для збудження процесу на його вхід необхідно подавати інформативний сигнал типу білого шуму, а з іншого – такий сигнал може бути недопустимим з точки зору фізики функціонування процесу (подача на вхід такого сигналу може призвести до руйнування процесу чи до створення аварійної ситуації). Тому в таких системах часто використовують як збуджуючий сигнал завдання регулятору, якщо він має форму сходинки, тобто має фронт прямокутного імпульсу. У багатьох випадках можна використовувати гармонічні сигнали, які сприймаються «легше» більшістю об’єктів, ніж білий шум або єдиний імпульс. Так, при дослідженні механічних систем часто використовують єдині імпульси, гармонічні збуджуючі сигнали та їх комбінації, а при дослідженні технологічних процесів до вхідного сигналу керування додають 10 … 15 % білого шуму, який забезпечує достатній рівень «збудження» процесу.

1.5.2. Вимоги до оцінювання параметрів моделі

Точність оцінювання параметрів моделі залежить від якості вимірюваних даних, коректності попереднього оброблення даних і від того, наскільки правильно вибрано метод оцінювання. Так, для оцінювання параметрів лінійних і псевдолінійних (нелінійних відносно змінних) моделей

можна застосовувати звичайний МНК [11] і його модифікації, а для оцінювання моделей, нелінійних відносно параметрів, – нелінійний МНК, метод максимальної правдоподібності та інші методи, розроблені для оцінювання параметрів нелінійних моделей.

Існують такі стандартні вимоги до оцінювання параметрів математичних моделей:

1. Оцінки мають бути *незміщеними*. Це означає, що оцінки параметрів не мають містити систематичної похибки, яка збільшує або зменшує оцінки параметрів на всіх вибірках даних або на різних відрізках однієї вибірки. Формально незміщеність оцінок параметрів записують так:

$$\boldsymbol{E}\left[\hat{\theta} = \theta\right].$$

де $\boldsymbol{E}$ – символ математичного сподівання; $\hat{\theta}$ – вектор оцінок параметрів; θ – істинне значенні вектора параметрів.

2. Оцінки мають бути *консистентними*, тобто оцінка $\hat{\theta}$ вектора параметрів повинна наближатись до свого істинного значення θ у міру збільшення обсягу вибірки даних. Оскільки оцінка $\hat{\theta}$ – це випадкова величина, то наближення до істинного значення можливе тільки в імовірнісному розумінні. Консистентна оцінка має задовольняти таке співвідношення:

$$\boldsymbol{p}\left(\left|\hat{\theta}_{\boldsymbol{k}} - \theta\right| < \varepsilon\right) \to \boldsymbol{1} \text{ при } \boldsymbol{k} \to \infty,$$

де $\varepsilon > \boldsymbol{0}$ – мале число; $\boldsymbol{p}$ – символ імовірності; $\hat{\theta}_{\boldsymbol{k}}$ – оцінка вектора параметрів у момент $\boldsymbol{k}$. Відомо, що довжина перехідного процесу при оцінюванні моделі залежить від кількості оцінюваних параметрів і вимірності моделі, а тому обсяги вибірок даних усіх випадках повинні бути якомога більшими. Проблеми з необхідними обсягами даних виникають, як правило, при моделюванні економічних і соціальних систем; при побудові моделей технічних систем такі проблеми досить рідкісні.

3. Оцінки повинні бути *ефективними*, а це означає, що з множини допустимих, незміщених і консистентних оцінок слід вибрати найближчі до оцінюваних параметрів, тобто такі, що мають найменші відхили від середнього значення.

Іншими словами, це вимога мінімальності дисперсії оцінки, яка формально записується так:

$$\boldsymbol{var}(\hat{\theta}) \to \boldsymbol{min}.$$

Незміщені ефективні оцінки параметрів лінійної моделі можна

отримати, наприклад, за допомогою методу найменших квадратів, якщо при оцінюванні виконуються такі умови:

– *похибка моделі* $e(k) = y(k) - \hat{y}(k)$ є центрованою величиною, де $y(k)$ – значення ряду, отримане експериментально (або статистичні дані), $\hat{y}(k)$ – оцінка змінної, отримана за допомогою побудованої моделі;

– *похибка моделі* – це некорельований процес, тобто відсутня автокореляція похибок:

$$cov[\, e(k)] = E\,[e(k)e(k-l)] = \begin{cases} \sigma_e^2, k = l, \\ 0, \quad k \neq l. \end{cases}$$

Корельованість похибки означає, що вона містить інформацію про процес. Тобто необхідно коригувати структуру моделі таким чином, щоб похибка стала некорельованою.

1.5.3 Вимоги до математичної моделі

1. Модель має бути адекватною процесу чи об'єкту. Адекватність означає, що модель повинна:

а) відображаюти найбільш характерні зв'язки та взаємодію між змінними процесу;

б) ураховувати можливі керуючі дії (сигнали);

в) ураховувати вплив зовнішніх збурень і шуми вимірювань;

г) ураховувати початкові значення змінних та обмеження на них.

Формально адекватність визначають за допомогою ряду статистичних величин. Наприклад, досить часто використовують середньоквадратичну похибку моделі (СКП), яку обчислюють за формулою

$$СКП(x_S, x_m) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[x_S(k) - x_m(k)]^2},$$

де $x_S(k)$ – вимір вихідного сигналу об'єкта в момент k; $x_m(k)$ – оцінка вихідного сигналу об'єкта, отримана за оціненою моделлю. Для лінійних моделей запропоновано кілька статистичних параметрів, використовуваних при оцінюванні адекватності, які розглядатимуться нижче. Використання одного параметра для визначення ступеня адекватності моделі є некоректним підходом, *оскільки оцінки параметрів – це випадкові величини*, а тому збільшення кількості критеріїв адекватності сприяє підвищенню ймовірності вибору адекватної моделі.

2. Рівняння моделі повинні мати *розв'язок*, тобто бажано мати аналітичний або, якщо це неможливо, то числовий розв'язок.

Одним з принципів, яких необхідно дотримуватись при побудові моделі, є такий: *"у моделі не повинно бути нічого зайвого, крім необхідного"*. Звичайно, що дотримуватися цього принципу досить непросто, і на практиці буває так, що модель дійсно має надзвичайно складну структуру, що також можна виправдати необхідністю досягнення

високого ступеня її адекватності процесу. Це особливо стосується нелінійних процесів. Разом з тим при побудові лінійних моделей у вигляді авторегресії чи авторегресії з ковзним середнім достатньо побудувати модель, статистичні характеристики якої збігаються зі статистичними характеристиками часового ряду, на основі якого вона оцінюється. Такі спрощені моделі виявляються цілком придатними для прогнозування і керування процесами. Загалом питання складності моделі вирішується у кожному випадку окремо.

3. Модель має бути досить *універсальною*, щоб її можна було застосувати для описання класу однотипних процесів або для описання функціонування процесу в різних умовах.

Наприклад, для описання моторної функції людини (реакція на зовнішні збуджуючі сигнали) застосовують звичайне диференціальне рівняння другого порядку, яке подають у вигляді функції передачі такого ж порядку:

$$W(s) = \frac{Ke^{-\tau s}}{(1-T_1 s)(1-T_2 s)},$$

де K – статичний коефіцієнт передачі об'єкта; τ – час запізнення щодо входу, який у середньому дорівнює для людини 300…350 мс; T_1, T_2 – сталі часу. Така передавальна функція може використовуватись, наприклад, для описання реакції людини на зовнішні відео- або аудіосигнали, що надходять через систему візуального сприйняття або аудіосистему (поширений приклад – водіння автомобіля чи іншої машини). Значення параметрів моделі можуть бути різними для різних людей, але структура моделі залишається незмінною. Таким чином, наведена модель описує широкий клас біологічних систем і цілком відповідає умові універсальності.

При моделюванні технічних систем широко застосовують ланки першого і другого порядків, що відповідають звичайним диференціальним рівнянням таких же порядків. На основі таких простих ланок можна побудувати моделі будь-якої складності. Доволі відомий у техніці та екології клас систем з розподіленими параметрами. Це, наприклад, процес поширення домішок в атмосфері та водному середовищі, механічні коливання сонячних батарей та антен супутників, крила літака, локомотива з вагонами на залізниці, автомобіля з причепом і багато інших. Динаміку таких систем описують диференціальними рівняннями з частинними похідними.

4. Вимога *робасності* (robust – сильний, міцний). Робасність означає, що модель повинна давати прийнятний прогноз вихідної змінної не тільки на тому відрізку часового ряду, на основі якого вона побудована, але й на будь-якому іншому відрізку, що відповідає вибраному режиму роботи. Робасність може розглядатися також як стійкість моделі до збурень, похибок і пропусків вимірів. Вимога робасності є особливо критичною для

систем, що працюють у реальному часі, оскільки нестійка модель може стати причиною створення аварійної ситуації.

5. Вимога *адаптивності*. Ця вимога означає, що хоча б частину параметрів моделі можна уточнювати в міру надходження нових вимірів від об'єкта. Ця вимога є обов'язковою при побудові моделей нестаціонарних систем, тобто систем, параметри яких є функціями часу. Системи керування, побудовані для нестаціонарних процесів, називаються адаптивними. Такі системи є досить складними з точки зору аналізу збіжності оцінок параметрів і похибок керування, а тому при проектуванні адаптивних систем необхідно особливу увагу приділяти питанням достатнього збудження процесу та вибору методу оцінювання параметрів.

1.6 Спрощена класифікація математичних моделей

У спеціальній літературі можна знайти класифікацію математичних моделей за різними критеріями, що приводять до визначення багатьох класів і підкласів моделей. Розглянемо спрощену класифікацію, яка може бути корисною на практиці.

1. Описові (вербальні) моделі

Технічне завдання, звіт, логічна схема взаємодії змінних, яка супроводжується словесним описом.

2. Математичні моделі

Вони будуються, наприклад, на основі рівнянь таких типів:

- диференціальні рівняння;
- різницеві рівняння;
- алгебраїчні рівняння.

Математичні моделі поділяють на два широких класи:

а) *аналітичні* – моделі, які описують вибрані змінні процесу; як правило, такі моделі відтворюють один з аспектів функціонування процесу чи об’єкта, наприклад, динаміку валового внутрішнього продукту, моторну функцію людини чи рух автомобіля, потяга або літака;

б) *імітаційні* – моделі, які приблизно відтворюють поточне функціонування процесу у вибраному масштабі часу.

Імітаційні моделі (ІМ) нагадують активний фізичний експеримент з використанням фактичних даних, отриманих безпосередньо з процесу.

Переваги імітаційних моделей:

- наочність результатів (проміжних та остаточних);
- динамічний характер відображення перебігу процесу;
- можливість урахування детермінованих і випадкових факторів, а також складних залежностей від них;
- простота корекції моделі (ввести додаткове рівняння, правило тощо);
- можливість дослідити процес на множині його реалізацій (тобто провести статистичний експеримент);

– практично необмежені можливості введення в модель таких елементів, як: рівняння будь-якого типу; логічні правила; нечітка логіка; статистичні розрахунки; оптимізаційні процедури; процедури прийняття рішень;

– об’єдання чітких алгоритмів з нечіткими (це непросто, але корисно при аналізі функціонування процесів з невизначеностями та формуванні рішень).

Види імітаційних моделей:

– висока вартість побудови та використання;

– вимагають багато часу для розроблення.

1.7 Деякі типи регресійних і різницевих рівнянь

Авторегресія. Рівняння авторегресії описує пам'ять процесу, тобто вплив значень попередніх станів на його поточний стан:

$$y(k) = a_0 + a_1 y(k-1) + ... + a_p y(k-p) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \varepsilon(k),$$

де a_i, $i = 1,..., p$ – коефіцієнти моделі, які оцінюються на основі значень часового ряду; p – порядок авторегресії, який визначається числом задіяних у часі значень ряду, що використовуються у правій частині рівняння для опису динаміки змінної у момент k; $k = 1, 2, ...$ дискретний час; $\varepsilon(k)$ – випадкова величина, поява якої зумовлена такими причинами:

– вплив випадкових збурень на процес, що моделюється;

– похибки рівняння, зумовлені неточно вибраною структурою (можливо, що не враховано деякі регресори, введено непотрібні незалежні змінні або робиться спроба моделювати нелінійний процес за допомогою лінійного рівняння);

– методичні й обчислювальні похибки, які з'являються при обчисленні оцінок коефіцієнтів рівняння.

Парна регресія має у правій частині незалежну змінну (регресор):

$$y(k) = a_0 + a_1 x(k) + \varepsilon(k),$$

де $x(k)$ – регресор (незалежна або екзогенна змінна), тобто $x(k)$ має три назви. Залежну змінну $y(k)$ називають ще ендогенною змінною.

Множинна регресія відображує вплив декількох незалежних змінних на залежну:

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + ... + a_p x_p(k) + \varepsilon(k),$$

де $x_1(k),..., x_p(k)$, $x_p(k)$ – регресори рівняння. Таке рівняння може містить також авторегресійну частину.

Авторегресія + множинна регресія = змішана регресія

$$y(k) = a_0 + \sum_{i=1}^{l} a_i y(k-i) + b_1 x_1(k) + b_2 x_2(k) + ... + b_p x_p(k) + \varepsilon(k).$$

Авторегресія з ковзним середнім порядку (p,q) $(АРКС(p,q))$

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{q} b_j \varepsilon(k-j) + \varepsilon(k).$$

Регресія, нелінійна відносно змінних (псевдолінійна регресія)

$$y(k) = a_0 + a_1 x(k) + a_2 x^2(k) + ... + a_p x^p(k) + \varepsilon(k).$$

Тобто в цьому разі це поліноміальна регресія порядку p. Коефіцієнти псевдолінійної регресії оцінюються за такими ж методами, що й лінійної, наприклад, за методом найменших квадратів (МНК) або методом максимальної правдоподібності (ММП).

Регресія, нелінійна відносно параметрів. Моделі, нелінійні відносно параметрів, містять адитивні члени, які мають добутки параметрів моделей або інші види зв'язку (крім адитивного) між параметрами:

$$y(k) = a_0 + a_1 e^{bx(k)} + \varepsilon(k).$$

Для оцінювання параметрів таких моделей необхідно застосовувати методи нелінійного оцінювання – нелінійний метод найменших квадратів, метод максимальної правдоподібності, метод Монте – Карло та інші.

Моделі гетероскедастичних процесів, тобто процесів, дисперсія яких змінюється у часі. Рівняння для умовної дисперсії (авторегресія першого порядку), має вигляд

$$h(k) = \beta_0 + \varepsilon^2(k-1) + \varepsilon_1(k),$$

де $h(k)$ – умовна дисперсія процесу в момент k; $\varepsilon^2(k)$ – квадрат залишків; $\varepsilon_1(k)$ – похибка моделі в момент k. Докладно моделі гетероскедастичних процесів розглядаються нижче.

Авторегресійна умовно гетероскедастична модель порядку $p(АРУГ(p))$ така:

$$h(k) = \beta_0 + \sum_{i=1}^{p} \beta_i \varepsilon^2(k-i) + \varepsilon_1(k).$$

Узагальнена авторегресійна умовно гетероскедастична модель $(УАРУГ(p,q))$ має вигляд

$$h(k) = \beta_0 + \sum_{i=1}^{p} \beta_i \varepsilon^2(k-i) + \sum_{i=1}^{q} \alpha_i h(k-i) + \varepsilon_1(k),$$

де $\alpha, \beta, \gamma \geq 0$ (щоб уникнути появи від'ємних значень умовних дисперсій).

Експоненціальна модель УАРУГ (умовна дисперсія як асиметрична функція ε, тобто моделювання впливу попередніх значень $\varepsilon(k-i)$ на волатильність)

$$\log(h(k)) = \alpha_0 + \sum_{i=1}^{p} a_i \frac{|\varepsilon(k-i)|}{h(k-i)} + \sum_{i=1}^{p} \gamma_i \frac{\varepsilon(k-i)}{h(k-i)} + \sum_{i=1}^{q} \beta_i \log(h(k-i)).$$

Модель УАРУГ-М (модифікована) – моделювання премії за ризик

$$y(k) = \beta + \gamma h(k) + \varepsilon(k),$$

$$h(k) = a_0 + a \sum_{i=1}^{p} \varepsilon^2(k-i) + \sum_{i=1}^{q} h(k) + \varepsilon_2(k).$$

Модель для прогнозування волатильності за допомогою УАРУГ

$$h(k+1) = \beta_0 + \beta_1 \varepsilon^2(k) + \gamma\, h(k),$$

$$h(k+j) = \beta_0 + (\beta_1 + \gamma_1) h(k+j-1).$$

Рівняння для умовної дисперсії і коваріації такі:

$$h(s(k)) = \alpha_0 + \alpha_1 \varepsilon^2(s(k-1)) + \alpha_2 h(s(k-1)),$$

$$h(f(k)) = \beta_0 + \beta_1 \varepsilon^2(f(k-1)) + \beta_2 h(f(k-1)),$$

$$cov[s(k), f(k)] = \gamma_0 + \gamma_1 \varepsilon\ (s(k-1)) + \gamma_2\, cov[s(k-1), f(k-1)]\ .$$

Коефіцієнт хеджування при використанні двофакторної моделі

$$\hat{b}^{*}(k) = \frac{cov[s(k), f(k)]}{\hat{h}^{2}(f)}.$$

Моделі коінтегрованих процесів, тобто нестаціонарних процесів, які можна об'єднати в межах однієї моделі, що має стаціонарні статистичні характеристики.

Концепція коінтегрованості змінних передбачає існування довгострокового зв'язку між значеннями змінних. Тобто ми припускаємо існування спільної урівноваженої траєкторії руху цих змінних, від яких вони можуть відхилятися на коротких проміжках часу. Однак економічні механізми в цілому діють таким чином, що рівновага відновлюється і зберігається на довгих часових інтервалах шляхом коригування відповідних відхилень від зрівноваженого стану.

У випадку коінтегрованості змінних $x(k)$ і $y(k)$ для них можна побудувати модель коригування похибки, яка поєднує динаміку змінних на коротких проміжках часу з довгостроковим зрівноваженим зв'язком і має такий вигляд:

$$\Delta x(k) = a_{10} \sum_{i=1}^{p} b_{1i} \Delta x(k-i) + \sum_{i=1}^{p} c_{1i} \Delta y(k-i) + \lambda_1 e(k-1) + \varepsilon_1(k),$$

$$\Delta y(k) = a_{20} + \sum_{i=1}^{p} b_{2i} \Delta y(k-i) + \sum_{i=1}^{p} c_{2i} \Delta x(k-i) + \lambda_2 e_x(k-1) + \varepsilon_2(k).$$

Коефіцієнти λ_1, λ_2 у наведених рівняннях називаються швидкістю пристосування (коригування).

Вище ми навели тільки деякі структури (типи) математичних моделей стаціонарних і нестаціонарних нелінійних процесів, що широко використовуються при описанні динаміки процесів різної природи. Зараз існує багато інших різновидів дискретних моделей, які описують різні складові реальних процесів. У подальшому ми намагатимемося виконати аналіз якомога більшої кількоств типів існуючих моделей, але слід зазначити, що зробити це в межах однієї книги досить складно.

1.8 Запитання і вправи

1) Назвіть три напрями застосування математичних моделей. Наведіть приклади практичного застосування.

2) Чому математична модель, призначена для застосування у замкненому контурі керування процесом (об'єктом), може мати нижчий ступінь адекватності, ніж модель, побудована для прогнозування?

3) Які вимоги висуваються до математичних моделей, призначених для застосування у реальному часі?

4) Сформулюйте необхідну умову для визначення оцінки періоду дискретизації вимірів (теорема Котельникова – Шеннона). Чому ця теорема є тільки необхідною умовою?

5) Назвіть п’ять елементів структури математичної моделі, поясніть їх на прикладі.

6) Поясніть фізичну суть запізнення на вході, наведіть приклади. Чи існують у природі процеси без запізнення і що необхідно для їх існування? Який середній час запізнення (реакції) для людини?

7) Яким чином ураховується запізнення щодо входу в дискретних і неперервних моделях?

8) У чому полягає фізична суть існування авторегресії?

9) Назвіть основні типи нелінійностей процесів. Які методи застосовують для їх оцінювання?

10) Сформулюйте узагальнений алгоритм побудови математичної моделі на основі експериментальних даних. Поясніть необхідність оцінювання декількох моделей-кандидатів.

11) Поясніть вимогу неперервності та синхронності до експериментальних даних.

12) Що означають вимоги представництва й інформативності до даних?

13) Сформулюйте три основних вимоги до оцінювання параметрів математичної моделі. За яких умов оцінки, отримані за методом найменших квадратів, задовольняють три загальні вимоги до оцінювань моделей?

14) Які вимоги має задовольняти модель процесу?

15) Що означає адекватність моделі процесу? Яким чином можна її визначити?

16) Для чого необхідно знаходити розв'язки рівнянь, що описують динаміку процесу?

17) Що означають універсальність і робастність моделі?

18) Яким чином досягається адаптивність моделі?

19) Сформулюйте переваги та вади імітаційного моделювання.

20) У чому полягає відмінність між парною і множинною регресіями?

21) Чому регресійні рівняння називають стохастичними?

22) Якими методами можна коректно оцінювати параметри моделей, нелінійних стосовно параметрів?

23) Який процес називають гетероскедастичним? Яку іншу назву мають такі процеси?

24) У чому полягає суть інтегрованості та коінтегрованості випадкових процесів?

2 МЕТОДИКА ПОБУДОВИ МАТЕМАТИЧНОЇ МОДЕЛІ СКЛАДНИХ ОБ'ЄКТІВ

Вище було розглянуто узагальнений алгоритм побудови математичної моделі, який можна застосовувати до будь-якого класу процесів. Тепер розглянемо методику, що більше орієнтована на побудову моделей фінансово-економічних, соціальних та екологічних процесів, для яких, як правило, набагато складніше поставити експеримент і отримати інформативні експериментальні дані в достатньому обсязі. Хоча в наведеному нижче вигляді методика також може бути успішно застосована до побудови моделей динаміки технічних систем і технологічних процесів.

Методика побудови математичної моделі процесу, в основу якої покладено ідеї Бокса і Дженкінса [1], складається з таких кроків:

- аналіз процесу, для якого будується модель, на основі літературних джерел, експертних оцінок перебігу процесу, візуального дослідження вимірів вхідних і вихідних змінних, поданих часовими рядами, та іншої доступної інформації.
- попереднє оброблення експериментальних даних;
- аналіз часових рядів на можливу наявність нелінійностей за допомогою множини статистичних критеріїв;
- вибір структур моделей-кандидатів, для чого необхідно виконати такі дії: обчислити та проаналізувати кореляційну матрицю для часових рядів залежної і незалежних змінних з метою визначення тих екзогенних змінних, які потрібно включити в модель. Обчислити автокореляційну (АКФ) і часткову автокореляційну (ЧАКФ) функції залежної змінної з метою оцінювання порядку авторегресійної частини моделі; оцінити характеристики інших складових структури математичної моделі;
- вибір методу (методів) оцінювання параметрів математичних моделей вибраних структур (найчастіше це є метод найменших квадратів (МНК), метод максимальної правдоподібності (ММП) та їхні модифікації");
- вибір найкращої з оцінених моделей-кандидатів за допомогою множини статистичних критеріїв.

Тепер розглянемо докладно кожний з етапів побудови моделі.

2.1 Аналіз процесу

Аналіз процесу – це надзвичайно важливий етап, коректне виконання якого потребує досвіду дослідження реальних процесів різної природи. Ігнорування цього етапу призводить до неможливості побудувати модель високого ступеня адекватності процесу. Як і в узагальненому алгоритмі побудови моделі, аналіз процесу спрямовується на виконання таких завдань:

- визначення кількості входів і виходів, тобто визначення вимірності процесу;

– установлення логічних зв'язків між змінними і аналіз можливостей їх математичного опису (коректного об'єднання в одному математичному виразі);

– визначення кількості зовнішніх збурень і їхнього типу (детерміноване чи стохастичне);

– установлення можливості декомпозиції процесу на окремі підпроцеси, які є простішими як з погляду їхнього функціонування, так і з погляду математичного опису; декомпозиція – досить складний процес, який ґрунтується на спеціальних математичних методах;

– якщо процес має ієрархічну структуру (верхній і нижній рівні функціонування), то необхідно чітко розмежувати ці рівні, визначити функції кожного з них і встановити, які типи зв'язків існують між ними; наприклад, технологічні процеси часто можна розмежувати на два та більше рівнів;

– використання знань зі спеціальної літератури щодо особливостей функціонування процесу, відомих законів і закономірностей його перебігу, виявлення існуючих моделей процесу та досвіду його теоретичного або експериментального дослідження;

– за наявності розроблених моделей досліджуваного процесу необхідно встановити їхні вади та переваги, а також визначити можливість подальшого використання (модифікації); аналіз і використання існуючих моделей дає можливість суттєво скоротити час та інші затрати на побудову та використання моделі.

Отриману інформацію максимально використовують для попереднього оцінювання структури моделі або декількох моделей-кандидатів, параметри оцінюють за допомогою експериментальних даних. Під час виконання аналізу функціонування досліджуваного процесу доцільно використовувати інформацію і порівнювати з різних джерел.

2.2 Попереднє оброблення даних

Процес попереднього оброблення експериментальних (статистичних) даних, як правило, містить такі операції:

– нормування і візуальна перевірка даних, а у разі потреби - їх коригування; нормування даних означає їх логарифмування або призведення до зручного діапазону їх зміни, наприклад, від 0 до 1; від -1 до +1; від +10 до -10 тощо;

– коригування даних полягає у заповненні пропусків і зменшенні викидів (екстремальних значень), що виходять за основний діапазон значень змінних;

– формування перших різниць або різниць вищих порядків, необхідних для аналізу відповідних складових часового ряду.

Поширеним методом нормування даних є їх логарифмування з наступним формуванням додаткових часових рядів з перших чи других

різниць. Нагадаємо, що перші різниці являють собою наближений дискретний аналог першої похідної, а другі різниці – другої похідної. Часто зі значень ряду віднімають його середнє, для того щоб отримати можливість працювати з відхилами, а не з повними значеннями змінних. Застосування того чи іншого методу нормування даних визначається у кожному випадку по-своєму.

Оброблення екстремальних (аномальних) значень

Хоча виявлення і оброблення екстремальних значень – це велика окрема тема для дослідження, розглянемо деякі можливості щодо розв'язання цієї проблеми. У подальшому будемо вважати дані аномальними, якщо вони виникли внаслідок впливу значних похибок вимірів або похибок, пов'язаних з некоректним збором статистичних даних. Якщо можна встановити факт наявності аномальних даних, то їх просто вилучають з розподілу.

Екстремальні значення – це правильно виміряні (зібрані) дані, які характеризують фактичні раптові (стрибкоподібні) зміни процесу. Підхід до розв'язання задачі дослідження екстремальних значень спостережень залежить від поставленої мети. Якщо дослідника цікавить тільки факт наявності таких значень (наприклад, з метою виявлення умов, що призводять до появи екстремальних значень), то досить мати надійний критерій для виявлення таких спостережень.

Якщо ж ставиться завдання виявити і вилучити екстремальні значення (наприклад, з метою покращання оцінок статистичних параметрів і моделей), то виникає задача – як правильно обробити дані. Спираючись на критерій для визначення екстремальних значень, можна було б визначити величину зміщення оцінок параметрів.

Критерії аналізу екстремальних значень застосовують з метою:

- вирівняти спостереження перед аналізом (як правило, суттєво зменшити великі значення);
- переконатися у тому, що дані містять аномальні значення, що свідчить про необхідність перегляду процедури отримання даних;
- виділити спостереження, які є цікавими з точки зору їх аномальності та по можливості описати встановлений ефект математично.

Класичний підхід до виявлення аномальних спостережень полягає у тому, що вибіркові спостереження розглядають як випадкові нормально розподілені величини. При цьому для аналізу (виявлення екстремальних значень) створюється статистика (статистичний тест), яка є чутливою до різких відхилень такого типу. Необхідно встановити розподіл цієї статистики при нульовій гіпотезі про те, що всі спостереження належать нормальній сукупності, а потім відхилити цю гіпотезу, якщо виявиться, що обчислена статистика їй суперечить.

Розглянемо можливий критерій відкидання екстремальних значень. Нехай дано деяку вибірку $\{x_1, x_2, ..., x_N\}$, $N \geq 3$, яка (за припущенням) є

випадковою для випадкової змінної X з нормальним розподілом $\{X\} \sim \mu_X, \sigma_X^2$. Позначимо відхилення від середнього через

$$\tilde{x}_i = x_i - \bar{x},\ i = 1,2, ..., N,$$

де $\bar{x} = \frac{1}{N} \sum_{k=1}^{N} x(k).$

Якщо виділити одне значення із спостережень, то вибіркове середнє для спостережень, що залишились, визначається як

$$\sum_{\substack{k=1 \\ k \neq 1}}^{N} \frac{x_i}{N-1} = \bar{x} - \frac{\tilde{x}}{N-1}. \qquad (2.1)$$

Якщо виділити декілька значень $x_1, x_2, ..., x_r$, то вибіркове серед них

$$\sum_{\substack{k=1 \\ k \neq 1}}^{N} \frac{x_i}{N-1} = \bar{x} - \frac{\tilde{x}_1 + \tilde{x}_2 + ... + \tilde{x}_r}{N-r}. \qquad (2.2)$$

Позначимо максимальний відхил через $\tilde{x}_m = x_m - \bar{x}$.

Тепер правило визначення екстремального значення можна сформулювати так: при заданому значенні c спостереження x_m відкидається, якщо $|\tilde{x}_m| > cSx$, де Sx – вибірковий стандартний відхил змінної X.

Якщо вибірка має досить великий обсяг, то значення x_m вилучається й аналіз продовжується. Величина константи c може змінюватися зі зміною довжини вибірки; вона пов'язана неявно зі t-статистикою:

$$\sqrt{\frac{Nc^2(\nu + \nu_0 - 1)}{\nu(\nu + \nu_0 - \frac{Nc^2}{\nu})}} \approx t_{1-a/2}^{\nu_0 + \nu - 1}, \qquad (2.3)$$

де $\nu = N - 1$; α – рівень значущості; ν_0 – будь-яке інше число додаткових степенів вільності, яке пов'язане з оцінюванням σ_x^2 за вибіркою, обсяг якої не дорівнює N ($\nu_0 = 0$, якщо такої інформації немає). Також існує наближений вираз для c через розподіл F:

$$c \approx \left(\frac{\nu}{N}\right)^{1/2} \left(\frac{3F_{1-q}}{1 + (3F_{1-q} - 1)/(\nu + \nu_0)} \right)^{1/2}, \qquad (2.4)$$

де $q = \Delta\hat{\sigma}_x^2 \frac{\nu}{N}$; $\Delta\hat{\sigma}_x^2$ – очікуваний приріст дисперсії внаслідок появи значень. При використанні (2.4) значення c визначаємо за виразом квадратного кореня.

Виразом (2.4) можна скористатися таким чином: якщо з ряду значення не вилучалися, то допустимий (очікуваний) відносний приріст дисперсії («премії») $\Delta\hat{\sigma}_x^2$ необхідно помножити на ν / N і, отже, отримаємо q. За його допомогою знайдемо відповідну верхню процентну точку для відношення дисперсій F_{1-q} при трьох і $\nu + \nu_0 - 1$ степенях вільності. За виразом (2.4) обчислимо значення c і застосуємо критерій до x_m. Очікуваний відносний приріст дисперсії («премія») залежить від того, наскільки ймовірною є поява значень; скажімо, можна взяти невеликий відносний приріст $\Delta\hat{\sigma}_x^2$ у межах 0,01 … 0,03.

Наприклад, якщо $N = 4, \nu = 3$ і $\nu / N = 0{,}75,$ то при $\Delta\hat{\sigma}_x^2 = 0{,}02$ маємо $q = 0{,}02 \cdot 0{,}75 = 0{,}015$. При трьох ступені вільності $F_{1-q} = F_{1-0{,}05} = 9{,}28$. Тепер знайдемо значення c:

$$c = (0{,}75)^{1/2}\left(\frac{3F_{0{,}95}}{1 + (3F_{0{,}95} - 1)/3}\right)^{1/2} = 1{,}449.$$

Спостереження x_m необхідно вилучити, якщо $|\tilde{x}_m| > 1{,}449\, S_x$. Можливі інші підходи до аналізу екстремальних значень.

Приклад 2.1. Обчислення критерію для виявлення екстремального значення. Є ряд значень $X = \{23{,}2;\ 23{,}4;\ 23{,}5;\ 24{,}1;\ 25{,}5\}$. Установити, чи можна вважати значення 25,5 екстремальним і чи потрібно його вилучити з вибірки.

Розв'язання

Обчислимо: $\bar{x} = 23{,}9$; $\tilde{x}_5 = 25{,}5 - 23{,}9 = 1{,}6$; $S_x = 0{,}77$. Для $\alpha = 0{,}05$, $\nu = 4$, $\nu_0 = 0$ і $N = 5$ за виразом (2.3) маємо

$$\left(\frac{15c^2}{9 - 5c^2}\right)^{1/2} = 2{,}776^3.$$

Звідси за методом спроб і помилок $c = 1{,}49$. Згідно з критерієм c

$$|1{,}6| > 1{,}49 \cdot 0{,}77 = 1{,}05,$$

тобто спостереження x_5 вилучається.

2.3 Аналіз наявності нелінійностей

Однією з проблем при визначенні структури моделі є установлення факту нелінійностей у досліджуваному процесі та їхнього типу. Для розв'язання цієї проблеми обов'язково використовують *візуальний аналіз* даних і формальні тести на наявність нелінійностей. Досвідченому фахівцеві з моделювання візуальний аналіз дозволяє оперативно виявити ділянки з лінійним або нелінійним трендом, які певним чином визначають наявність гетероскедастичних та значних викидів (імпульсів), що можуть суттєво впливати на якість моделі. Слід зазначити, що існують навіть окремі навчальні курси з візуального аналізу даних. Це свідчить про те, що не варто нехтувати такою доступною і ефективною можливістю дослідження даних.

Існує також ряд формальних тестів на наявність нелінійності. Розглянемо простий тест для її виявлення. Цей тест застосовується у тому разі, якщо можна набрати кілька груп (вибірок) спостережень для одного і того ж процесу:

$$\hat{F} = \frac{\frac{1}{m-2}\sum_{i=1}^{m} n_i (\bar{y} - \hat{y}_i)^2}{\frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2},$$

де $\bar{y}_i$ – середнє значення для i-ї групи (вибірки або групи вибірок) даних; $\hat{y}_i$ – середнє для лінійної апроксимації даних; m – кількість груп даних; n_i – кількість вимірів у i-й групі; n – загальна кількість вимірів. Фактично наведена статистика являє собою таке відношення:

$$\hat{F} = \frac{\textit{Відхилення середніх значень від прямої регресії}}{\textit{Відхилення значень y(k) від групових середніх}}.$$

Якщо статистика $\hat{F}$ з $\nu_1 = m - 2$, $\nu_2 = n - m$ степенями вільності досягає рівня значущості або перевищує його, то гіпотезу щодо лінійності слід відхилити. Недоліком даного підходу є те, що для його застосування необхідно мати декілька (не менше трьох) груп даних для одного і того ж процесу, які можна отримати в результаті виконання повторних експериментів. Очевидно, що це не завжди можливо.

Наявність нелінійності можна встановити також за допомогою вибіркових *нелінійних кореляційних функцій* (НКФ) [14, 15], тобто кореляційних функцій, розрахованих за вибірками експериментальних (статистичних) даних.

Наприклад, якщо дискретна НКФ

$$r_{yx^2}(s) = r_{y(k)x^2(k-s)} = \frac{1}{N} \frac{\sum_{k=s+1}^{N} \left\{ [y(k) - \bar{y}] [x(k-s) - x]^2 \right\}}{\sigma_y \sigma_x^2}, \; s = 0, 1, 2, 3, \ldots$$

містить значення, які суттєво відрізняються від нуля у статистичному розумінні, то процес має квадратичну нелінійність відносно регресора x.

Наявність нелінійного детермінованого тренду в процесі можна визначити шляхом оцінювання рівняння

$$y(k) = a_0 + c_1 k + c_2 k^2 + \ldots + c_m k^m,$$

яке є поліномом порядку m відносно часу. Якщо хоча б один з коефіцієнтів c_i, $i = 1,\ldots, m$ є статистично значущим, то гіпотеза щодо відсутності тренду відхиляється. Якщо тренд досить швидко змінює свій напрям розвитку і для нього важко знайти адекватне функціональне описання, то застосовують моделі випадкових трендів, які ґрунтуються на комбінаціях випадкових величин.

Автоматично оцінює структуру математичної моделі метод *групового врахування аргументів* (МГВА) [12, 13], запропонований академіком О. Г. Івахненком (Інститут кібернетики НАН України). Цей метод уже багато разів застосовувався до широкого класу процесів; його успішно використовують і зараз для моделювання процесів різної природи з нелінійностями та нестаціонарностями. Подальшим розвитком цього методу є нечіткий МГВА, який ґрунтується на нечіткому поданні параметрів оцінюваної моделі.

2.4 Формування структури моделі

На наступному етапі необхідно вибрати структури моделей-кандидатів. Поняття структури моделі розглянуто вище, але нагадаємо, що воно містить: *порядок моделі* (найвищий порядок рівнянь, що його утворюють): *вимірність* (кількість рівнянь моделі); час запізнення щодо входу (лаг) і його оцінювання; *можливі нелінійності* та їхній тип; *зовнішні збурення* та їхній тип (детерміновані чи випадкові, адитивні та мультиплікативні).

Для того щоб визначити, які незалежні змінні (регресори) необхідно включити в праву частину рівняння, обчислюють коефіцієнт кореляції між залежною та відповідною незалежною змінними.

Коефіцієнт кореляції, а в загальному випадку – кореляційна функція дозволяє установити факт існування зв'язку між змінними. Кореляція може бути лінійною або нелінійною залежно від типу функціональної залежності, яка фактично існує між змінними. У більшості практичних випадків розглядають лінійну кореляцію (взаємозв'язок) між змінними, але глибшого аналізу потребує використання нелінійних залежностей. Складну нелінійну залежність можна спростити, але знати про її існування необхідно для того, щоб у разі потреби побудувати складнішу за структурою модель процесу з

вищим ступенем адекватності.

Кореляційна матриця дає можливість установити існування зв'язку між залежною (ендогенною) змінною і незалежними (екзогенними) змінними в правій частині. Розглянемо кореляційну матрицю R вимірності 3×3, яка будується для трьох змінних x, y, z:

$$R = \begin{bmatrix} r_{yy} & r_{xy} & r_{zy} \\ r_{yx} & r_{xx} & r_{zx} \\ r_{yz} & r_{xz} & r_{zz} \end{bmatrix}, \quad де\ r_{yx} = r_{xy},\ r_{yz} = r_{zy}, r_{xz} = r_{zx}.$$

Нехай y – показник якості технологічного процесу; x, z – технологічні параметри, які за припущенням впливають на показник якості. Тобто ставиться завдання установити існування залежності вигляду

$$y = f(x,z),$$

яку можна подати у формі регресії змінної y на незалежні змінні x, z:

$$y(k) = a_0 + a_1 x(k) + a_2 z(k) + \varepsilon(k),$$

де k – дискретний час (наприклад, у частках секунди, секундах, хвилинах, годинах, днях, тижнях, місяцях тощо); $\varepsilon(k)$ – випадкова змінна, уведення якої у модель пояснюється такими причинами:

– часто буває неможливо встановити всі незалежні змінні, які впливають на залежну змінну, а тому наведене рівняння описує процес з похибкою;

– можуть існувати такі незалежні змінні, які неможливо виміряти і включити в модель, а тому їх розглядають як збурення і вважають, що їхній спільний вплив на залежну змінну описується випадковою змінною $\varepsilon(k)$;

– у наведеному вище регресійному рівнянні можуть бути пояснювальні змінні, які є формально корельованими із залежною змінною, але фактично не впливають на неї;

– будь-якому методу оцінювання параметрів рівнянь притаманні методичні похибки, які слід урахувати, якщо це можливо, в моделі.

Уважається, що сукупний вплив усіх зазначених факторів можна описати певною мірою за допомогою випадкової змінної $\varepsilon(k)$. Оскільки вона не враховується, то оцінити її значення (похибку моделі або залишок) можна лише після оцінювання коефіцієнтів моделі, тобто

$$\hat{\varepsilon}(k) = e(k) = y(k) - \hat{y}(k),$$

де $\hat{y}(k)$ – оцінка змінної $y(k)$, отримана за допомогою моделі; $y(k)$- фактичний вимір.

Для обчислення елементів матриці R необхідно мати синхронізовані в часі вибірки значень усіх трьох змінних y, x, z. Формула для розрахунку з кореляції має вигляд

$$r_{yx} = \frac{\frac{1}{N-1}\sum_{k=1}^{N}\{[x(k)-\bar{x}][y(k)-\bar{y}]\}}{\sigma_x \sigma_y},$$

де $\bar{x}, \bar{y}$ – вибіркові середні значення змінних x, y; σ_x, σ_y – стандартні відхили змінних, тобто корені квадратні з їх дисперсії. Наприклад,

$$\sigma_y = \sigma_y^2 = \left[\frac{1}{N-1}\sum_{k=1}^{N}[y(k)-y]^2\right]^{1/2},$$

де N – кількість вимірів змінної y; $\bar{y}$ – вибіркове середнє значення ряду $\{y(k)\}$, яке обчислюється за відомою формулою

$$\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y(k).$$

Очевидно, що перед формальним обчисленням коефіцієнтів кореляції необхідно проаналізувати процес і визначити наявність (або відсутність) логічних зв'язків між змінними. Це дозволяє увести до розгляду тільки ті змінні, які дійсно впливають на залежну змінну, наприклад, на показник якості. Очевидно, що для правильного вибору незалежних (екзогенних) змінних необхідно знати технологічний або інший процес, який моделюється. На основі значень коефіцієнтів кореляції приймається рішення про включення їх у рівняння регресії

$$y(k) = a_0 + b_1 x(k) + b_2 z(k) + \varepsilon(k),$$

яке в загальному вигляді може бути подано як

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + a_3 x_3(k) + \ldots + a_{p-1} x_{p-1}(k) + \varepsilon(k).$$

Можна показати, що між коефіцієнтами регресії b_1, b_2 і коефіцієнтами кореляції r_{yx}, r_{yz} існує однозначний взаємозв'язок.

Останнє рівняння являє собою *лінійну регресію* p-го порядку, але досить часто необхідно застосовувати більш складні нелінійні моделі. Характерним представником нелінійної відносно змінних регресії є поліноміальна регресія порядку $p-1$

$$y(k) = a_0 + a_1 x(k) + a_2 x^2(k) + a_3 x^3(k) + \ldots + a_{p-1} x^{p-1}(k) + \varepsilon(k).$$

Хоча в це рівняння включено тільки одну незалежну змінну $x(k)$, очевидно, що воно може бути розширене будь-якими іншими змінними.

Для визначення необхідності введення у рівняння регресії авторегресійної складової необхідно обчислити і дослідити вибіркову *автокореляційну функцію* змінної $y(k)$. Рівняння з авторегресійною складовою має вигляд

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + b_1 x(k) + b_2 z(k) + \varepsilon(k),$$

тобто в рівняння регресії уведено авторегресійну (АР) складову другого

порядку. Порядок авторегресії визначається за допомогою автокореляційної функції. Число коефіцієнтів автокореляційної функції, відмінних від нуля в статистичному значенні, і становитиме оцінку порядку авторегресії.

Природа авторегресії пояснюється існуванням так званої «пам'яті» процесу, яка проявляється у тому, що його поточний стан значною мірою визначається попередніми станами. Наприклад, стан людини вранці залежить від того, яким було самопочуття увечері та в попередні дні. На формування ринкових цін суттєво впливають їхні значення у попередні періоди часу. Поточний стан технічної системи або технологічного процесу також залежить від її стану в попередні моменти часу.

АКФ і ЧАКФ використовують для визначення попередньої оцінки порядку авторегресійної частини моделі, тобто скільки затриманих у часі значень необхідно брати для описання процесу. При цьому слід урахувати, що АКФ дає менш «чітку» оцінку порядку процесу, ніж ЧАКФ. Наприклад, для процесу АР(1) значення основної змінної $y(k)$ і $y(k-2)$ будуть корельованими, незважаючи на те, що $y(k-2)$ немає у моделі. Кореляція між $y(k)$ і $y(k-2)$, тобто ρ_2, дорівнює коефіцієнту кореляції між значеннями $y(k)$ і $y(k-1)$, помноженому на коефіцієнт кореляції між $y(k-1)$ і $y(k-2)$, або $\rho_2 = \rho_1\rho_1 = \rho_1^2$. Подібні «непрямі» кореляції наявні в АКФ будь-якого процесу авторегресії.

Вибіркову АКФ обчислюють за виразом

$$r_y(s) = r_{y(k)y(k-s)} = \frac{1}{N-1}\frac{\sum_{k=s+1}^{N}\{[y(k)-\bar{y}Iy(k-s)-\bar{y}]\}}{\sigma_v^2},\ s = 1, 2, 3,...,$$

де σ_v^2 – вибіркова дисперсія змінної $y(k)$; $\bar{y}$ – середнє значення вибірки даних.

Число коефіцієнтів АКФ, відмінних від нуля в статистичному розумінні, указує на порядок авторегресійної частини моделі. Для стаціонарного процесу (це процес із постійними середнім значенням, дисперсією і коваріацією) коефіцієнти $r_y(s)$ мають нормальний розподіл і нульове середнє.

На відміну від АКФ часткова АКФ між значеннями $y(k)$ і $y(k-s)$ включає вплив величин $y(k-1),..., y(k-s+1)$, а це означає, що коефіцієнти ЧАКФ чіткіше відображують зв'язок між окремими значеннями основної змінної. Так, для процесу АР(1) ЧАКФ між $y(k)$ і $y(k-2)$ дорівнює нулю за визначенням, що підтверджується обчисленими значеннями ЧАКФ.

Для того щоб знайти попередню оцінку порядку моделі, вибіркові

коефіцієнти ЧАКФ (тобто коефіцієнти, знайдені за вибіркою даних) можна обчислити також за допомогою простого методу, поданого нижче.

1. Формують додатковий часовий ряд з відхилів основної змінної

$$\{y'(k)\} = \{y(k)\} - \mu,$$

де μ – середнє значення ряду.

2. Формують рівняння першого порядку

$$y'(k) = \Phi_{11} y'(k-1) + e(k),$$

де $e(k)$ – похибка моделі.

У такому рівнянні Φ_{11} відіграє роль коефіцієнта АКФ і ЧАКФ між $y(k)$ і $y(k-1)$. Для оцінювання двох коефіцієнтів можна сформувати рівняння другого порядку

$$y'(k) = \Phi_{11} y'(k-1) + \Phi_{22} y'(k-2) + e(k),$$

де Φ_{22} – коефіцієнт ЧАКФ між $y(k)$ і $y(k-2)$.

Коефіцієнти ЧАКФ можна обчислити також за допомогою коефіцієнтів АКФ, використовуючи такі вирази [2]:

$$\Phi_{11} = r(1), \Phi_{22} = \frac{r_2 - r_1^2}{1 - r_1^2}; \quad \Phi_{ss} = \frac{r_s - \sum_{j=1}^{s-1} \Phi_{s-1,j} r_{s-j}}{1 - \sum_{j=1r}^{s-1} \Phi_{s-1,j} r_j}.$$

У загальному випадку коефіцієнти ЧАКФ стаціонарного процесу АРКС(p,q) мають наближатися до нуля, починаючи з p-го значення. АКФ процесу АРКС(p,q) починає наближатися до нуля при значеннях зміщення $s \geq q$.

Коли ми говоримо, що значення коефіцієнтів автокореляційної функції повинні відрізнятися від нуля у статистичному розумінні, це означає, що існує вираз (формула), який дозволяє підтвердити або спростувати цей факт. Одним із загальноприйнятих підходів до встановлення того факту, що коефіцієнти АКФ суттєво відмінні від нуля у статистичному розумінні, є обчислення та аналіз значущості статистичного параметра Льюнга – Бокса $Q(r_k)$ за формулою [3, 4]

$$Q(r_k) = N(N+2) \sum_{k=1}^{S} r_k^2 / (N-k),$$

де N – довжина вибірки даних змінної, для якої обчислено значення автокореляційної функції r_k; s – число коефіцієнтів АКФ, які досліджуються на суттєву відмінність від нуля. Якщо дані згенеровано процесом АР, то значення $Q(r_k)$ асимптотично мають розподіл χ^2 з s степенями вільності, а тому для перевірки їхньої значущості необхідно

користуватися відповідними статистичними таблицями. Очевидно, що більші значення вибіркової автокореляційної функції приводять до більших значень $Q(r_k)$.

Так, для ідеального процесу білого шуму $Q(r_k)=0$. Якщо значення $Q(r_k)$, обчислене за наведеним виразом, перевищує критичне значення з розподілу χ^2 з s степенями вільності, то існує щонайменше одне значення $r(k)$, яке є відмінним від нуля у статистичному розумінні. Статистику Люнга – Бокса [16] можна застосовувати також для установлення близькості залишків моделі до білого шуму. Однак слід пам'ятати, що при обчисленні s значень кореляційної функції число степенів вільності зменшується на число коефіцієнтів моделі. Таким чином, при аналізі залишків моделі $АРКС(p,q)$ статистика $Q(r_k)$ має розподіл χ^2 з $s-p-q$ степенями вільності, а з урахуванням константи – $s-p-q-1$.

Цей етап закінчується формуванням структур декількох моделей-кандидатів з векторами параметрів $\theta_1,...,\theta_m$, де m – кількість кандидатів. Кандидатів може бути декілька, оскільки встановити структуру точно за один раз, як правило, неможливо. Загалом побудова моделі високого ступеня вільності – це трудомісткий ітераційний процес, який потребує значних зусиль. На наступному етапі оцінюють параметри моделей-кандидатів.

2.5 Оцінювання коефіцієнтів моделей-кандидатів

Коефіцієнти (параметри) рівняння оцінюють, використовуючи принцип економії, або збереження. Цей принцип означає, що *кількість оцінюваних коефіцієнтів не повинна перевищувати їх необхідне число* («необхідність» можна визначити, наприклад, як потребу зберегти в моделі основні статистичні характеристики процесу).

При моделюванні процесів будь-якої природи слід пам'ятати, що поведінку процесу необхідно *апроксимувати* за допомогою рівнянь, а не намагатися описати її до найменших дрібниць. Треба враховувати також, що різні за структурою моделі можуть мати однакові властивості. Наприклад, рівняння авторегресії першого порядку

$$y(k)=0{,}5\,y(k-1)+\varepsilon(k)$$

еквівалентне ковзному середньому у вигляді

$$y(k)=\varepsilon(k)+0{,}5\varepsilon(k-1)+0{,}25\varepsilon(k-2)+0{,}125\varepsilon(k-3)+\ldots .$$

Оцінювана модель повинна задовольняти принцип інверсії, тобто щоб за допомогою отриманого рівняння можна було згенерувати початковий ряд, на основі якого оцінювались коефіцієнти. Це означає, що хоча модель і спрощена, вона має збігатися з досліджуваним процесом за такими

основними характеристиками, як середнє, дисперсія і коваріація.

У процедурі оцінювання часто використовують не абсолютні значення змінних, а їхні відхили від середнього, тобто

$$y(k) = Y(k) - \mu_y,$$

де $Y(k)$ – значення виміру; μ_y – середнє значення ряду. Якщо для оцінювання параметрів використовують рекурсивну процедуру, то поточне середнє можна обчислювати за формулою

$$\mu_y(k) = \mu_y(k-1) + \frac{1}{k}[y(k) - \mu_y(k-1)].$$

Найбільш поширеними методами оцінювання параметрів моделі є такі:

- метод найменших квадратів (МІЖ) [11];
- метод максимальної правдоподібності (ММП) [17];
- метод допоміжної (інструментальної) змінної (МДП) [18];
- нелінійний метод найменших квадратів (НМНК) [19, 20];
- та їхні рекурсивні версії (РМНК, РММП, РМДП). Оцінки (звичайного) МНК обчислюють за допомогою такого виразу:

$$\hat{\theta} = \left[X^T X\right]^{-1} X^T y,$$

де $\theta[p]$ – вектор оцінок параметрів вимірності p; $X[N \times p]$ – матриця вимірів; $Y[N]$ – вектор вимірів залежної змінної. У квадратних дужках указано вимірність векторів і матриці. Елементи матриці вимірів обчислюються по-своєму для кожної конкретної моделі. Так, для моделі

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + a_3 x_3(k) + \varepsilon(k)$$

матриця вимірів має вигляд

$$X = \begin{bmatrix} 1 & x_1(1) & x_2(1) & x_3(1) \\ 1 & x_1(2) & x_2(2) & x_3(2) \\ \dots & \dots & \dots & \dots \\ 1 & x_1(N) & x_2(N) & x_3(N) \end{bmatrix}.$$

Одиниці в першому стовпчику матриці X означають, що вимір при коефіцієнті a_0 завжди дорівнює одиниці.

Елементи матриці вимірів дещо ускладнюються при використанні поліноміальної моделі, але її також можна оцінювати за допомогою лінійних методів. Безпосереднє застосування методу мінімізації суми квадратів похибок до поліноміальної моделі порядку p приводить до формування такої матриці вимірів:

$$X' = \begin{bmatrix} N & \sum_{k=1}^{N} x(k) & \sum_{k=1}^{N} x^2(k) & \dots & \sum_{k=1}^{N} x^p(k) \\ \sum_{k=1}^{N} x(k) & \sum_{k=1}^{N} x^2(k) & \sum_{k=1}^{N} x^3(k) & \dots & \sum_{k=1}^{N} x^{p+1}(k) \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^{N} x^p(k) & \sum_{k=1}^{N} x^{p+1}(k) & \sum_{k=1}^{N} x^{p+2}(k) & \dots & \sum_{k=1}^{N} x^{2p}(k) \end{bmatrix}.$$

У такому разі векторно-матричне рівняння для N вимірів залежної і незалежної змінних можна записати так:

$$y' = X'\theta; \text{ звідси } \hat{\theta} = [X']^{-1} y',$$

де $y' = \left[\sum_{1}^{N} y(k) \ \sum_{1}^{N} x(k)y(k) \ \dots \ \sum_{1}^{N} x^p(k)y(k) \right]^T$;

$\hat{\theta}$ – вектор оцінок параметрів моделі. Тобто оцінку вектора параметрів можна знайти, розв'язавши системи лінійних (нормальних) рівнянь.

Для отримання незміщених, консистентних та ефективних оцінок параметрів θ лінійної регресійної математичної моделі, наприклад, моделі змішаної регресії

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + b_1 x(k) + b_2 z(k) + \varepsilon(k),$$

за допомогою методу найменших квадратів необхідно задовольнити такі умови:

а) $\varepsilon(k)$ – некорельована послідовність випадкових чисел із нульовим середнім, тобто

$$E[\varepsilon(k)] = 0, cov[\varepsilon(k)] = E[\varepsilon(k)\varepsilon(j)] = \begin{Bmatrix} \sigma_\varepsilon^2, k = j \\ 0, \ k \neq j \end{Bmatrix};$$

б) послідовності $\varepsilon(k)$ і $y(k)$ не повинні бути корельовані між собою.

Зазначимо, що перевірити виконання наведених умов ми можемо тільки після оцінювання коефіцієнтів моделі, а до оцінювання можна лише постулювати їх виконання. Тобто після оцінювання моделі оцінка значень випадкового процесу визначається похибками моделі

$$\hat{\varepsilon}(k) = e(k) = y(k) - \hat{y}(k),$$

що дає можливість виконати аналіз характеристик випадкового процесу $\{\varepsilon(k)\}$.

2.6 Діагностика моделей – вибір найкращої з множини оцінених кандидатів

На цьому етапі аналізується якість моделі, тобто виконується перевірка оцінених кандидатів на адекватність процесу. Діагностика

складається з таких кроків:

1. *Візуальне дослідження графіка похибок* моделі $e(k)=y(k)-\hat{y}(k)$, де $\hat{y}(k)$ – оцінка змінної, отримана за допомогою побудованої моделі. На графіку не повинно бути значних викидів і довгих інтервалів, на яких похибка має великі значення (тобто довгих інтервалів суттєвої неадекватності). У разі застосування рекурсивних методів оцінювання найбільші похибки будуть у перехідному процесі, коли інформаційна матриця ще не містить достатньо інформації про процес.

2. *Похибки моделі не повинні бути корельовані* між собою. Для аналізу наявності кореляції між значеннями похибок необхідно обчислити АКФ і ЧАКФ для ряду $\{e(k)\}$ і за допомогою Q-статистики визначити ступінь корельованості (наприклад, Q-статистика вважається несуттєвою до рівня 10%).

Крім того, корельованість похибок визначають за допомогою статистики Дарбіна – Уотсона (DW) [5, 6], яку розраховують за формулою

$$DW=2-2\rho,$$

де $\rho=E[e(k)e(k-1)]/\sigma_e^2$ – коефіцієнт кореляції між сусідніми значеннями похибки; σ_e^2 – дисперсія послідовності похибок $\{e(k)\}$. Таким чином, при повній відсутності кореляції між похибками $DW=2$ – це ідеальне значення. Гриничними значеннями для DW є 0 (при $\rho=1$) і + 4 (при $\rho=-1$).

Отримати формулу $DW=2-2\rho$ можна доволі просто. Автори цієї статистики Durbin і Watson запропонували скористатися для перевірки корельованості похибок моделі таким виразом:

$$DW=\frac{\sum_{k=2}^{N}[e(k)-e(k-1)]^2}{\sum_{k=1}^{N}e^2(k)}=\frac{\sum_{k=2}^{N}[e(k)-e(k-1)]\,[e(k)-e(k-1)]}{\sum_{k=1}^{N}e^2(k)},$$

тобто DW можна певною мірою трактувати як коефіцієнт автокореляції для їх різниць) приростів похибок.

Розкривши квадрат різниці в чисельнику, отримаємо

$$DW=\frac{\sum_{k=2}^{N}e^2(k)}{\sum_{k=1}^{N}e^2(k)}+\frac{\sum_{k=2}^{N}e^2(k-1)}{\sum_{k=1}^{N}e^2(k)}-2\frac{\sum_{k=2}^{N}e(k)e(k-1)}{\sum_{k=1}^{N}e^2(k)},$$

де $\frac{\sum_{k=2}^{N} e^2(k)}{\sum_{k=1}^{N} e^2(k)} \approx 1$, $\frac{\sum_{k=2}^{N} e^2(k-1)}{\sum_{k=1}^{N} e^2(k-1)} \approx 1$, а $\frac{\sum_{k=2}^{N} e^2(k)e(k-1)}{\sum_{k=1}^{N} e^2(k-1)} = \rho$.

Тому можна записати, що $DW = 2 - 2\rho$.

3. Для лінійної моделі 2-3 порядку *оцінки параметрів мають збігатися до усталених значень* після 30-40 (не більше) ітерацій алгоритму оцінювання. Якщо кількість ітерацій набагато перевищує указані числа, то це свідчить про те, що процес може бути нестаціонарним.

4. Перевірка значущості параметрів моделі. *Статистика Стьюдента [21],* або t*-статистика* (випадкова величина, що має t-розподіл), яка використовується для визначення значущості оцінки кожного коефіцієнта в статистичному розумінні, визначається за виразом

$$t = \frac{\hat{a} - a^0}{SE_{\hat{a}}},$$

де $\hat{a}$ – оцінка коефіцієнта моделі; a^0 – нуль-гіпотеза (початкова гіпотеза) щодо цієї оцінки; $SE_{\hat{a}}$ – стандартна похибка оцінки. За нуль-гіпотезу щодо значущості оцінки можна висувати будь-яку: як що коефіцієнт значущий $(H_0 : a^0 \neq 0)$ чи незначущий $(H_0 : a^0 = 0)$. Статистична теорія перевірки гіпотез пропонує висувати нуль-гіпотезу, яка є протилежною бажаному результату. У даному разі очікуваним результатом є значущість коефіцієнтів математичної моделі. Отже, необхідно висувати нульову гіпотезу, що коефіцієнт незначущий. Це дає можливість коректно підійти до визначення значущості оцінок коефіцієнтів і дещо спростити розрахунки.

Для того щоб установити, чи є оцінка коефіцієнта значущою, треба знати довжину вибірки даних N (потужність вибірки); число степенів вільності $f = N - n$, де n – кількість коефіцієнтів моделі, які оцінюються на основі ряду даних, і вибрати рівень значущості $a = 1\%$, або $a = 5\%$, або $a = 10\%$ (для цих значень існують розрахункові таблиці для критичних значень t-статистики). Фактично рівень значущості означає ймовірність припуститися *помилки першого роду* при перевірці гіпотези. Згадаємо, що

$$\alpha = p\{ X \in G / \omega \mid H_0 \} = \int_{n-m(G/\omega)} L_{H_0}(X)dx,$$

де $X = [x_1,...,x_n] \in R^n$ – уся вибірка, яка розбивається на дві множини, що перетинаються: ω і G / ω (ω – область прийняття нуль-гіпотези); G / ω – критична область: якщо $X \in G / \omega$, то H_0 відкидається; $L_{H_0}(X)$ –

закон розподілу X. Помилка першого роду означає відхилення правильної гіпотези.

Користуючись значеннями N, f і $a,$ з таблиць для t-розподілу знаходять критичне значення t-статистики, тобто $t_{кр}$. Для перевірки правильності висунутої гіпотези розраховане значення t порівнюють з критичним $t_{кр}$. Якщо

$$-t_{кр} < t < t_{кр} \text{ або } |t| < |t_{кр}|,$$

то нуль-гіпотеза щодо незначимості коефіцієнта приймається (його можна не враховувати в регресії). Звідси випливає, що чим більше значення t-статистики для оцінювання коефіцієнта, тим імовірніше, що цей коефіцієнт є значущим.

Загалом послідовність дій при перевірці значущості оцінок коефіцієнтів побудованої моделі можна окреслити так:

- сформулювати нуль-гіпотезу щодо значущості коефіцієнта;
- обчислити значення t-статистики для кожного коефіцієнта регресії (це виконує кожний пакет для математичного моделювання);
- за допомогою значень N, f і $a,$ знайти з таблиць для t-статистики її критичне значення;
- перевірити нуль-гіпотезу за наведеним вище простим правилом (аналіз виконання нерівності $-t_{кр} < t < t_{кр}$).

5. *Коефіцієнт множинної детермінації* R_2, який обчислюють так:

$$R^2 = \frac{var(\hat{y})}{var(y)} = 1 - \frac{SSE}{SST},$$

де $var(\hat{y})$ – дисперсія залежної змінної, оціненої за допомогою побудованої моделі; $var(y)$ – дисперсія вимірів залежної змінної; $SSE = \sum_{k=1}^{N} [y(k) - \hat{y}(k)]^2$ – сума квадратів похибок (залишків) моделі (*sum of squared errors*); $SST = \sum_{k=1}^{N} [y(k) - \bar{y}]^2$ – загальна сума квадратів (*total sum of squares*); де $\bar{y}$ – середнє значення; $SST = SSE + SSR$, де $SSE = \sum_{k=1}^{N} [\hat{y}(k) - \bar{y}]^2$ – загальна сума квадратів для регресії (*sum of squares for regression*).

Очевидно, що найкращим значенням є те, коли $R^2 = 1$, тобто коли дисперсії вимірів змінної і цієї ж змінної, оціненої за рівнянням, збігаються. Цей параметр можна трактувати також як міру інформативності моделі,

якщо вибрати мірою інформативності дисперсію. Таким чином, R^2 показує рівень інформативності моделі стосовно інформативності вибірки даних, за допомогою якої вона була

6. *Сума квадратів похибок* для вибраної моделі повинна бути мінімальною, тобто

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2 \to \min_{\hat{\theta}}.$$

7. Для оцінювання адекватності моделі також використовують *інформаційний критерій Акайке* [22, 23]

$$AIC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n$$

і *критерій Байсса – Шварца* [24]

$$BSC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + n \ln(N),$$

де $n = p + q + 1$ – число параметрів моделі, які оцінюються за допомогою статистичних даних (p – число параметрів авторегресійної частини моделі; q – число параметрів ковзного середнього; 1 з'являється тоді, коли оцінюється зміщення (або перетин), тобто a_0).

Критерії Акайке і Байєса – Шварца містять у правій частині суму квадратів похибок, а тому за цими критеріями вибирають ту модель, для якої критерії набувають найменших значень. Уведення нового регресора приводить до збільшення критерію (при цьому збільшується n), але разом з тим зменшується сума квадратів похибок і критерій у цілому зменшується. Якщо регресор не покращує модель, то критерій збільшується. Слід також зазначити, що асимптотичні властивості для довгих вибірок кращі в критерії Байєса – Шварца, тобто його рекомендують застосовувати при відносно великих значенні $N (N > 100)$.

8. Окрім згаданих параметрів, для визначення адекватності моделі в цілому використовують F*-статистику Фішера* [25, 26], яка пропорціональна відношенню

$$F \sim \frac{R^2}{1 - R^2},$$

а для множинної (багатофакторної) регресії вона визначається за формулою

$$F = \frac{R^2}{1 - R^2} \cdot \frac{(N - p - 1)}{p},$$

де, як і раніше, N – кількість значень ряду; p – кількість параметрів моделі

без урахування перетину (константи).

Таким чином, якщо $R^2 \to 1$, то $F \to \infty$. Порядок застосування F-статистики такий самий, як і для t-статистики. Нуль-гіпотезою є у цьому разі припущення про те, що модель неадекватна в цілому, тобто

$$H_0 : a_1 = a_2 = ... = a_p = 0.$$

Проте альтернативна гіпотеза H_1 така: *хоча б одне значення* a_i *відмінне від нуля у статистичному розумінні*.

Значення $F_{крит}$ знаходять з таблиць для F-розподілу. Послідовність застосування цієї статистики можна подати таким чином:

1. Сформулювати нуль-гіпотезу щодо адекватності моделі в цілому Наприклад, H_0: *модель неадекватна в цілому* (або $H_0 : a_1 = a_2 = ... = a_p = 0$).

2. Розрахувати значення для оціненої моделі (як правило, воно розраховується усіма пакетами статистичного оброблення даних).

3. Задати рівень значущості $a = 1\%$, або $a = 5\%$, або $a = 10\%$.

4. Користуючись значеннями N, f і a, знайти критичне значення $F_{крит}$ з таблиць для F-розподілу при (p, $N - p - 1$) степенях вільності.

5. Перевірити нуль-гіпотезу: якщо $F > F_{крит}$, то нуль-гіпотеза щодо неадекватності моделі в цілому відкидається на вибраному рівні значущості.

Коректне застосування методики Бокса – Дженкінса забезпечує побудову адекватної математичної моделі процесу, якщо експериментальні дані відповідають вимогам **представництва** та **інформативності**. Перша вимога означає, що вибірка даних повинна охоплювати тривалий проміжок часу, щоб повністю відображати поведінку режиму функціонування процесу, для якого будується модель. Вимога інформативності означає, що вибірка має містити в собі обсяг інформації, достатній для оцінювання коефіцієнтів моделі.

Наприклад, якщо моделюється процес другого порядку, то вибірка має забезпечувати коректне обчислення першої і другої похідних. Іноді формально інформатнвність оцінюють також за допомогою величини дисперсії процесу, а також за вмістом гармонічних складових. Чим більше гармонік містить вибірка, тим інформативнішою вона є.

Умову інформативності даних пов'язують з умовою *достатнього збудження* процесу. Достатнє збудження означає, що вхідний сигнал повинен охоплювати всю смугу частот, які може пропускати на вихід процес (об'єкт). Тобто вхідний сигнал має охоплювати всю амплітудно-частотну характеристику процесу. Ця вимога залишається однаковою для процесів будь-якої природи.

2.8 Запитання і вправи

1) Назвіть етапи пообудови математичних моделей за методикою Бокса – Дженкінса.

2) Яка мета аналізу функціонування процесу? Чи можна його пропустити при побудові математичної моделі?

3) Яка мета попереднього оброблення даних? Назвіть основні операції, які виконують у процесі попереднього оброблення даних. До чого приводить визначення значень змінних у великому діапазоні?

4) Які два основних типи нелінійностей трапляються в аналізі часових рядів?

5) Яким чином можна встановити наявність нелінійностей у процесі?

6) Який метод дає можливість автоматизувати процес визначення і врахування нелінійностей процесу?

7) Яку інформацію можна отримати на основі візуального аналізу даних? Як можна нею скористатися?

8) Яким чином можна знайти оцінку порядку авторегресійної частини моделі?

9) У чому полягає відмінність між автокореляційною і частковою автокореляційною функціями процесу?

10) На чому ґрунтується відбір незалежних змінних (регресорів, екзогенних змінних) для включення в праву частину математичної моделі?

11) Назвіть три умови коректного застосування методу найменших квадратів до оцінювання параметрів математичної моделі.

12) У чому полягає принципова різниця між методом найменших квадратів (МНК), призначеним для оцінювання лінійних моделей, і МНК для оцінювання моделей, нелінійних відносно параметрів? Які критерії якості мінімізують ці методи?

13) Про що свідчить корельованість похибок моделі між собою? За допомогою якої статистики можна визначити ступінь корельованості похибок? Яких значень набуває ця статистика в ідеальному випадку?

14) Що означають помилки першого і другого роду при перевірці статистичних гіпотез?

3 ЗАСТОСУВАННЯ РІЗНИЦЕВИХ РІВНЯНЬ ДО ОПИСАННЯ СТАТИСТИЧНИХ ДАНИХ

3.1 Загальні відомості про різницеві рівняння

При використанні дискретних рівнянь незалежну змінну час t замінюють дискретним часом, тобто вважають $t = kT_S$, де T_S – період дискретизації вимірів, який у техніці набуває значень від десятків мікросекунд до десятків секунд і навіть хвилин, а при моделюванні фінансово-економічних процесів – від декількох хвилин до одного року, залежно від того, яку статистику можна отримати. Період дискретизації, як правило, нормують до одиниці і незалежною змінною залишається k (дискретний час), що набуває цілих значень від **0** до ∞. При цьому для кожної прикладної задачі дискретна одиниця часу має відповідне реальне значення.

Завдяки простоті структури та наявності надійних методів оцінювання параметрів різницеві рівняння (РР) набули широкого застосування при моделюванні процесів у техніці, економіці та фінансах, екології, біології та інших прикладних і наукових галузях. Простим прикладом різницевого рівняння є стохастичне рівняння авторегресії першого порядку з одиничним коефіцієнтом (окремий випадок) при затриманому в часі значенні основної змінної:

$$y(k) = y(k-1) + \varepsilon(k),$$

де $y(k)$ – основна змінна; $\varepsilon(k)$ – випадкова величина, яка відображує вплив різноманітних невимірюваних факторів на основну змінну. У першу чергу це випадкові збурення, що діють на процес. Найчастіше припускають, що випадкова величина має нормальний розподіл: $\{\varepsilon(k)\} \approx N_n(0, \sigma_\varepsilon^2)$, тобто вона має нульове середнє і скінченну дисперсію σ_ε^2. Наприклад, якщо $y(k)$ – ціни в деякі біржові акції в k-й день, то $\varepsilon(k)$ характеризує коливання ціни під впливом багатьох факторів, які неможливо ввести в модель, оскільки неможливо отримати їхні виміри або ж ці впливи мають якісний характер.

Як правило, у наведеній моделі процес $\{\varepsilon(k)\}$ має такі обмеження:

$$E[\varepsilon(k)] = 0, E[\varepsilon(k)\varepsilon(l)] = \begin{cases} \sigma_\varepsilon^2, k = l, \\ 0, \quad k \neq l. \end{cases}$$

Докладніше роль випадкової змінної розглядатиметься окремо в кожному конкретному випадку.

За його допомогою описують, наприклад, ціну акції на біржі в момент часу, що відповідає аргументові k. Його називають ще рівнянням, яке описує *процес випадкового кроку* (випадкове блукання або *random walk*). Таку назву воно отримало з тієї причини, що приріст значення основної

змінної визначається фактично випадковою змінною. Воно може бути записане також у формі першої різниці

$$\Delta y(k) = \varepsilon(k),$$

де $\Delta y(k) = y(k) - y(k-1).$

Для рівняння випадкового кроку можна записати однорідне рівняння

$$y(k) - y(k-1) = 0,$$

яке має характеристичне рівняння $a - 1 = 0$ (як отримати характеристичне рівняння у загальному випадку, ми розглянемо нижче в цьому розділі). Таким чином, дане характеристичне рівняння має один корінь $a = 1$. Якщо характеристичне рівняння має хоча б один одиничний корінь, то кажуть, що відповідне йому різницеве рівняння описує *процес з одиничним коренем*. Нижче покажемо, що процеси з одиничними коренями – це *процеси з трендами* або *інтегровані процеси*. Надалі під трендом розумітимемо загальний довгостроковий напрям розвитку процесу; фактично він збігається з поточним середнім значенням.

Більш загальною формою різницевого рівняння є така:

$$\Delta y(k) = \alpha_0 + \alpha_1 y(k-1) + \varepsilon(k), \qquad (3.1)$$

але для того щоб воно відповідало процесу випадкового кроку, необхідно покласти $a_0 = a_1 = 0$, інакше це рівняння уже не відповідатиме своєму визначенню. Зазначимо, що порядок авторегресійної частини різницевого рівняння визначається кількістю минулих (затриманих) вимірів залежної змінної, які використовуються в його правій частині для пояснення її зміни в часі.

Різницеві рівняння, у правій частині яких наявні минулі (затримані) виміри основної змінної, називаються авторегресійними (АР), тобто регресією змінної на саму себе. Рівняння авторегресії i-го порядку має вигляд

$$y(k) = a_0 + \sum_{i=1}^{n} a_i y(k-1) + \varepsilon(k). \qquad (3.2)$$

Якщо для процесу, що моделюється, можна виявити вхідну змінну, то вона записується у правій частині рівняння авто регресії, тобто

$$y(k) = a_0 + \sum_{i=1}^{n} a_i y(k-1) + \sum_{j=1}^{q} b_j x(k-j) + \varepsilon(k).$$

Якщо $x(k)$ – випадковий процес, то таке рівняння називається авторегресією з ковзним середнім (АРКС). У разі ковзного середнього необхідно також виконати таку умову: $\sum_{j=1}^{q} b_j = 1.$

Характеристичне рівняння, записане для (3.1), може мати одиничні корені, тобто один або більше коренів характеристичного рівняння можуть

мати значення «1». Такі процеси називаються процесами авторегресії з інтегрованим ковзним середнім $АРІКС\ (p,d,q)$, де d – кількість одиничних коренів характеристичного рівняння. Процеси цього класу *нестаціонарні* – вони мають тренд, порядок якого визначається кількістю одиничних коренів. Якщо $d=1$, то тренд лінійний; якщо $d=2$, то тренд квадратичний і т. д. Таким чином, *процеси з трендом, інтегровані процеси і процеси з одиничними коренями* – це назви процесів, що мають тренд.

Застосування перших різниць і різниць вищих порядків

Разом з основною змінною для описання процесів використовують перші та другі різниці, наприклад:

$$\Delta y(k)=y(k)-y(k-1),$$

$$\Delta y(k+1)=y(k+1)-y(k),$$

$$\Delta y(k+2)=y(k+2)-y(k+1).$$

Наведені перші різниці відображують швидкість зміни основної змінної, що відповідає першій похідній для рівнянь, записаних у неперервному часі, тобто для диференціальних рівнянь.

Обчислення перших різниць приводить до вилучення лінійного тренду з процесу. Наприклад, нехай тренд описується лінійним рівнянням

$$y(k)=a_0+a_1k.$$

Перша різниця для цього процесу

$$\Delta y(k)=y(k)-y(k-1)=a_0+a_1k-a_0-a_1(k-1)=a_1,$$

тобто після дискретного диференціювання залишилась константа.

Другі різниці відображують швидкість зміни в часі перших різниць (тобто прискорення) і їх записують так:

$$\Delta^2 y(k)=\Delta(\Delta y(k))=\Delta[y(k)-y(k-1)]=[y(k)-y(k-1)]-[y(k-1)-y(k-2)]=$$

$$=y(k)-2y(k-1)+y(k-2);$$

$$\Delta^2 y(k+1)=\Delta(\Delta y(k+1))=y(k+1)-2y(k)+y(k-1).$$

Останній вираз називають різницевою схемою другої похідної. На практиці другі різниці використовують досить рідко, а різниці вищого порядку не використовуються. Застосування других різниць до процесу приводить до вилучення квадратичного тренду, що можна легко проілюструвати на прикладі описання тренду поліномом другого порядку від k. Наприклад, нехай

$$y(k)=a_0+a_1k+a_2k^2.$$

Перші різниці мають вигляд

$$\Delta y(k)=a_0+a_1k+a_2k^2-[a_0+a_1(k-1)+a_2(k-1)^2]=$$

$$=a_0+a_1k+a_2k^2-a_0-a_1k+a_1-a_2k^2+2a_2k-a_2=$$

$$=a_1+2a_2k-a_2,$$

другі різниці – $\Delta^2 y\ (k) = a_1 + 2a_2 k - a_2 - [a_1 + 2a_2(k-1) - a_2] = 2a_2$.

Приклад 3.1. Знайдемо загальний розв'язок РР другого порядку:

$$y(k) = 2{,}1 + 0{,}8y(k-1) - 0{,}15y(k-2).$$

Сформуємо однорідне рівняння:

$$y(k) - 0{,}8y(k-1) + 0{,}15y(k-2) = 0.$$

Оскільке це рівняння другого порядку, то воно має два однорідних розв'язки, які знаходять за допомогою розв'язку характеристичного рівняння, записаного для однорідного (більш детально цю методику розглянуто нижче). Характеристичне рівняння має вигляд

$$\lambda^2 - 0{,}8\lambda + 0{,}15 = 0.$$

Це квадратне рівняння має два корені: $\lambda_1 = 0{,}5,\ \lambda_2 = 0{,}3$. Таким чином, два розв'язки можна записати як

$$y_1^h(k) = A_1(0{,}5)^k;\ y_2^h(k) = A_2(0{,}3)^k.$$

Правильність знаходження однорідних розв'язків можна перевірити шляхом їх підстановки в однорідне рівняння. Наприклад, підставимо перший однорідний розв'язок $y_1^h(k) = A_1(0{,}5)^k$ в однорідне рівняння

$$A_1(0{,}5)^k - 0{,}8A_1(0{,}5)^{k-1} + 0{,}15A_1(0{,}5)^{k-2} = 0.$$

Якщо розділити всі члени на $A_x\ A_1(0{,}5)^{k-2}$, то отримаємо

$$(0{,}5)^2 - 0{,}8(0{,}5) + 0{,}15 = 0{,}25 - 0{,}40 + 0{,}15 = 0.$$

Другий однорідний розв'язок перевіряється аналогічно.

Частковий розв'язок для детермінованого збурення знайдемо так:

$$y^p = 2{,}1 + 0{,}8y^p - 0{,}15y^p.$$

Звідси $y^p = 6{,}0$, де верхній індекс $p = partial$ (частковий). Загальне визначення часткового розв'язку наведено нижче. Тепер об'єднаємо і частковий розв'язки:

$$y(k) = 6{,}0 + A_1(0{,}5)^k + A_2(0{,}3)^k,$$

де A_1, A_2 – довільні константи, які знайдемо за допомогою початкових умов. Оскільки маємо дві невідомі константи, то необхідно мати дві умови. Нехай $y(0) = 1{,}0;\ y(1) = 2{,}0.$ Тепер можемо записати систему з двох рівнянь для констант:

$$k = 0 : 1{,}0 = 6{,}0 + A_1 + A_2,$$

$$k = 1 : 2{,}0 = 6{,}0 + 0{,}5A_1 + 0{,}3A_2.$$

Розв'язуючи систему, знайдемо, що $A_1 = -12{,}5; A_2 = 7{,}5$, і розв'язок набуває вигляду

$$y(k) = 6{,}0 - 12{,}5(0{,}5)^k + 7{,}5(0{,}4)^k.$$

Таким чином, знайдений розв'язок має середнє значення (його називають ще зміщенням) 6,5, а два інші члени наближаються до нуля при $k \to \infty$.

Приклад 3.1. Для порівняння дискретної моделі з неперервною розглянемо розв'язок диференціального рівняння другого порядку, яке описує вимушені коливання

$$\ddot{y}(t) + 2h\dot{y}(t) + r^2 y(t) = x(t), \qquad (3.3)$$

де h – коефіцієнт, що характеризує властивості опору середовища (демпфірування); r – коефіцієнт жорсткості. Нехай збуджувальна сила змінюється за гармонічним законом $x(t) = M\sin(\omega t)$, а опір середовища покладемо для простоти нульовим, $h = 0$. Тепер рівняння (3.2) набуває вигляду

$$\ddot{y}(t) + r^2 y(t) = M\sin(\omega t). \qquad (3.4)$$

Власні коливання тіла визначаються однорідним диференціальним рівнянням

$$\ddot{y}(t) + r^2 y(t) = 0 \qquad (3.5)$$

й описуються рівнянням гармонічних коливань $\ddot{y}(t) = A\sin(rt + \varphi)$.

Необхідно знайти частковий розв'язок рівняння (3.3). Вигляд часткового розв'язку залежить від того, чи буде число $a + ib = i\omega$, $i = \sqrt{-1}$ коренем характеристичного рівняння для рівняння (3.3). Оскільки характеристичне рівняння $\lambda^2 + r^2 = 0$ має корені $\lambda_{1,2} = \pm ir$, то аналіз зводиться до випадків: коли частота впливу збуджуюча сили збігається з частотою власних коливань тіла – це резонанс і коли не збігається – це відсутність резонансу.

Розглянемо нерезонансний випадок $(\omega \neq r)$. Згідно з теорією диференціальних рівнянь частковий розв'язок рівняння (3.4) слід шукати за допомогою пробної функції

$$y^p(t) = A\cos(\omega t) + B\sin(\omega t), \qquad (3.6)$$

де A і B – коефіцієнти, які потрібно визначити.

З останнього рівняння отримаємо:

$$y^p(t) = A\cos(\omega t) + B\sin(\omega t),$$

$$\dot{y}^p(t) = -A\omega\sin(\omega t) + B\omega\cos(\omega t),$$

$$\ddot{y}^p(t) = -A\omega^2\cos(\omega t) + B\omega^2\sin(\omega t).$$

Підставивши другу похідну в (3.4), дістанемо

$$A\left(r^2-\omega^2\right)\cos(\omega t)+B\left(r^2-\omega^2\right)\sin(\omega t)=M\sin(\omega t).$$

З початкових умов знайдемо невідомі коефіцієнти:

$$\left.\begin{aligned}A\left(r^2-\omega^2\right)&=0\\B\left(r^2-\omega^2\right)&=M\end{aligned}\right\}\Rightarrow A=0,\ B=\frac{M}{r^2-\omega^2}.$$

Частковий розв'язок:

$$y^p(t)=\frac{M}{r^2-\omega^2}\sin(\omega t). \tag{3.7}$$

Загальний розв'язок має вигляд

$$y^p(t)=\frac{M}{r^2-\omega^2}\sin(\omega t)+A\sin(rt+\varphi). \tag{3.8}$$

Таким чином, коливання, наведені в рівнянні (3.7), визначаються лінійною комбінацією гармонічних коливань різної частоти.

Резонансні коливання. Явище резонансу виникає, якщо частота зміни збурюючої сили збігається з частотою власних коливань тіла $(\omega=r)$. Рівняння коливань має вигляд

$$\ddot{y}(t)+r^2y(t)=M\sin(rt). \tag{3.9}$$

У цьому разі частковий розв'язок міститиме множник t:

$$y^p(t)=t\left[A\cos(rt)+B\sin(rt)\right]. \tag{3.10}$$

Отримуємо такий результат:

$$y^p(t)=t\left[A\cos(rt)+B\sin(rt)\right],$$

$$\dot{y}^p(t)=-A\cos(rt)+B\sin s(rt)+r\left[-Ar\sin(rt)+Br\cos(rt)\right],$$

$$\ddot{y}^p(t)=2\left[-Ar\sin(rt)+Br\cos(rt)\right]+t\left[-Ar^2\cos(rt)-Br^2\sin(rt)\right],$$

а звідси –

$$-2Ar\sin(rt)+2Br\cos(rt)=M\sin(rt),$$

$$\left.\begin{aligned}-2Ar&=M\\2Br&=0\end{aligned}\right\}\Rightarrow A=-\frac{M}{2r},\ B=0.$$

Неоднорідний розв'язок має вигляд

$$y^t(t)=\frac{M}{2r}t\cos(rt). \tag{3.11}$$

Частковий розв'язок (3.11) являє собою вимушені коливання з необмеженою (розбіжною) амплітудою. Гармонічні коливання лежать між прямими

$$y(t)=\frac{M}{2r}t \quad i \quad y(t)=-\frac{M}{2r}t.$$

Загальний розв'язок має вигляд

$$y(t)=-\frac{M}{2r}t\cos(rt)+A\sin(rt+\varphi)$$

являє собою лінійну комбінацію коливань з необмеженою амплітудою і гармонічних коливань з постійною амплітудою тієї ж частоти.

Приклад 3.2. Знайдемо повний розв'язок рівняння третього порядку АРКС(3,2):

$$y(k)=0{,}5+y(k-1)+0{,}25\,y(k-2)-0{,}25\,y(k-3)+\varepsilon(k)-0{,}125\,\varepsilon(k-1)+$$
$$+0{,}125\,\varepsilon(k-2).$$

Запишемо характеристичне рівняння

$$\lambda^3-\lambda^2-0{,}25\lambda+0{,}25=0$$

і знайдемо його корені: $\lambda_1=1;\ \lambda_2=-0{,}5;\ \lambda_3=0{,}5$ (процес з єдиним коренем). Однорідний розв'язок має вигляд

$$y^h(k)=A_1+A_2(-0{,}5)^k+A_3(0{,}5)^k.$$

Для знаходження часткового розв'язку виберемо пробну функцію у вигляді

$$y^p_{проб}(k)=b_0+b_1k+\sum_{i=0}^{\infty}a_i\varepsilon(k-i)$$

і підставимо її в рівняння процесу, скориставшись такими загальними позначеннями для коефіцієнтів:

$$b_0+b_1k+a_0\varepsilon(k)+a_1\varepsilon(k-1)+a_2\varepsilon(k-2)+a_3\varepsilon(k-3)+\ldots=$$
$$=a_0+a_1\left[b_0+b_1(k-1)+a_0\varepsilon(k-1)+a_1\varepsilon(k-2)+a_2\varepsilon(k-3)+a_3\varepsilon(k-4)+\ldots\right]+$$
$$+a_2\left[b_0+b_1(k-2)+a_0\varepsilon(k-2)+a_1\varepsilon(k-3)+a_2\varepsilon(k-4)+a_3\varepsilon(k-5)\ldots\right]+$$
$$+a_3\left[b_0+b_1(k-3)+a_0\varepsilon(k-3)+a_1\varepsilon(k-4)+a_2\varepsilon(k-5)+a_3\varepsilon(k-6)+\ldots\right]+$$
$$+\varepsilon(k)+\beta_1\varepsilon(k-1)+\beta_2\varepsilon(k-2).$$

Прирівняємо константи в лівій і правій частинах, а також коефіцієнти при однакових змінних:

$b_0=a_0+a_1b_0-a_1b_1+a_2b_0-2a_2b_1+a_3b_0-3a_3b_1;$

при k $\quad b_1=a_1b_1+a_2b_1+a_3b_1;$

при $\varepsilon(k)$ $\quad a_0=1;$

при $\varepsilon(k-1)$ $\quad a_1=a_1a_0+\beta_1;$

при $\varepsilon(k-2)$ $\quad a_2=a_1a_1+a_2a_0;$

при $\varepsilon(k-3)$ $\quad a_3=a_1a_2+a_2a_1+a_3a_0.$

Таким чином, $b_1\approx 0{,}67$, а коефіцієнт b_0 необхідно визначити за допомогою початкових мов. Повний розв'язок такий:

$$y(k) = A + A_2(-0{,}5)^k + A_3(0{,}5)^k + 0{,}67k + \sum_{i=0}^{\infty} a_i \varepsilon(k-i),$$

де $A = A_1 + b_0$, $a_i = a_{i-1} + 0{,}25a_{i-2} - 0{,}25a_{i-3}$.

3.2 Запитання і вправи

1) Для чого необхідно знаходити розв'язки різницевих рівнянь? Чи можна скористатися моделлю ***АРКС***(p,q) для асимптотичного аналізу поведінки процесу?

2) Який процес називають процесом з одиничними коренями? Які синоніми цієї назви? Дайте визначення тренду.

3) На яку частину розв'язку впливають початкові умови? У чому проявляється вплив початкових умов?

4) Що характеризують перші і другі різниці? Запишіть рівняння АР(2) через перші різниці. Для розв'язання якої задачі можна скористатися цією моделлю?

5) У якому випадку неоднорідний розв'язок дорівнює нулю?

6) Запишіть загальний вигляд однорідного рівняння для випадку, коли характеристичне рівняння має трикратний корінь. Чи буде такий розв'язок стійким (збіжним)? Дайте пояснення.

4 ПРОГНОЗУВАННЯ ДИНАМІКИ РОЗВИТКУ ПРОЦЕСІВ ЗА ДОПОМОГОЮ РІЗНИЦЕВИХ РІВНЯНЬ

4.1 Для чого потрібні прогнози?

Прогнозування подій у процесі приватної і ділової активності – це суттєва складова нашого повсякденного життя, яка стосується кожного. Навряд чи знайдеться такий напрям людської діяльності, де не треба робити прогнозів. Для того щоб продемонструвати важливість коректного розв'язання задачі прогнозування, розглянемо ієрархію процесу прийняття рішень на рівні виробничої, торговельної або іншої компанії. В управлінському процесі, що стосується діяльності цих компаній, можна виділити такі рівні ієрархії планування:

1. Стратегічне бізнес-планування. На цьому рівні визначаються стратегічні цілі функціонування організації (фірми, підприємства) на тривалому часовому інтервалі: від 3 до 20 років. План складається по *місяцях, кварталах або роках*. Основними цілями, як правило, є такі: отримання певного рівня прибутку; досягнення необхідного рівня якості продукції; визначення своєї участі (частки) на ринку конкретних товарів; задоволення потреб покупців; досягнення певного рівня автоматизації виробництва; досягнення необхідного рівня відносин між роботодавцем і найманими робітниками.

2. Середньострокове планування управлінської діяльності. На цьому рівні визначаються управлінські дії, потрібні для досягнення стратегічних цілей на часовому інтервалі від 3 місяців до 3 років. План складається по *тижнях* або по *місяцях*. Планування стосується необхідних фінансів; маркетингу; обсягів продажу, перевезення (розповсюдження) продукції; кількості, місцезнаходження і характеру необхідного устаткування; агрегування (концентрації) виробництва – яка конкретна продукція вироблятиметься на конкретних підрозділах підприємства; формування основних календарних планів.

3. Оперативне планування і управління. Головна мета цього рівня управління полягає у тому, щоб ефективно розподілити ресурси для досягнення наведених вище цілей. Оперативне планування та управління реалізується на часових інтервалах від одного місяця до двох років; план складається по *тижнях* або по *місяцях*. Воно стосується управління і контролю розподілу матеріальних ресурсів, виробництва та розподілу продукції, необхідних виробничих потужностей (приміщення, устаткування, робоча сила).

4. Оброблення транзакцій (щоденне управління). На цьому рівні реалізуються управлінські операції на часовому інтервалі від 1 дня до 1 місяця. При цьому контролюються такі операції: закупівля необхідних матеріалів: формування замовлень; виставлення і надходження рахунків; оперативний контроль якості продукції; управління транспортним

господарством підприємства.

Докладний аналіз кожного рівня прийняття рішень на підприємстві показує, що управлінський персонал не може ефективно виконувати планування і контроль без належного *прогнозування* показників розвитку підприємства та інших процесів, пов'язаних з діловою активністю. При цьому горизонт прогнозу (проміжок часу, на який робиться прогноз) залежить від рівня ієрархії і може бути тривалим (до 20 років) або коротким (один день). Таким чином, ієрархія процесу прийняття рішень потребує створення і використання ієрархічної системи прогнозування.

Необхідно зазначити, що рух вниз по управлінській ієрархії супроводжується суттєвими змінами характеру інформації і її докладності. Так, на вершині піраміди необхідно приймати *неструктуровані* та *недостатньо структуровані* рішення при неповній інформації (або її відсутності). Чим нижчий рівень управління, тим більш програмованими (структурованими, упорядкованими, прозорими) є рішення, для яких існує інформація у дедалі повнішому обсязі.

Наявність на підприємстві комп'ютеризованої системи підтримки прийняття рішень уможливлює автоматизацію процесів управління і контролю. Наприклад, автоматизуються операції обліку на складах, контролю якості, аналізу номенклатури обладнання, формування технологічних карт для реалізації технологічних процесів, накопичення та підтримки необхідних запасів сировини тощо. У цілому, рухаючись униз по ієрархії прийняття рішень, можна сформулювати такі висновки:

- рішення і прогнози стають менш стратегічними;
- зменшується участь у прийнятті рішень управлінського персоналу верхнього рівня;
- скорочується горизонт прогнозування від років до місяців, тижнів і днів;
- рішення і прогнози стають більш програмованими (стандартизованими, структурованими, чітко визначеними);
- значно зростає докладність описання ситуацій і завдань, щодо яких приймаються рішення;
- зростає ступінь автоматизації процесу прийняття рішень;
- для прийняття рішень, а особливо тих, що повторюються, доцільно застосовувати відповідні комп'ютерні системи підтримки прийняття рішень при прогнозуванні і плануванні діяльності.

Зі сказаного випливає, що прогнозуюча система також повинна бути побудована за ієрархічною архітектурою (структурою). Такий підхід дає можливість структуризувати (упорядкувати) задачі прогнозування, створити множину необхідних для прогнозування спеціалізованих моделей, ситуацій і методів прийняття рішень на кожному рівні, а також наблизити комп'ютеризований процес прийняття рішень до того, що є звичним для нас.

Крім бізнес-процесів досить часто постає проблема оцінювання прогнозу при прийнятті приватних рішень. Наприклад, необхідно

спрогнозувати сімейний бюджет на 4-5 років з урахуванням оплати навчання в навчальному закладі та виплат за житловими кредитами або спрогнозувати затрати часу на виконання різних робіт залежно від їхньої складності.

4.2 Які складові процесу можна прогнозувати?

У загальному випадку прогноз можна подати одним (точковим) значенням змінної, інтервалом, у який потрапляє випадкова змінна, а також імовірністю того змінна (чи подія) матиме певне значення у вибраному інтервалі. Якщо для описання процесу застосовують лінгвістичні змінні, то прогнозом буде нечітке значення, але його також можна перетворити в чітке число.

Можна по-різному ставити задачу – прогнозування залежно від рівня приняття рішення і конкретної задачі управління чи контролю. Прогнозування може стосуватися таких складових процесу:

- дермінованого тренду як індикатора довгострокових змін процесу;
- випадкового (нерегулярного) тренду як показника коротко – і седньострокових змін;
- короткострокових змін, тобто прогнозування коливань (відхилень), що надаються на тренд;
- сезонних ефектів;
- приростів (швидкості) зміни процесу, які визначаються першими різницями;
- десперсії або стандартного відхилу як міри розсіювання процесу;
- якісних змінних (за допомогою нечітких множин, мереж Байєса тощо);
- комбінацій указаних елементів процесів.

Відповідно до того які складові процесу необхідно прогнозувати, ставиться задача побудови математичної, ймовірнісної (статистичної) або логічної моделі, що має на меті забезпечити високу якість прогнозу на заданий горизонт. Розглянемо деякі можливості математичного опису складових процесів різної природи.

Детермінований тренд

Якщо описати детермінований тренд за допомогою полінома від часу k -го порядку

$$y(k) = a_0 + a_1 k + a_2 k^2 + ... + a_p k^p + \varepsilon(k),\ E[\varepsilon(k)] = 0,$$

то визначення прогнозу тренду зводиться до підстановки в це рівняння бажаного значення часу k і застосування безумовного математичного сподівання. Прогнозування значень тренду вважається довгостроковим прогнозом, оскільки детермінований тренд указує на довгострокові зміни процесів. Обмеження на випадковий процес $K[\varepsilon(k)] = 0$ необхідне для коректного застосування методів оцінювання моделей, а також для подальшого аналізу якості оцінок прогнозів.

Стохастичний тренд

Для описання і прогнозування стохастичного тренду можна скористатись, наприклад, відомим рівнянням випадкового кроку з перетином (константою)

$$y(k) = a_0 + y(k-1) + \varepsilon(k),$$

розв'язок якого має вигляд

$$y(k) = y_0 + ka_0 + \sum_{i=1}^{k} \varepsilon(i).$$

Сума $\sum_{i=1}^{k} \varepsilon(i)$ у правій частині останнього рівняння описує випадкову складову тренду. Цю складову називають ще *нерегулярною*.

Прогнозування коливань, що накладаються на тренд

Коливання, що накладаються на тренд, або короткострокові зміни можна описати рівняннями авторегресії з ковзним середнім

$$y(k) = \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{q} b_i \varepsilon(k-j) + \varepsilon(k).$$

Техніку обчислення такого прогнозу за допомогою умовного та безумовного математичного сподівання ми розглянемо нижче в цьому розділі.

При переході до перших різниць і різниць вищих порядків з процесу вилучається тренд відповідного порядку. Наприклад, якщо процес містить лінійний тренд, то перші різниці вилучають його і після переходу до різниць ми маємо справу з коливаннями, що накладаються на тренд.

Можливість прогнозування *сезонних ефектів* досягається за рахунок введення у модель процесу відповідних значень основної змінної із затримками (лагами), що відповідають періодичності ефекту. Як буде показано нижче, сезонний ефект можна враховувати як за допомогою основної змінної, так і за допомогою процесу ковзного середнього.

Прогнозування дисперсії

Якщо дисперсія процесу змінюється у часі, то для її описання можна вибрати рівняння для формування відповідної функції прогнозування та обчислення у подальшому оцінки прогнозованого значення. Процеси зі змінною дисперсією, що мають назву гетероскедастичних, розглядатимуться в іншому розділі. Методика побудови моделей гетероскедастичних процесів передбачає такі кроки:

- математичне описання самої змінної рівнянням авторегресії невисокого порядку (наприклад, першого);
- математичне описання умовної дисперсії як динамічної змінної за допомогою рівняння прийнятної (за якістю прогнозу) структури.

4.3 Умовні та безумовні статистичні характеристики

При виконанні статистичного аналізу випадкових процесів використовують *умовні* та *безумовні статистичні характеристики*. Зокрема, для знаходження короткострокових і довгострокових прогнозів розвитку процесів застосовують умовне E_k і безумовне E математичне сподівання. У визначенні цих характеристик існують відмінності.

Безумовні статистичні характеристики – це константи, які розглядають і оцінюють на довільних часових інтервалах, не накладаючи умов на змінні функції та обсяги інформації, необхідної для визначення цих характеристик.

Тобто інформація для обчислення дисперсії вважається відомою на всьому інтервалі, що розглядається у процесі аналізу.

Так, безумовне математичне сподівання використовують для знаходження довгострокових прогнозів або умов економічної рівноваги. Наприклад, безумовне вибіркове середнє і дисперсія обчислюються за відомими формулами

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} y(k),\ var(x) = \frac{1}{N-1}\sum_{i=1}^{N}\left[x(k) - \bar{x}\right]^2,$$

а безумовне математичне сподівання позначається символом E. У виразі для дисперсії немає необхідності зменшувати значення N на одиницю, якщо потужність ряду досить велика, наприклад декілька десятків.

Умовні статистичні характеристики в аналізі динаміки процесів, поданих часовими рядами, – це *функції часу*, які визначаються на кожний конкретний момент часу k. При цьому для їх обчислення необхідно, щоб на вибраний (заданий) момент була інформація щодо значень змінних і функцій, необхідних для використання відповідних обчислень.

Умовне математичне сподівання застосовують для визначення короткострокових і середньострокових прогнозів. Умовну дисперсію процесу і стандартний відхил також часто використовують як міру ризику, наприклад, при аналізі фінансових процесів, формуванні правил торгівлі на біржі, аналізі банківських та економічних ризиків. У технічних системах дисперсія також відіграє звичну роль при визначенні ступеня відхилення вузла (наприклад, підшипника, корпусного елемента, крила) від нормального (заданого) стану. Тому вміння правильно аналітично описати дисперсію дає можливість з високою точністю описати і спрогнозувати значення відхилень від норми.

Так, умовне математичне сподівання стохастичного процесу АР(1) визначається за виразом

$$E_k\left[y(k+1) \mid y(k), y(k-1), ..., y(0), \varepsilon(k), \varepsilon(k-1), ..., \varepsilon(0)\right] =$$
$$E_k\left[a_0 + a_1 y(k) + \varepsilon(k+1)\right] = a_0 + a_1 y(k)$$

за умови, що $E_k\left[\varepsilon(k+1)\right] = 0, l \geq 1$.

Умовне вибіркове середнє та умовну вибіркову дисперсію можна наближено обчислити за такими виразами:

$$\bar{x}(k)=\frac{1}{k}\sum_{i=1}^{k}x(k),\ var_k(x)=\frac{1}{k-1}\sum_{i=1}^{k}\left[x(i)-\bar{x}(k)\right]^2,\ k=2,...,N.$$

У результаті застосування цих виразів отримаємо ряд значень умовного середнього та умовної дисперсії, тобто ще дві характеристики процесу, якими можна скористатися при побудові математичних і статистичних моделей.

Наближені значення умовних вибіркових статистичних характеристик можна визначити також за аналогією з обчисленням ковзного середнього. При такому підході потрібно вибрати ширину ковзного вікна й обчислити значення статистичного параметра, рухаючись крок за кроком від початку до кінця часового ряду. Наприклад, якщо вибрати вікно завширшки п'ять значень ряду, то умовна дисперсія обчислюватиметься за виразом

$$var_k(x)=\frac{1}{4}\sum_{i=1}^{k+2}\left[y(i)-\bar{x}\right]^2,\ k=3,...,N-2,$$

або в загальному вигляді –

$$var_k(x)=\frac{1}{d-1}\sum_{i=k-(d-1)/2}^{k+(d-1)/2}\left[y(i)-\bar{y}\right]^2,\ k=\frac{d-1}{2},...,N-\frac{d-1}{2},$$

де d – ширина ковзного вікна. Ширина вікна залежить від того, наскільки швидко змінюється дисперсія. Якщо вона має високу динаміку, то вибрана ширина вікна має дорівнювати 5 … 9. Незважаючи на наближеність таких розрахунків, практика моделювання свідчить про те, що обчислені значення виявляються досить корисними при побудові моделей.

У складніших випадках обчислюють функцію умовної дисперсії, яка точніше відображує характер її зміни в часі. Наприклад, можна розглянути таку модель процесу:

$$y(k)=\sqrt{f\left[x(k)\right]}\,\varepsilon(k), \tag{4.1}$$

$$\varepsilon(k)=\beta\varepsilon(k-1)+\nu(k), \tag{4.2}$$

де $y(k)\in\Re,\ k=1,2,...,N;$ $\{\nu(k)\}$ – множина незалежних однаково розподілених (НОР) величин, що мають нормальний розподіл з параметрами $N(0,1);\ \beta\in\Theta=(-1;1);\ f\left[x(k)\right]\in C^p[0,1]$. Припущення про те, що $\nu(k)$ – гауссів процес, зроблено для зручності викладок. Змінна $x(k)\in[0,1]$ є упорядкованою за значеннями, тобто

$$x(1)\le x(2)\le ...\le x(N),$$

де $x(k)=\frac{k}{N}, k=1,2,...,N$. Функцію $f[x(k)]$ називають функцією дисперсії, хоча вона не повністю описує дисперсійно-коваріаційну структуру процесу (4.1) – (4.2). Припустимо, що $f[x(k)]$ має p неперервних похідних. Безумовну дисперсію процесу $y(k)$ можна визначити таким чином:

$$var[y(k)]=f[x(k)]E\left[\varepsilon^2(k)\right]=f[x(k)]E\left\{[\beta\varepsilon(k-1)+\nu(k)]^2\right\}=$$
$$=\left\{\beta^2 E\left[\varepsilon^2(k-1)\right]+E\left[\nu^2(k)\right]\right\}f[x(k)]=\beta^2 f[x(k)]E\left[\varepsilon^2(k-1)\right]+f[x(k)]=$$
$$=\beta^2 var[y(k)]+f[x(k)].$$

Звідси отримаємо

$$var[y(k)]=\frac{f[x(k)]}{1-\beta^2}.$$

Таким чином, процес (4.1) – (4.2) є умовно і безумовно гетероскедастичним. Задача полягає у тому, щоб оцінити функцію $f[x(k)]$. Вона розглядатиметься нижче.

4.4 Оцінювання якості прогнозу

Важливим моментом процесу прогнозування є об'єктивне визначення якості отриманого прогнозу. Оскільки прогнозовані значення – випадкові величини, то для оцінювання їхньої якості необхідно використовувати декілька статистичних критеріїв. Рисунок 4.1 ілюструє часову вісь і відрізки часу, на яких виконуються оцінювання моделі та перевірка якості прогнозу.

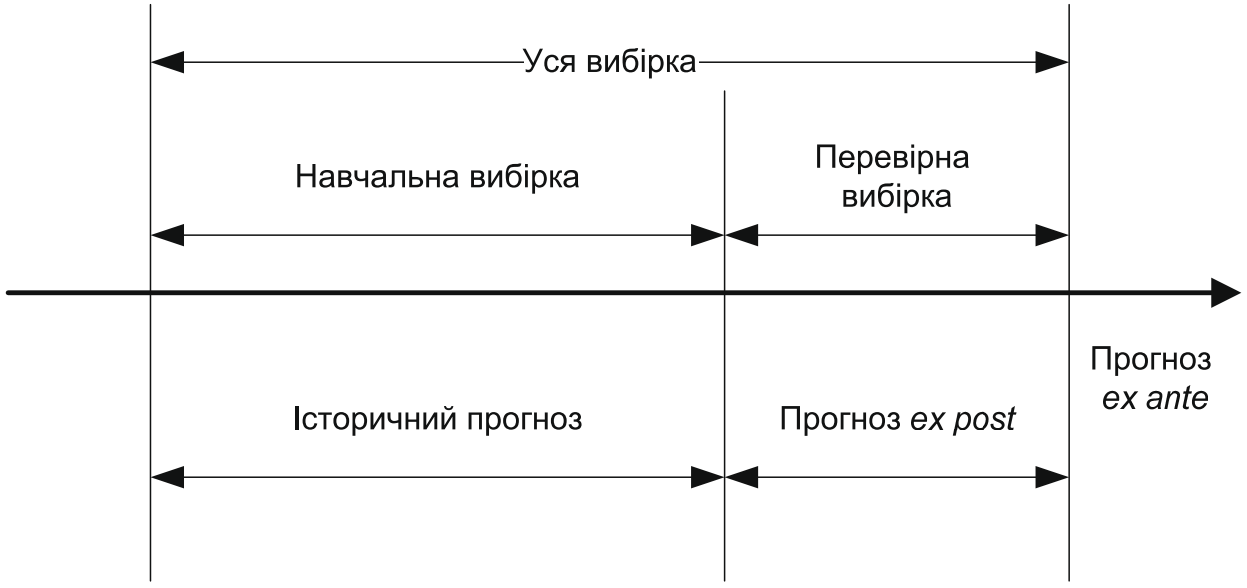


Рисунок 4.1 – Види прогнозування за часовим рядом

Наявну вибірку даних доцільно розділити на навчальну та перевірну. На начальній вибірці виконується оцінювання параметрів моделі процесу та реалізується так званий «історичний» прогноз, який дає змогу

встановити якість однокрокового прогнозу на цій ділянці ряду. Прогноз на перевірній частині вибірки даних у науковій літературі називають ще прогнозом *ex post*. У різних емпіричних дослідженнях рекомендують залишати для перевірки 5…40% значень ряду даних, хоча при аналізі коротких рядів доцільно значно більшу частину ряду використовувати для оцінювання параметрів моделі.

Прогнозування значень поза вибіркою даних називають прогнозом *ex ante* (див. рисунок 4.1).

Як правило, для оцінювання якості прогнозів використовують множину доповнювальних статистичних критеріїв. Наприклад, значення середньоквадратичної похибки залежить від масштабу даних, а тому недостатньо використовувати тільки цей статистичний параметр для аналізу якості прогнозу. Розглянемо деякі статистичні критерії якості прогнозу та їх призначення.

Дисперсія і стандартний відхил прогнозу

Досить часто проста сума похибок прогнозів дорівнює нулю, оскільки похибки мають різні знаки, а тому необхідно використовувати іншу міру похибки. Ступінь розсіювання значень змінної навколо її середнього можна виміряти за допомогою стандартного відхилу $\sigma(k)$, який дорівнює квадратному кореню з дисперсії, тобто

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left[y(k)-\hat{y}(k)\right]^2}, \qquad (4.3)$$

де $y(k)$ – фактичне значення змінної; $\hat{y}(k)$ – прогноз (ділення на $N-1$, а не на N слушне при невеликій потужності ряду).

Стандартний відхил залишків – один з основних показників якості прогнозу. Це досить широко застосовувана статистична характеристика яка є корисною при аналізі поведінки процесів різної природи. Вона використовується, наприклад, як міра ризику (волатильність) при аналізі фінансових процесів, як характеристика надійності в управлінні запасами та в інших прикладних задачах управління, а також як міра розсіювання значень змінних стану при аналізі систем різної природи.

Оскільки дисперсія є квадратичною характеристикою процесу, то в під час її обчислення не відбувається взаємної компенсації значень відхилів від середнього. Крім того, дисперсію $\sigma^2(k)$ і стандартний відхил можна перевіряти на статистичну значущість, що певною мірою сприяє поглибленому аналізу похибки прогнозу. Вище було показано, що при застосуванні для прогнозування рівнянь АРКС прогноз незміщений (математичне сподівання похибок прогнозів дорівнює нулю), але дисперсія прогнозованих значень прямо пропорційна числу кроків прогнозування. Однак вона збігається до скінченної величини для стаціонарних процесів.

При відносно невеликому горизонті прогнозування можна стверджувати, що майбутнє значення прогнозованого показника потрапляє

в інтервал, який визначається як ± два стандартних відхили від обчисленого значення прогнозу.

Очевидно, що дисперсії і стандартного відхилу недостатньо для аналізу якості прогнозів. Наприклад, якщо прогноз попиту дорівнює 2500 одиниць товару зі стандартним відхилом похибки 100, то інтервал 2300 – 2700 є досить інформативним. Однак, якщо при такому ж значенні прогнозу стандартний відхил становитиме 400 одиниць, то відповідний інтервал 1700 – 3300 навряд чи буде корисним для практичного використання.

Середнє абсолютне значення похибки

Обчислення середнього абсолютного значення похибки (САП) ґрунтується на виразі для експоненціального середнього і має вигляд

$$САП\ (k) = a\,|\,y(k) - \hat{y}(k)\,| + (1 - a)САП\ (k - 1) = a\,|\,e(k)\,| + (1 - a)САП\ (k - 1), \tag{4.4}$$

де $0 < a < 1$; $e(k)$ – похибка прогнозу. При такому обчисленні середнє абсолютне значення завжди невід'ємне.

САП можна обчислити також за іншим виразом

$$САП = \frac{1}{N}\sum_{k=1}^{N} \left|y(k) - \hat{y}(k,k)\right|.$$

Для досить широкого класу статистичних розподілів випадкових величин значення стандартного відхилу є дещо більшим за САП і строго пропорційним йому. Коефіцієнт пропорційності коливається для різних розподілів між 1,2 і 1,3 (для нормального розподілу це значення дорівнює $\sqrt{\pi / 2} = 1{,}2533$) [7, 8]. Таким чином, можна записати, що

$$\sigma(k) \approx 1{,}25САП. \tag{4.5}$$

Виходячи з цього процедуру оцінювання *стандартної похибки* прогнозу можна сформулювати так:

- обчислити поточне значення похибки прогнозу як $e(k) = y(k) - \hat{y}(k)$;
- обчислити нове значення САП за допомогою рівняння (4.4);
- знайти поточний стандартний відхил за виразом (4.5).

Середню абсолютну похибку та її стандартний відхил використовують, як правило, для визначення якості оцінки прогнозу одночасно.

Середній квадрат похибки і сума квадратів похибок

Якщо середній квадрат похибок (СКП) визначається для поширеного випадку однокрокового прогнозування при довжині часового ряду N, то СКП визначається за формулою

$$САП = \frac{1}{N}\sum_{k=1}^{N} \left[y(k) - \hat{y}(k)\right]^2. \tag{4.6}$$

У випадках, коли прогнозування виконується на s кроків відносно

моменту часу k , то СКП обчислюється за виразом

$$САП = \frac{1}{s}\sum_{k=1}^{s}\left[y(k+i)-\hat{y}(k+i,k)\right]^2. \quad (4.7)$$

Очевидно, що формули (4.6) і (4.7) однакові, але кожна форма чітко відображує тип прогнозу.

Більшість відомих пакетів для математичного моделювання і виконання статистичних розрахунків обчислюють суму квадратів похибок за виразом

$$СмКП = \sum_{k=1}^{N}\left[y(k)-\hat{y}(k)\right]^2. \quad (4.8)$$

Це корисний інтегральний показник якості, який використовують для порівняльного аналізу різних методів прогнозування.

Саме СКП і СмКП є найбільш поширеними критеріями, що використовуються при порівняльному аналізі та виборі кращої моделі для обчислення оцінок прогнозів. Однак їх також обмаль для поглибленого аналізу результатів.

Середня абсолютна похибка у процентах

Середня абсолютна похибка в процентах (САПВ) – це середнє абсолютних значень похибок оцінок прогнозу в процентах відносно фактичного значення показника:

$$САПВ = \frac{1}{N}\sum_{k=1}^{N}\frac{\left|y(k)-\hat{y}(k)\right|}{\left|y(k)\right|}\cdot 100\% = \frac{1}{N}\sum_{k=1}^{N}\frac{\left|e(k)\right|}{\left|y(k)\right|}\cdot 100\%, \quad (4.9)$$

або у разі прогнозування на s кроків відносно k -го моменту:

$$САПВ = \frac{1}{s}\sum_{k=1}^{N}\frac{\left|y(k+i)-\hat{y}(k+i,k)\right|}{\left|y(k+i)\right|}\cdot 100\% = \frac{1}{N}\sum_{k=1}^{N}\frac{\left|e(k+i)\right|}{\left|y(k+i)\right|}\cdot 100\%. \quad (4.10)$$

Оскільки ця міра характеризує відносну якість прогнозу, то її використовують в основному для порівняння точності прогнозів різнорідних об’єктів (процесів) прогнозування. Однак вона є завжди корисною при використанні порівняльного аналізу якості прогнозування одного й того ж процесу різними методами, оскільки відносна міра є чіткою і зрозумілою для дослідника і практичного користувача. Типові значення САПВ і їх пропоновану інтерпретацію дано в таблиці 4.1 [7, 8].

Таблиця 4.1 – Інтерпретація типових значень критерію САПВ

САПВ, %	Інтерпретація
< 10	Висока точність
10 – 20	Добра точність
20 – 50	Задовільна точність
> 50	Незадовільна (неприйнятна) точність

Якщо у формулах (4.9) і (4.10) $y(k)$ або $y(k+i)$ наближаються до нуля, то значення САПВ прямуватиме до нескінченності. Про це слід пам'ятати при застосуванні даного критерію якості прогнозу. Для того щоб виконати обчислення цього критерію у таких випадках, нульові значення $y(k)$ або $y(k+i)$ необхідно пропускати з відповідним коригуванням значення N або s. Можливо, що такий підхід не відповідає деяким вимогам статистичного аналізу даних, але він дає можливість наближено і більш повно виконати аналіз якості прогнозування.

Середня похибка (СП) і середня похибка в процентах (СПВ)

Середня похибка – це не відносний показник, вона характеризує ступінь зміщення прогнозованих значень від фактичних і розраховується за формулою

$$СП = \frac{1}{N}\sum_{k=1}^{N}\left[y(k) - \hat{y}(k)\right] = \frac{1}{N}\sum_{k=1}^{N} e(k), \qquad (4.11)$$

або

$$СП = \frac{1}{s}\sum_{i=1}^{s}\left[y(k+s) - \hat{y}(k+s,k)\right]. \qquad (4.12)$$

Очевидно, що СП зменшуватиметься тоді, коли похибки мають різні знаки. Середню похибку в процентах (СПВ) обчислюють за виразом

$$СПВ = \frac{1}{N}\sum_{k=1}^{N}\frac{\left[y(k) - \hat{y}(k)\right]}{y(k)} \cdot 100\%, \qquad (4.13)$$

або

$$СПВ = \frac{1}{s}\sum_{k=1}^{s}\frac{\left[y(k+s) - \hat{y}(k+s,k)\right]}{y(k+s)} \cdot 100\%. \qquad (4.14)$$

СПВ також характеризує *зміщеність* прогнозу. Якщо втрати при прогнозуванні, пов'язані із завищенням фактичного майбутнього значення, врівноважуються заниженням, то ідеальний прогноз має бути незмінним. У такому разі СП і СПВ повинні наближатися до нуля. Очевидно, що нуль – це ідеальне значення і забезпечити його на практиці неможливо. Емпірично встановлено, що прийнятними значеннями для СПВ (так само, як і для САПВ) є ≤ 5.

Очевидно, що максимальна абсолютна похибка (МАП) може бути визначена як

$$МАП = \max_{k}\left\{\left|y(k) - \hat{y}(k)\right|\right\},\ 1 \leq k \leq N, \qquad (4.15)$$

або

$$МАП = \max_{i}\left\{\left|y(k+1) - \hat{y}(k+1,k)\right|,...,\left|y(k+s) - \hat{y}(k+s,k)\right|\right\},\ 1 \leq i \leq s, \qquad (4.16)$$

а мінімальна абсолютна похибка (МіАП) визначається як

$$МАП = \max_{k}\left\{ \left| y(k) - \hat{y}(k) \right| \right\},\ 1 \le k \le N, \tag{4.17}$$

або

$$МАП = \max_{i}\left\{ \left| y(k+1) - \hat{y}(k+1,k) \right|, ..., \left| y(k+s) - \hat{y}(k+s,k) \right| \right\},\ 1 \le i \le s. \tag{4.18}$$

Критерії МАП і МіАП також можуть бути корисними при виконанні порівняльного аналізу кількох методів прогнозування, особливо якщо нас цікавлять максимально або мінімально можливі відхилення прогнозів від фактичних значень на заданому інтервалі.

Коефіцієнт нерівності Тейла [27].

Коефіцієнт нерівності Тейла U – це важливий індикатор якості моделі та прогнозу. За визначенням $0 \le U \le 1$. Якщо $U = 1$, то модель має практично нульові (неприйнятні) прогнозуючі властивості, що випливає з формули для обчислення U :

$$U = \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left[y(k) - \hat{y}(k) \right]^2}}{\sqrt{\frac{1}{N}\sum_{k=1}^{N} y^2(k)} + \sqrt{\frac{1}{N}\sum_{k=1}^{N} \hat{y}^2(k)}}. \tag{4.19}$$

При $U = 1$ прогнозовані значення збігаються з фактичними значеннями ряду, тобто модель ідеальна. Отже, U дає можливість установити придатність моделі (методу) для оцінювання прогнозу.

Коефіцієнт Тейла можна розкласти на три складові:

1) складова, пропорційна зміщенню,

$$U^M = \frac{(\bar{y} - \bar{\hat{y}})^2}{\frac{1}{N}\sum_{i=1}^{N}\left[y(i) - \hat{y}(i) \right]^2}, \tag{4.20}$$

2) складова, пропорційна дисперсії,

$$U^S = \frac{(\sigma_{факт} - \sigma_{моделі})^2}{\frac{1}{N}\sum_{i=1}^{N}\left[y(i) - \hat{y}(i) \right]^2}, \tag{4.21}$$

3) складова, пропорційна коваріації

$$U^C = \frac{2(1-\rho)(\sigma_{факт} \cdot \sigma_{моделі})}{\frac{1}{N}\sum_{i=1}^{N}\left[y(i) - \hat{y}(i) \right]^2}, \tag{4.22}$$

де ρ – коефіцієнт кореляції між залишками моделі.

Величина U^M використовується для перевірки факту, чи є

систематичним відхилення середніх фактичного і прогнозованого ряду. Іншими словами, чи існує систематичне зміщення на виході моделі в той чи інший бік. Чим меншим є обчислене значення U^M, тим краща модель. Якщо $U^M = 0$, то прогнозовані значення не містять зміщення, а також модель адекватна процесу за цим показником.

Величина U^S використовується для тестування динамічних властивостей моделі: тобто, чи відповідає її дисперсія дисперсії фактичного ряду. Наприклад модель може відтворювати систематично менші коливання, ніж ті, які має фактичний ряд. Чим меншим є значення U^S, тим менше відхилення дисперсії виходу моделі від дисперсії ряду.

Нарешті, складова, пропорційна коваріації, є мірою корельованості фактичного та прогнозованого за моделлю ряду. Зазначимо, що за побудовою останнього критерію для нього виконується рівність

$$U^M + U^S + U^C = 1.$$

Якість моделі також визначається тим, наскільки точно вона може прогнозувати зміни напряму розвитку процесу, тобто нахил або знак тренду Моделі можуть мати високу точність відтворення ряду, але погано прогнозувати тренди або цикли. Інші моделі, навпаки, можуть мати меншу точність (адекватність), але кращі динамічні властивості. Таким чином, завжди необхідно шукати компроміс між точністю моделі, тобто якістю апроксимації ряду та її динамічними властивостями. Однак для аналізу цієї властивості формальних тестів немає. Можна наближено встановити якість моделі щодо відтворення динаміки ряду шляхом візуального аналізу фактичного і спрогнозованого ряду.

Ще однією характеристикою якості моделі є її чутливість до початкового (стартового) періоду імітаційного моделювання. Якщо модель генерує результати, які можна наближено класифікувати як інваріантні до початкових умов, то вона вважається якісною. Інакше, якщо результати імітаційного моделювання залежать від початкових умов, можна припустити, що модель неякісна. Наприклад, вона може мати нестаціонарність певного типу.

4.5 Довірчий інтервал для прогнозу

При використанні для обчислення прогнозу регресійних рівнянь і рівнів інших типів ми отримуємо точкову оцінку. Однак така оцінка є далеко не завжди значущою. При прогнозуванні необхідно визначити інтервал, усередині якого з достатнім ступенем упевненості можна очікувати появу фактичного значення показника. У регресійному аналізі межі цього інтервалу задаються за допомогою довірчого інтервалу.

Довірчий інтервал – це інтервал, який містить сам прогноз і в якому з визначеним ступенем упевненості можна очікувати появу фактичного

значення прогнозованої змінної.

Так, значення прогнозу $\hat{y}(k+1)=1500$ з довірчим інтервалом ±150 і ступенем упевненості 95 % означає, що з імовірністю $\approx$0,95 очікується те, що наступне значення прогнозованої змінної лежатиме в межах 1350 – 1650. Якби довірчий інтервал становив ±500, то з імовірністю $\approx$0,95 очікується, що майбутнє значення лежатиме в інтервалі 1000 – 2000.

У регресійному аналізі *мінімальна ширина довірчого інтервалу* відповідає точці $(\bar{y},\bar{k})$ – середині спостереження. По обидва боки від цієї середини довжина інтервалу збільшується. Для того щоб визначити довірчий інтервал, необхідно знайти стандартну похибку рівняння регресії S_r яка обчислюється усіма пакетами статистичного оброблення даних. Величину S_r ще називають середньоквадратичним відхилом і обчислюють за формулою

$$S_r=\sqrt{\frac{1}{N-p}\sum_{i=1}^{N}\left[y(k)-\hat{y}(k)\right]^2}\ , \tag{4.23}$$

де p – кількість параметрів моделі, які оцінюються в процесі її побудови. Суму $\sum\left[y(k)-\hat{y}(k)\right]^2$ можна обчислити також як різницю між загальною сумою квадратів і сумою квадратів значень, знайденою за регресією.

Тепер можна визначити стандартну похибку прогнозу як

$$S_{\hat{y}(k)}=S_r=\sqrt{1+\frac{1}{N}+\frac{(k-\bar{k})^2}{\sum_{i=1}^{N}(k-\bar{k})^2}}, \tag{4.24}$$

тобто стандартна похибка прогнозу залежить від довжини ряду N і віддалі (у часі) від середини часового періоду, що розглядається, до моменту прогнозування. Видно, що збільшення довжини ряду приводить до зменшення стандартної похибки прогнозу. Якщо число кроків прогнозування відносно N -го значення ряду дорівнює τ , то член $(k-\bar{k})$ у виразі (4.24) необхідно замінити на $(\tau+(N-1)/2)$. Таку процедуру введення нового позначення називають ще *стандартним прогностичним перепозначенням*.

З урахуванням введеного позначення вираз для стандартної похибки набуває вигляду

$$S_{\hat{y}(k+\tau)}=S_r\sqrt{1+\frac{1}{N}+\frac{(\tau+(N-1)/2)^2}{\sum_{k=1}^{N}k^2-(\sum k)^2/N}}, \tag{4.25}$$

де S_r - рівняння регресії $y(k)=a_0+bk$, яке можна обчислити за виразом

$$S_r=\sqrt{\frac{\sum y^2(k)-a_0\sum y(k)-b\sum k\,y(k)}{N-2}}. \tag{4.26}$$

Скориставшись стандартною похибкою прогнозу, можна визначити

довірчі інтервали. Так, наближеними 99-, 95- і 68-процентним довірчими інтервалами будуть значення $\pm 3S_{\hat{y}(k+\tau)}$, $\pm 2S_{\hat{y}(k+\tau)}$ і $\pm S_{\hat{y}(k+\tau)}$ відповідно.

4.6. Запитання і вправи

1) Наведіть приклади практичних задач, при розв'язанні яких необхідно користуватися прогнозами. Яка різниця між прогнозом і передбаченням?

2) Наведіть приклади задач прогнозування при прийнятті особистих рішень.

3) На які частини і в якій пропорції доцільно ділити вибірку даних при побудові моделі та оцінюванні прогнозів?

4) Що означає термін «історичний прогноз»? Чи є необхідність у його виконанні? Обґрунтуйте відповідь. Які існують інші види прогнозів?

5 МЕТОД ГРУПОВОГО ВРАХУВАННЯ АРГУМЕНТІВ

5.1 Особливості методу

Відмінність алгоритмів методу групового врахування аргументів (МГВА) [29] від інших алгоритмів структурної ідентифікації і селекції кращої регресії полягає у такому:

- використання *зовнішнього критерію*, що ґрунтується на поділі вибірки даних на навчальну та перевірну;
- зменшення вимог до обсягу первісної інформації;
- більша *різноманітність структур:* використання, як у регресійних алгоритмах, шляхів повного чи зменшеного перебору варіантів структур і застосування оригінальних багаторядних ітераційних процедур;
- вищий *ступінь автоматизації* – достатньо лише ввести первісні дані та вказати зовнішній критерій;
- автоматична *адаптація* структури оптимальної моделі та зовнішніх критеріїв до рівня завад у системі чи порушень – ефект завадостійкості зумовлює робасність підходу;
- запровадження принципу *незавершених рішень* у процес поступового ускладнення моделей.

5.2 Основні принципи і загальна схема методу

Метод запропоновано наприкінці 60-х років XX століття академіком О. Г. Івахненком (Інститут кібернетики НАН України). Він ґрунтується на ідеях самоорганізації і механізмах живої природи – схрещуванні (гібридизації) і селекції (доборі).

Нехай є вибірка з N спостережень вхідних $X(i)$ і вихідних $Y(i)$ векторів:

$$\{X(1)\ \ldots Y(1)\},$$
$$\{X(2)\ldots X(2)\},$$
$$\ldots$$
$$\{X(N)\ldots Y(N)\}.$$

За результатами спостережень потрібно визначити $F(x)$ (рисунок 5.1). При цьому структура моделі $F(x)$ невідома. Найбільш повну залежність між входами $X(i)$ і виходами $Y(i)$ можна подати за допомогою узагальненого полінома Колмогорова – Габора [30]. Якщо $X = \{x_1,\ldots, x_N\}$, то поліном має вигляд

$$Y = a_0 + \sum_{i=1}^{N} a_i x_i + \sum_{j=1}^{N} \sum_{i\le j} a_{ij} x_i x_j + \sum_{i=1}^{N} \sum_{j\le i} \sum_{k\le j} a_{ijk} x_i x_j x_k + \ldots,$$

де всі коефіцієнти a невідомі.

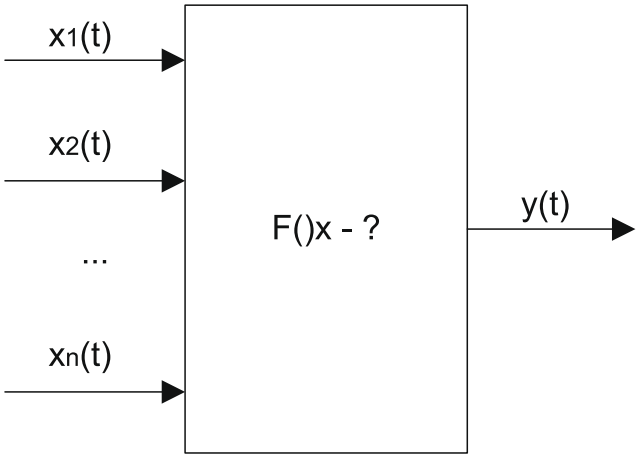


Рисунок 5.1 – Схематичне зображення об'єкта з багатьма входами

При побудові моделі (при оцінюванні її коефіцієнтів) як критерій використовують критерій регулярності (точності)

$$\bar{\varepsilon}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2.$$

Необхідно мінімізувати середньоквадратичну похибку $\bar{\varepsilon}^2 \to min$.

Принцип множинності моделей: існує множина моделей на даній вибірці, що забезпечують нульову похибку (достатньо підвищити степінь полінома моделі). Тобто якщо є N вузлів інтерполяції, то можна побудувати сімейство моделей, кожна з яких при проходженні через експериментальні точки даватиме нульову похибку:

$$\bar{\varepsilon}^2 = 0.$$

Як правило, степінь нелінійності беруть не вище за $n-1$, де n – число точок вибірки. Позначимо через S складність моделі (визначається числом членів полінома Колмогорова – Габора [30]).

Значення похибки $\bar{\varepsilon}^2$ залежить від складності структури моделі. При цьому в міру зростання складності спочатку вона падатиме, а потім зростатиме. Нам же потрібно вибрати таку оптимальну структуру, при якій похибка буде мінімальною. Крім того, якщо враховувати дію завад, то можна виділити такі моменти:

1. При різному рівні завад залежність $\bar{\varepsilon}^2$ від складності S змінюватиметься, зберігаючи при цьому загальну спрямованість, тобто із зростанням складності вона спочатку буде зменшуватись, а потім зростатиме.

2. При збільшенні рівня завад величина $\min\limits_{S} \bar{\varepsilon}^2$ зростатиме.

3. Із зростанням рівня завад $S_0 = arg\ min\, \bar{\varepsilon}^2$ буде зменшуватись (оптимальне значення складності зміщуватиметься вліво). При цьому

$\bar{\varepsilon}^2(S_0) > 0$, якщо рівень завад ненульовий (рисунок 5.2).

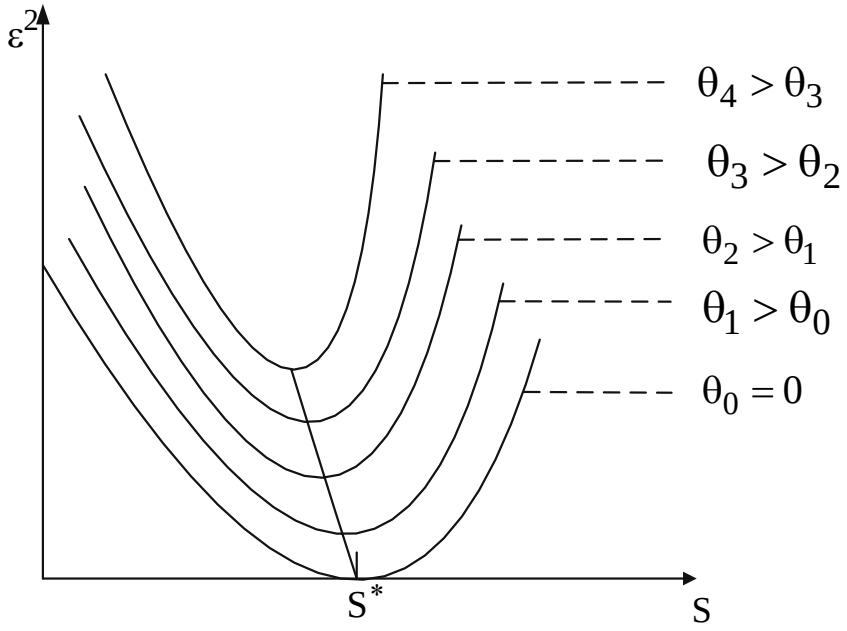


Рисунок 5.2 – Варіанти зміщення екстремуму

Теорема неповноти Ґеделя [31, 32]. У будь-якій формальній логічній системі міститься ряд тверджень і теорем, які не можна ні спростувати, ні довести, залишаючись у рамках цієї системи аксіом.

У цьому разі ця теорема означає, що вибірка завжди неповна. Один із способів подолання цієї неповноти полягає у застосуванні принципу зовнішнього доповнення. Як зовнішнє доповнення використовується додаткова вибірка (перевірна), значення якої не використовувались при навчанні системи (тобто при пошуку оцінок значень коефіцієнтів полінома Колмогорова – Габора).

Пошук найкращої моделі здійснюється у такий спосіб:

- уся вибірка поділяється на навчальну і перевірну:

$$N_{виб} = N_{навч} + N_{перев};$$

- на навчальній вибірці $N_{навч}$ визначаються значення a_0, a_i, a_{ij};
- на перевірній вибірці $N_{навч}$ відбираються кращі моделі.

Вхідний вектор має розмірність $N\left(X = \{x_1, ..., x_N\}\right)$.

Принцип свободи вибору (неостаточності проміжного розв'язку):

1. Для кожної пари x_i і x_j будуються часткові описи (усього C_N^2) вигляду

$$y^{(s)} = \varphi(x_i, x_j) = a_0 + a_i x_i + a_j x_j, s = 1..C_N^2 (лінійні),$$

або

$$y^{(s)} = \varphi(x_i, x_j) = a_0 + a_i x_i + a_j x_j + a_{ii} x_i^2 + a_{ij} x_i x_j + a_{jj} x_j^2,$$

$s = 1..C_N^2$ (квадратичні).

2. Визначаємо коефіцієнти цих моделей за МНК, використовуючи навчальну вибірку. Тобто знаходимо $\hat{a}_0, \hat{a}_1, ..., \hat{a}_j, ..., \hat{a}_N, \hat{a}_{11}, ..., \hat{a}_{ij}, ... \hat{a}_{NN}$.

3. Далі на перевірній вибірці для кожної з цих моделей шукаємо оцінку

$$\bar{\varepsilon}_s^2 = \frac{1}{N_{перев}} \sum_{k=1}^{N_{перев}} \left[Y(K) - \hat{Y}_k^{(s)} \right]^2,$$

де $Y(K)$ – дійсне вихідне значення в k-й точці перевірної вибірки; $\hat{Y}_k^{(s)}$ – вихідне значення у k - й точці перевірної вибірки для моделі складності S, і визначаємо, таким чином, F кращих моделей.

Обрані y_i подаються на другий ряд, де шукаємо залежність:

$$z_I = \varphi^{(2)}(x_i, x_j) = a_0^{(2)} + a_1^{(2)} x_i + a_2^{(2)} x_j + a_3^{(2)} x_i^2 + a_4^{(2)} x_i x_j + a_5^{(2)} x_j^2.$$

Оцінка тут така ж, як і на першому ряді. Добір кращих моделей здійснюється знову так само, але $F_2 < F_1$.

Процес конструювання рядів повторюється доти, доки середній квадрат похибки зменшується. Коли на шарі m одержимо збільшення похибки $\bar{\varepsilon}^2$, то процес пошуку припиняється (рисунок 5.3).

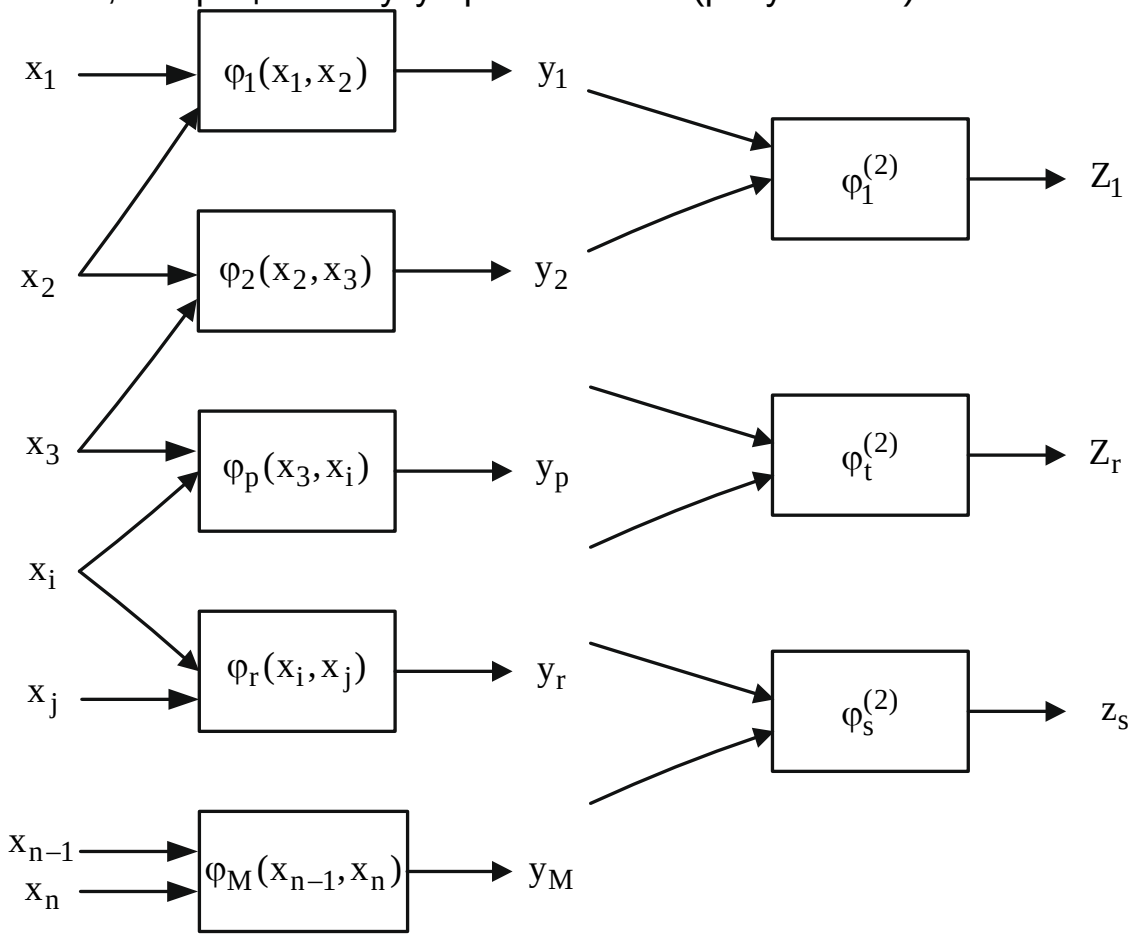


Рисунок 5.3 – Схема формування часткових описів

Якщо часткові описи квадратичні і число рядів дорівнює L, то одержуємо, що степінь полінома $k = 2L$. На відміну від звичайних методів статистичного аналізу при такому підході можна одержати досить складну залежність, навіть маючи коротку вибірку.

Існує проблема: на першому ряді можуть відсіятися деякі змінні x_i і x_j, котрі впливають на вихідні дані. У зв'язку з цим запропоновано таку модифікацію – на другому шарі подавати y_i і y_j, тобто

$$z_I = a_0^{(2)} + a_1^{(2)} y_i + a_2^{(2)} x_j + a_3^{(2)} y_i^2 + a_4^{(2)} y_i x_j + a_5^{(2)} x_j^2.$$

Це важливо при більшому рівні збурень для того, щоб забезпечити незміщеність оцінок моделей.

Виникають два критерії добору кращих кандидатів часткових описів, які передаються на певному шарі на наступний ряд (шар).

1. Критерії регулярності (точності) $\bar{\varepsilon}^2_{перев}$ такі:

а) $$\bar{\varepsilon}^2 = \frac{1}{N_{перев}} \sum_{i=1}^{N_{перев}} (y_i - y^{*(i)})^2;$$

б) $$\bar{\Delta}^2_{перев} = \frac{\sum_{i=1}^{N_{перев}} (y_i - y^{*(i)}(x))^2}{\sum_{i=1}^{N_{перев}} (y_i - \bar{y})^2},$$

де $N_{перев}$ – довжина (потужність) перевірної вибірки; y_i – фактичне значення змінної перевірної вибірки; $y^{*(i)}$ – оцінка значення змінної, отримана за моделлю; $\bar{y}$ – середнє значення змінної.

2. Критерій незміщеності.

Беремо всю вибірку, поділяємо її на дві частини $R = R_1 + R_2$.

Перший експеримент: R_1 – навчальна вибірка, R_2 – перевірна; визначаємо виходи моделі $y^{*(i)}$, $i = 1..R$.

Другий експеримент: R_2 – навчальна вибірка, R_1 – перевірна; визначаємо виходи моделі $y^{**(i)}$, $i = 1..R$ порівнюємо.

Критерій незміщеності

$$n_{зм} = \frac{1}{N} \sum_{i=1}^{N} (y_i^* - y_i^{**})^2.$$

Чим менше $n_{зм}$, тим меншим є зміщення моделі. Такий критерій визначається для кожного часткового опису першого рівня і потім

знаходиться $n_{зм}$ для рівня в цілому:

$$n_{зм}\frac{1}{F}\sum_{i=1}^{F} n_{зм,i}^{(1)},$$

для F кращих моделей. У ряді варіантів $F=1$. Так само на другому шарі обчислюємо $n_{зм}^{(2)}$.

Процес селекції триває доти, доки цей критерій перестане зменшуватися, тобто до досягнення умови

$$n_{зм}^{(2)} \to min.$$

Переваги методу групового врахування аргументів

1. Можна встановити невідому довільно складну залежність за обмеженою вибіркою. Кількість невідомих параметрів моделі може бути більшою за кількість точок навчальної послідовності.

2. Можливість адаптації параметрів моделі при одержанні нових даних експериментів. (Зокрема, використовуючи рекурсивний метод найменших квадратів (РМНК)).

5.3 Алгоритм самоорганізації МГВА і його застосування у задачах прогнозування

Багаторядний МГВА. Існують два підходи при виборі часткових описів і побудові МГВА:

1) точнісний.
2) робасний.

При першому підході в алгоритмі МГВА при виборі описів використовують *критерій регулярності послідовності*, або точнісний критерій, що визначається у такий спосіб:

$$AB = \frac{1}{n_B}\sum_{i=1}^{N_B}(\hat{y}_i^B - y_i^B)^2 \to min,$$

де y_i^B – фактичний вихід на вибірці B; $\hat{y}_i^B$ – прогноз згідно з моделлю.

Навчання відбувається на вибірці A, перевірка – на B.

Цей критерій часто застосовують на практиці. Перший підхід використовують для одержання найбільш точної моделі за вибіркою даних.

В основі *робасного* підходу лежить використання диференціального критерію несуперечності

$$CN = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i^A - \hat{y}_i^B)^2 \to min,$$

де $\hat{y}_i^A$, $\hat{y}_i^B$ – це виходи моделі, побудовані за вибірками A і B відповідно. Цей критерій – критерій узгодженості моделей, який спрацьовує при зашумлених даних.

Для того щоб одержати найбільш гострий глибокий мінімум за цим критерієм, вибірки A і B вибираються так, щоб виконувалась умова

$$|A| \approx |B|,$$

їхні дисперсії були приблизно однакові, а взаємна дисперсія була максимальна. Наступне питання стосується вибору предикатів (змінних), що вводяться в модель. Насамперед для кожного $\tilde{x}_i = \{x_i^1, x_i^2, ..., x_i^N\}$ – вектора-стовпця (певне спостереження) – виконують процедуру нормування:

1) $\tilde{x}_i = \dfrac{x_i}{x_{i,max} - x_{i,min}}$;

2) $x_i' = \dfrac{x_i - x_{i,min}}{x_{i,max} - x_{i,min}} (\in [0;1])$;

3) $x_i' = \dfrac{x_i - \bar{x}_i}{x_{i,max} - x_{i,min}} (\in [-1;1]$.

Далі визначають коефіцієнти кореляції вхідних спостережень з виходом

$$\rho_{yx_i} = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(x_{iy} - \bar{x}_i)}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2 \sum_{i=1}^{N} (x_{ij} - \bar{x}_i)^2}}$$

і перевіряють гіпотезу про те, що коефіцієнт кореляції відмінний від нуля. Для моделі вибираємо ті змінні, для яких $\rho_{yx_i} \geq \Theta$, де Θ – певний поріг. Якщо є необхідність досліджувати нестаціонарні процеси, то можна використовувати підхід з виділенням трендів.

5.4 Описання алгоритму

Перша ітерація

Крок 1. З множини виходів $X = \{X_1, X_2, ..., X_n\}$ вибирають пару аргументів X_i, X_j і складають часткові описи вигляду

$$Y_k^{(1)} = \varphi(X_i, X_j), i \neq j, i, j = \overline{1}, \overline{N}.$$

При цьому використовують часткові описи квадратичного типу

$$Y_k^{(1)} = a_0 + a_i X_i + a_j X_j + a_{ij} X_i X_j + a_{ii} X_i^2 + a_{jj} X_j^2.$$

Число часткових описів 1-го ряду дорівнює $M = n(n-1)/2$.

Крок 2. Застосовуючи метод найменших квадратів для кожного опису, за навчальною вибіркою знаходять оцінки невідомих коефіцієнтів $\hat{a}_0, \hat{a}_i, \hat{a}_j, \hat{a}_{ij}, \hat{a}_{ij}, \hat{a}_{ij}$.

Крок 3. За критерієм мінімуму $\bar{\varepsilon}^2$ на перевірній послідовності відбирають F_1 кращих моделей, тобто реалізують процедуру селекції. Величина F_1 називається свободою вибору, при цьому $F_1 < M$. Виходи цих моделей служать аргументами-входами для конструювання моделей другого ряду.

Крок 4. Знаходять $\bar{\varepsilon}^2(0) = \min\limits_k \bar{\varepsilon}_k^2(0)$.

m-та ітерація

Крок 1. Конструюють часткові описи вигляду

$$Y_k^{(m)} = a_0^{(m)} + a_i^{(m)} X_i + a_j^{(m)} X_j + a_{ij}^{(m)} X_i X_j + a_{ii}^{(m)} X_i^2 + a_{jj}^{(m)} X_j^2, k = 1..F_1(F_1 - 1)/2.$$

Крок 2. Для кожного опису, використовуючи МНК, знаходять відповідні оцінки $\hat{a}_0^{(m)}, \hat{a}_i^{(m)}, \hat{a}_j^{(m)}, \hat{a}_{ij}^{(m)}, \hat{a}_{ii}^{(m)}, \hat{a}_{jj}^{(m)}$.

Крок 3. На перевірній послідовності знаходять для кожного часткового опису значення критерію

$$\bar{\varepsilon}_k^2(m) = \frac{1}{N_{перев}} \sum_{i=1}^{N_{перев}} (Y_i - Y_{ki}^{(m)})^2,$$

де $N_{перев}$ – обсяг перевірної вибірки.

Крок 4. Знаходять $\bar{\varepsilon}^2(m) = \min\limits_k \bar{\varepsilon}_k^2(m)$. Перевіряють умову $\bar{\varepsilon}^2(m) > \bar{\varepsilon}^2(m-1)$, де $\bar{\varepsilon}^2(m), \bar{\varepsilon}^2(m-1)$ – значення критерію точності для найкращих моделей m-го і $(m-1)$-го рядів селекції відповідно. Якщо точність задовільна, то закінчення процедури. Шукана модель вибирається з часткових описів $(m-1)$-го рівня, на якому досягається мінімальна похибка $\bar{\varepsilon}^2(m-1)$. Інакше – перехід до конструювання наступного ряду часткових описів. При цьому проводиться добір (селекція) F_2 кращих описів.

Заключний етап

Рухаючись від кінця до початку і роблячи послідовну заміну змінних, формують вирази для шуканої моделі в початковому просторі описів.

5.5 Запитання і вправи

1) Поясніть, у чому полягає принципова різниця між методом групового врахування аргументів і регресійними методами.

2) Які види часткового описання використовують у МГВА? Яким чином формується структура математичної моделі при використанні МГВА?

3) Яке призначення зовнішнього критерію? У якому процентному відношенні рекомендується розділяти ряд даних на навчальну і перевірну вибірки при побудові моделі та оцінюванні прогнозів? Якими можуть бути

Ваші висновки та пропозиції з цього приводу? (Обґрунтуйте відповідь).

4) До якого типу методів відноситься МГВА – дедуктивних чи індуктивних? Поясніть на прикладі.

5) МГВА відтворює динаміку чи статику процесу? Які основні напрями застосування МГВА?

6) Чи можна використовувати МГВА в реальному часі? Чи можна використовувати моделі, отримані за МГВА, для проектування систем керування? Чому?

7) Наведіть інші можливі приклади застосування МГВА крім прогнозування і керування.

8) Чи можна порівнювати принцип функціонування МГВА з нейронними мережами? Назвіть інші методи моделювання еволюційного типу.

6 ОСНОВИ ПОБУДОВИ БАЙЄСОВИХ МЕРЕЖ

6.1 Методика оцінювання побудови байєсових мереж

Побудова байєсових мереж (БМ) [33, 34] пов'язана з необхідністю розв'язання декількох задач, зокрема це задачі обчислювального характеру, що трапляються при навчанні мережі. У загальному випадку навчання мережі відноситься до NP-повних задач, тобто обсяг обчислень зростає поліноміально із збільшенням числа вузлів (змінних) мережі.

Усей підрозділ присвячено розробленню практичної методики побудови байєсових мереж, яку можна використовувати за наявності достатньої статистичної інформації щодо досліджуваної системи, необхідної для побудови БМ. Пропонована методика може також використовуватися тими дослідниками, хто вже має уявлення про мережі, але не має достатнього досвіду їхньої побудови та застосування. Спочатку розглянемо загальні питання стосовно використання теореми Байєса, а потім перейдемо до загальних принципів побудови та навчання БМ на основі експериментальних (статистичних) даних.

Таким чином, необхідно розробити методику побудови (формування структури) мережі Байєса у вигляді напрямленого ациклічного графа, призначеного для моделювання і візуалізації інформації щодо конкретної задачі, навчання мережі на основі наявної інформації і формування статистичного висновку – прийняття рішення щодо поставленої задачі. БМ можна розглядати як модель подання ймовірнісних залежностей (взаємозв'язків) між його вершинами. Зв'язок $A \to B$ називають причинним, якщо подія A є причиною виникнення B, тобто якщо існує механізм впливу значень змінної A на значення, яких набуває змінна B. БМ називають причинною (каузальною) тоді, коли всі її зв'язки є причинними.

Формально байєсова мережа – це трійка $N = \langle V, G, J \rangle$, першою компонентою якої є множина змінних V; другою – напрямлений ациклічний граф G, вузли якого відповідають випадковим змінним модельованого процесу; J – спільний розподіл імовірностей змінних $V = \langle X_1, X_2, ..., X_n \rangle$.

При цьому стосовно множини змінних виконується марковська умова, тобто кожна змінна мережі не залежить від решти змінних, за винятком батьківських попередників цієї змінної.

Спочатку ставиться задача обчислення значень взаємної інформації між усіма вершинами (змінними) мережі. Потім необхідно знайти оптимальну структуру мережі, використовуючи критерій якості, тобто оцінку описання мережі мінімальної довжини (ОМД), що аналізується й оновлюється на кожній ітерації алгоритму навчання.

6.2 Формування висновку на основі теореми Байєса

Імовірність одночасної появи двох незалежних подій D і S визначається за виразом

$$p(D,S)=p(D)p(S). \qquad (6.1)$$

Якщо події D і S залежні, то поява однієї з них дає певну інформацію про можливість появи другої:

$$p(D,S)=p(D)p(S\mid D),$$

де $p(S\mid D)$ – імовірність появи події S за умови, що вже мала місце подія D. Наприклад, подію D можна інтерпретувати як захворювання, а S – як симптом. Якщо є інформація про те, що пацієнт має якесь захворювання, то можна присвоїти вищу ймовірність появи певного симптому. Ураховуючи комутативність наведеного вище виразу, можна записати

$$p(D,S)=p(S)p(D\mid S)=p(D)p(S\mid D),$$

звідки завжди маємо просту форму теореми Байєса (ТБ)

$$p(D\mid S)=\frac{p(D)p(S\mid D)}{p(S)}.$$

Теорему Байєса можна розглядати як механізм формування висновку (прийняття рішення). Припустимо, що розглядається проста задача постановки діагнозу. У даному разі маємо $p(D\mid S)$ – імовірність захворювання при наявності симптому S, тобто це подія, стосовно якої необхідно сформулювати висновок; $p(D)$ – імовірність захворювання на конкретну хворобу в межах деякої популяції, цю величину можна оцінити на підставі аналізу історії розвитку цієї популяції; $p(S\mid D)$ – імовірність появи симптому, якщо пацієнт уже хворий. Останню величину можна оцінити за допомогою даних, узятих з історій хвороб. Імовірність появи симптому S у вибраній популяції позначимо через $p(S)$; цю величину також можна обчислити на основі статистичних даних, але в цьому, як правило, немає потреби (покажемо це нижче).

Припустимо, що змінна захворювання D має два стани (або може набувати двох можливих значень): D_t – істинне значення ймовірності, яке означає, що пацієнт має хворобу; D_f – неістинне (протилежне) значення. Ці два значення ймовірності дають у сумі 1 незалежно від того, яке значення має S:

$$p(D_t,S)+p(D_f\mid S)=1.$$

Застосуємо до останньої рівності теорему Байєса

$$\frac{p(D_t)p(S\mid D_t)}{p(S)}+\frac{p(D_f)p(S\mid D_f)}{p(S)}=1, \qquad (6.2)$$

або

$$p(S)=p(D_t)p(S\mid D_f)+p(D_f)p(S\mid D_f).$$

Отже, знаючи оцінку $p(S)$, її можна виключити з подальшого розгляду. У цьому прикладі змінна D має тільки два стани, але очевидно, що $p(S)$ можна виключити з розгляду при довільному числі станів D.

Теорему Байєса можна розглядати як вираз (механізм), який об'єднує «апріорну» і «правдоподібну» інформацію, яку запишемо у вигляді

$$p(D|S) = a\ p(D)p(S|D),$$

де $a = 1/p(S)$ – нормуюча константа. Тепер $p(D)$ можна розглядати як апріорну інформацію, оскільки вона була відома до отримання будь-яких вимірів; $p(S|D)$ – правдоподібна інформація, оскільки ми отримуємо її з аналізу (вимірів) симптомів.

Запишемо послідовність дій (алгоритм) щодо формування байєсового висновку на відомій множині конкуруючих гіпотез, які пояснюють множину даних. Для кожної гіпотези необхідно виконати такі дії:

– перетворити апріорну та правдоподібну інформацію, що міститься в даних, у ймовірності;

– перемножити отримані ймовірності;

– нормувати результати з метою отримання апостеріорної ймовірності для кожної гіпотези при наявній інформації;

– вибрати гіпотезу, яка має максимальну ймовірність.

Апріорні знання. У деяких випадках ми можемо обчислити апріорні ймовірності за допомогою статистичних даних. Наприклад, апріорну ймовірність появи захворювання можна визначити в результаті ділення числа випадків захворювання на загальне число пацієнтів, які проходять огляд. Здебільшого це неможливо зробити внаслідок суб'єктивних труднощів отримання статистичних даних, але апріорні знання можуть бути в інших формах. Розглянемо ілюстративний приклад розпізнавання образів.

Приклад 6.1. Розглянемо задачу і принципи розпізнавання двох кіл у цифровому образі, які мають бути розташовані на певній відстані одне від одного. Алгоритми розпізнавання ґрунтуються, як правило, на обчисленні множини ознак і їх порівнянні з відомими. Для розпізнавання зображення кіл можна скористатися багатьма ознаками, але для прикладу виберемо простий варіант розпізнавання. Наприклад, розробимо алгоритм розпізнавання двох кіл у даному образі, які мають однакові радіуси та розміщені на певній відстані S одне від одного. Якщо вдається знайти два суміжних кола, то далі слід установити, чи це саме ті кола, які ми шукаємо.

Припустимо, що центри кіл лежать на відстані $S = 2(r_i + r_j)$, де r_i, r_j – радіуси кіл, знайдених в образі. Для простоти візьмемо, що радіуси однакові. Для кожної пари кіл, знайдених у цифровому образі, обчислимо міру M наближення до шуканої пари кіл за виразом

$$M = \frac{\left|r_i - r_j\right|}{r_i} + \frac{\left|S - 2\left(r_i - r_j\right)\right|}{r_i}.$$

Очевидно, що $M = 0$ при ідеальному узгодженні міри з вибраною парою кіл. Міру M можна перетворити за деякою логікою в імовірність, наприклад, за допомогою розподілу ймовірностей. Таким способом можна знайти суб'єктивну оцінку ймовірності за допомогою обчислених значень міри M .

Альтернативною стратегією є застосування об'єктивних методів. Для цього необхідно виконати деякі експерименти. Для даного прикладу необхідно знайти розміри фігур (кіл) для множини фотографій. Для кожного виміру параметрів двох кіл обчислюємо міру M, а також запитуємо експерта: чи являє собою вибрана пара кіл шукані кола? На основі цього експерименту можна побудувати гістограму та відповідний дискретний розподіл. Отриманий розподіл можна описати певною функцією, наприклад такою:

$$p(M) = a\ exp(-\beta M^2),$$

де параметри α і β розраховуються за допомогою експериментальних даних таким чином, щоб досягти найкращого опису даних (тобто максимізувати функцію правдоподібності). Якоюсь мірою такий розподіл є наближенням до нормального.

Суб'єктивні та об'єктивні ймовірності. Питання вибору суб'єктивного чи об'єктивного підходу до визначення апріорних ймовірностей є ще предметом дебатів між фахівцями в галузі теорії і практики застосування байєсових методів. На перший погляд об'єктивний підхід є надійнішим, але він потребує значних обсягів експериментальних даних, а остаточний результат є досить чутливим до похибок вимірів. Тому значна частина дослідників схиляються до суб'єктивного вибору апріорних ймовірностей. Надалі ми звертатимемося до того чи іншого підходу залежно від особливостей поставленої задачі.

Правдоподібність. Як правило, апріорні ймовірності ґрунтуються на фактах, які знову й знову підтверджуються з плином часу. Їх можна оцінювати на основі відомих обґрунтованих знань щодо проблеми, яка моделюється. Разом з тим експериментальні дані містять, як правило, похибки вимірів (або похибки збору статистичних даних), що призводить до невизначеності, яку виражають через правдоподібність. У прикладі, що розглядається, ці похибки можуть бути пов'язані з методичними та обчислювальними похибками алгоритму розпізнавання образів. Алгоритм розпізнавання не може взяти і виділити коло, але він може сказати, з яким ступенем наближення певна фігура наближається до кола. Наприклад, можна підрахувати число пікселів, що формують коло. Знаючи число пікселів, можна обчислити відповідну ймовірність наближення цієї фігури до кола. Тобто правдоподібність можна обчислити за аналогією з

обчисленням апріорних імовірностей.

Тепер можна сформулювати правило прийняття рішення (висновку) щодо наявності шуканого зображення двох кіл у деякому образі:

$$p(C \mid I) = ap(C)\ p(I \mid C),$$

де $p(C)$ – апріорна ймовірність того, що два кола являють собою шукані кола; вона визначається на основі міри M, а також апріорного знання щодо перетворення M у ймовірність; $p(I \mid C)$ – імовірність отримання необхідної інформації щодо образу за умови, що два кола являють собою шукані кола – це інформація щодо правдоподібності, отримана в процесі оброблення вимірів.

Існують різні погляди на проблему застосування суб'єктивних та об'єктивних методів. Одні школи схиляються до суб'єктивних, а інші – до об'єктивних методів. Суб'єктивний підхід ґрунтується на нашому розумінні предметної галузі та проблеми, на наявних даних; він дає можливість у подальшому сформулювати висновок. З другого боку, об'єктивний підхід може містити елементи суб'єктивізму. Тобто обидві форми можуть суттєво перетинатися щодо здобуття і застосування знань, і це цілком природно. При розв'язанні конкретних задач, по можливості, варто користуватись обома формами з метою виявлення кращої для даного випадку.

Проста мережа Байєса [33, 34]. Розглянутий спрощений підхід до формування байєсового висновку не дає можливості застосовувати його у більш складних ситуаціях оброблення апріорної інформації. Так, у виразі для міри подібності деякого образу до шуканої пари кіл

$$M = \frac{\left|r_i - r_j\right|}{r_i} + \frac{\left|S - 2\left(r_i - r_j\right)\right|}{r_i}$$

обидва члени в правій частині однаковою мірою впливають на значення M, але це не кращий спосіб формування міри. У цю міру можна ввести нові члени, які характеризують, наприклад, колір фону навколо кіл, які ми шукаємо. Тобто складнішою мірою подібності довільного образу до шуканого може бути така:

$$M = a\frac{\left|r_i - r_j\right|}{r_i} + \beta\frac{\left|S - 2\left(r_i - r_j\right)\right|}{r_i} + \gamma \cdot (\textit{ознака кольору}),$$

де α, β і γ – евристичні константи, які можна визначити, наприклад, експертним шляхом. Таким чином, процес аналізу стає евристичним, а тому необхідно спробувати знайти кращий (формальний) метод подання апріорних моделей.

Розглянемо випадок, коли дані щодо проблеми можуть надходити з декількох джерел. Тепер теорема Байєса набуває вигляду

$$p(D \mid S_1, S_2, \ldots, S_n) = \frac{p(D)p(S_1, S_2, \ldots, S_n \mid D)}{p(S_1, S_2, \ldots, S_n)}.$$

У цьому разі виникає проблема оцінювання умовної ймовірності $p(S_1, S_2, ..., S_n \mid D)$ при великих значеннях n. Однак якщо припустити незалежність подій $S_i, i = 1,...,n$ при відомому D, то отримаємо

$$p(S_1, S_2, ..., S_n \mid D) = p(S_1 \mid D)\, p(S_2 \mid D) ... p(S_n \mid D).$$

У результаті подальшого нормування можна позбутися знаменника $p(S_1, S_2, ..., S_n)$, що дещо спрощує задачу формування висновку. Таким чином, отримуємо таке рівняння для формування висновку за теоремою Байєса:

$$p(D \mid S_1, S_2, ..., S_n) = a\ p(D)\ p(S_1 \mid D)\, p(S_2 \mid D) ... p(S_n \mid D).$$

Це рівняння можна зобразити графічно, як показано на рисунку 6.1. На графі змінні подано колами, а стрілки вказують на зв'язок (умовні ймовірності) між незалежними і залежними змінними. Незалежні змінні називають *батьківськими,* або *попередниками*, а залежні – *дитячими,* або *нащадками*.

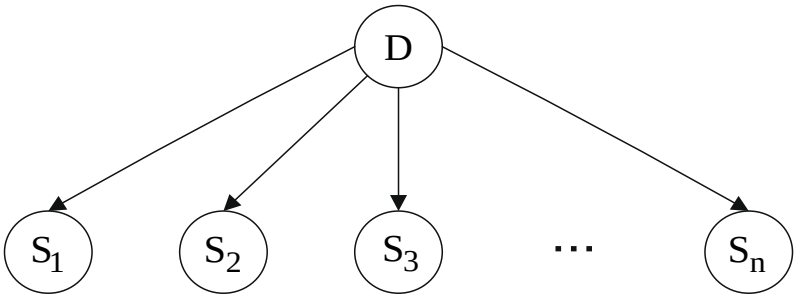


Рисунок 6.1 – Проста («наївна») мережа Байєса

Задачу розпізнавання шуканого образу двох кіл також можна подати у вигляді простої («наївної») мережі Байєса, зображеної на рисунку 6.2. Зазначимо, що використання деревоподібної структури дає можливість точніше виразити вплив кожного члена міри наближення деякого довільного образу до шуканого зображення. Відповідні змінні описано в таблиці 6.1, а висновок можна сформувати за виразом

$$p(M \mid S, D, F) = a\, p(M)\, p(S \mid M)\, p(D \mid M)\, p(F \mid M).$$

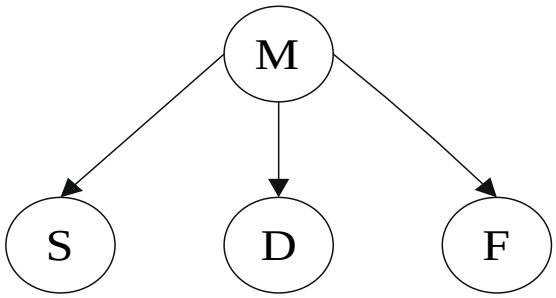


Рисунок 6.2 – Проста мережа Байєса для розпізнавання шуканого образу

Змінні, що характеризують цю задачу, є *дискретними* або *неперервними*. Дискретні змінні набувають одного значення зі скінченної множини значень або один із станів. При цьому кожний стан може бути поданий одним цілим числом або цілим числом у певному діапазоні

значень. Неперервні змінні можуть набувати будь-якого значення у межах певного діапазону значень, їх розглядають як дійсні числа. Мережа Байєса може містити дискретні та неперервні змінні.

За оцінку кольору фону навколо кіл, які являють собою шуканий образ, можна взяти гістограму пікселів для відтінків кольорів у безпосередній близькості до кіл. Це може бути дискретна змінна, яка набуває обмеженої кількості значень.

Таблиця 6.1 – Описання змінних простої мережі Байєса для розпізнавання шуканого образу

Змінна	Інтерпретація	Тип	Значення
M	Міра подібності до шуканого образу	Дискретна (два значення)	Істина або фальш
S	Відстань між центрами кіл	Неперервна	$(S-5(r_i+r_j))/r_i$
D	Різниця у розмірі кіл	Неперервна	$\left\|(r_i+r_j)/r_i\right\|$
F	Колір фону навколо кіл	Дискретна (20 значень)	За наближеною гістограмою пікселів для відтінків кольорів

З іншого боку, відстань між колами – це неперервна змінна, хоча точність її вимірювання можна обмежити точністю розміру пікселя. Можна дещо змінити вираз для визначення ступеня рознесення кіл у просторі, наприклад, увести додатні та від'ємні значення (шляхом видалення модуля):

$$\textit{Рознесення кіл} = \frac{S_i - 2(r_i - r_j)}{r_i} = \frac{2r_i - 2r_i - 2r_j}{r_i} \approx -2$$ при $r_i \approx r_j$ і $S_i = 2r_i$.

Це приведе до того, що міра рознесення кіл змінюватиметься приблизно від – 2 (кола розташовані дуже близько, $S_i = 2r_i$) до 2 (кола лежать далеко одне від одного, $S_i = 3r_i$). Діапазон значень змінної «*рознесення кіл*» можна поділити на будь-яке число станів, але для ілюстрації зупинимось на таких семи станах:

$$\{\textit{менше за } -2{,}0\},\ \{-2{,}0-(-1{,}5)\},\ \{-1{,}5-(-1{,}0)\},\ \{-1{,}0-(-0{,}5)\}.$$
$$\{-0{,}5-0\},\ \{0-0{,}5\},\ \{\textit{більше ніж } 0{,}5\}.$$

Залежно від конкретної постановки задачі число станів змінної можна визначати різними способами, а це вже може бути предметом окремого дослідження.

Кожній дузі мережі Байєса ставиться у відповідність матриця зв'язку – матриця умовних ймовірностей. Матриця, яка зв'язує вузол D з вузлом M, для кожної пари станів має такий вигляд:

$$P(D \mid M) = \begin{bmatrix} p(d_1 \mid c_1) & p(d_1 \mid c_2) \\ p(d_2 \mid c_1) & p(d_2 \mid c_2) \\ p(d_3 \mid c_1) & p(d_3 \mid c_2) \\ p(d_4 \mid c_1) & p(d_4 \mid c_2) \end{bmatrix}.$$

Значення елементів матриць умовних імовірностей можна знайти експериментально. Для цього необхідно мати результати великої кількості дослідів з відомими значеннями всіх змінних. Їх можна отримати шляхом цифрового оброблення реальних образів для вузлів-нащадків (іншими словами, листкових вузлів) S, D, F плюс експертний висновок щодо вузла M.

Отримані таким чином матриці зв'язку являють собою об'єктивні ймовірності, які визначаються так:

$$p(d_3 \mid c_1) = (\text{Кількість появ в образі } d_3 \text{ і } c_1) / \text{Загальна кількість появ } c_1.$$

Очевидно, що навіть для даного простого прикладу кілька значень умовних ймовірностей буде значним. Тому для отримання прийнятних оцінок умовних ймовірностей треба мати великі масиви даних.

Мережу Байєса, що розглядається у даному прикладі, називають по-різному: класифікатор Байєса, наївний класифікатор Байєса та проста мережа Байєса. Це проста і зручна форма мережі, яка застосовується у багатьох практичних задачах. Для того щоб скористатися мережею, слід задати значення змінних, поданих вузлами. Задавання значень вузлам (змінним) називають інстанціюванням. Формування висновку за допомогою мережі, зображеної на рисунку 6.3, можливе після того, як задано значення змінних S, D, F за допомогою інформації (вимірів), що міститься в образі, і вироблених правил дискретизації змінних, як показано вище. Для отримання висновку треба перемножити значення усіх умовних імовірностей для кожного стану M, які беруть з матриць зв'язку. Далі необхідно нормувати результат так, щоб сума умовних імовірностей дорівнювала б одиниці. Таким чином отримаємо ймовірність появи шуканого образу з двох кіл у конкретних експериментальних даних.

Звичайно, що змінні, які входять до мережі, можуть бути взаємозалежними. Так, для прикладу з розпізнаванням зображення двох кіл, змінні S = *«рознесення кіл»* і D = *«різниця у розмірі кіл»* можуть бути певною мірою корельованими. Зокрема, можна виставити контраргументи проти того, що S і D – це дійсно ті змінні, які можна використати для установлення факту наявності двох шуканих кіл у певному образі. Тобто

ідея розпізнавання може бути сформульована дещо по-іншому.

Розглянемо ускладнену мережу, зображену на рисунку 6.3. Ця структура являє собою кращу модель процесу розпізнавання, оскільки вона містить нову семантичну одиницю (вузол) *«кола»*. Тобто такий елемент може бути виявлений в образі, але він не обов'язково зумовлений появою шуканого зображення. Тепер вузол «кола» можна розглядати як загальну причину введення вузлів S = *«рознесення кіл»* і D = *«різниця у розмірі кіл»*, що дає можливість не розглядати проблему їх можливої залежності.

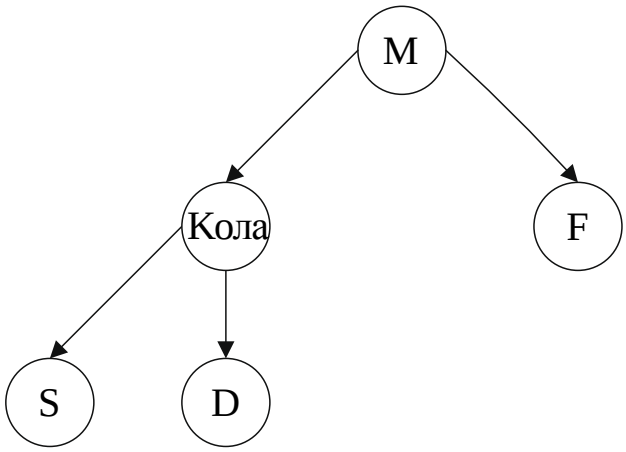


Рисунок 6.3 – Байєсове дерево прийняття рішень

На рисунку 6.3 вузли M і $Кола$ мають матрицю зв'язку $P(Кола/M)$; вузли M і F – матрицю $P(F|M)$; вузли $Кола$ і S – матрицю $P(S|Кола)$, а вузли $Кола$ і D – матрицю $P(D|Кола)$.

Для нового вузла необхідно встановити число його станів. У найпростішому випадку – це дихотомічна змінна з двома станами, але в цьому разі краще ввести три такі стани: o_1=«ймовірно це не шукані кола»; o_2=«це можуть бути шукані кола» і o_3=«ймовірно це шукані кола». Значення елементів матриці зв'язку можна знайти за експериментальними даними, як показано вище, але в цьому разі слід отримати експертну оцінку щодо значення нетермінального вузла O і вузла M, за допомогою якого формується гіпотеза.

Продемонструємо роботу мережі, починаючи з вузла O (кола). За теоремою Байєса маємо

$$p(O|S,D)=\frac{p(O)p(S|O)p(D|O)}{p(S)P(D)}.$$

Однак тут виникає проблема визначення ймовірності $p(O)$ – апріорної ймовірності появи шуканих кіл в образі. У цьому разі O є проміжною змінною, яка не вимірюється, але ймовірності її значень необхідно знати. Ми можемо обчислити правдоподібність значення O за умови, що S і O

отримують певні значення, тобто можна записати

$$l(O|S,D)=\frac{p(S|O)p(D|O)}{p(S)P(D)},$$

або в простішій формі

$$l(O)=a\ p(S|O)p(D|O).$$

Як і раніше, значення $p(S)$ і $p(D)$ можна виключити з розгляду шляхом нормування суми значень $l(O)$ до одиниці. Обчислена у такий спосіб правдоподібність – це ймовірність, обчислена за припущенням, що апріорні ймовірності кожного стану змінної O є однаковими, тобто $p(o_1)=p(o_2)=p(o_3)=1/3$. Тепер для кореневого вузла M можна записати

$$p(M|O,F)=\frac{p(M)p(O|M)p(F|M)}{p(O)P(F)},$$

або простіше

$$p(M|O,F)=a\ p(M)p(O|M)p(F|M).$$

Якщо відомо значення (вимір) F, наприклад $F=f_5$, то з матриці зв'язку можна визначити $p(F|M)$. Однак ми не маємо значення стану змінної O, а тільки оцінку правдоподібності для неї $l(O)$, яка є елементом розподілу можливих станів змінної O. Для того щоб знайти оцінку $p(O|M)$, необхідно знайти середнє цього розподілу. Це можна зробити так:

$$p(o|m_1)=p(o_1|m_1)l(o_1)+p(o_2|m_1)l(o_2)+p(o_3|m_1)l(o_3),$$
$$p(o|m_2)=p(o_1|m_2)l(o_1)+p(o_2|m_2)l(o_2)+p(o_3|m_2)l(o_3).$$

Тепер можна обчислити розподіл імовірностей для M :

$$p'(m_1)=p(m_1|O,f_5)=ap(m_1)\{p(o_1|m_1)l(o_1)+p(o_2|m_1)l(o_2)+p(o_3|m_1)l(o_3)\}\cdot p(f_5|m_1),$$
$$p'(m_2)=p(m_2|O,f_5)=ap(m_2)\{p(o_1|m_2)l(o_1)+p(o_2|m_2)l(o_2)+p(o_3|m_2)l(o_3)\}\cdot p(f_5|m_2);$$

де p' – середня апостеріорна ймовірність, тобто ймовірність набуття змінною певного значення за умови, що відома деяка інформація (у цьому разі це значення F,S,D).

Хоча ми не маємо апріорної ймовірності для вузла O, її можна оцінити за допомогою апріорної (або апостеріорної) ймовірності для M і матриці зв'язку $p(O|M)$. У векторній формі це рівняння має вигляд

$$p(O)=P(O|M)p(M).$$

На відміну від наведеної вище теореми Байєса (у скалярній формі) це є векторне рівняння, тобто $p(o_1)\neq p(o_1|m_2)p(m_2)$.

Припустимо, що $p(M)=\{0,4\ \ 0,6\}$; це означає, що

$$P(O)=\begin{bmatrix} p(o_1 \mid m_1) & p(o_1 \mid m_2) \\ p(o2 \mid m_1) & p(o_2 \mid m_2) \\ p(o_3 \mid m_1) & p(o_3 \mid m_2) \end{bmatrix}\begin{bmatrix} 0{,}4 \\ 0{,}6 \end{bmatrix}=\begin{bmatrix} 0{,}4\,p(o_1 \mid m_1)+0{,}6p(o_1 \mid m_2) \\ 0{,}4\,p(o_2 \mid m_1)+0{,}6p(o_2 \mid m_2) \\ 0{,}4\,p(o_3 \mid m_1)+0{,}6p(o_3 \mid m_2) \end{bmatrix}.$$

Оскільки суми елементів стовпчиків матриці зв'язку дорівнюють одиниці, то цей результат стосується також обчислених значень $p(O)$.

Тепер можна обчислити розподіл ймовірностей для значень станів змінної O за умови, що є виміри, скажімо, $\{s_3, d_2\}$:

$$p(o_1 \mid s_3,d_2)=a\ p(o_1)p(S_3 \mid o_1)p(d_2 \mid o_1),$$
$$p(o_2 \mid s_3,d_2)=a\ p(o_2)p(S_3 \mid o_2)p(d_2 \mid o_2),$$
$$p(o_3 \mid s_3,d_2)=a\ p(o_3)p(S_3 \mid o_3)p(d_2 \mid o_3).$$

а той факт, що

$$p(o_1 \mid s_3,d_2)+p(o_2 \mid S_3,d_2)+p(o_3 \mid S_3,d_2)=1,$$

дозволяє виключити з розгляду α. Очевидно, що наведена процедура обчислення ймовірностей є досить складною і громіздкою, а при збільшенні розмірів мережі вона стає недосяжною для сприймання. Тобто виникає необхідність розроблення спеціальних методів і відповідних обчислювальних алгоритмів для виконання подібних розрахунків.

6.3 Аналіз ефективності функціонування мережі Байєса

Як зазначалося вище, байєсова мережа (БМ) – це ймовірнісна графічна модель причинних зв'язків між якісними та кількісними змінними, яка створюється для описання статики або динаміки об'єктів різної природи з метою формування висновку щодо того чи іншого (поточного) стану досліджуваного об'єкта. Будь-яка нова інформація про об'єкт використовується для оновлення розподілів імовірностей станів, які характеризуються вузловими змінними мережі. На основі оновлених розподілів імовірностей формується статистичний висновок, який дає можливість особі, що приймає рішення (ОПР), прийняти рішення щодо виконання відповідних дій. Перевагою мереж Байєса, якщо порівнювати їх з іншими підходами до врахування і оброблення невизначеностей різної природи, є достатня формалізація усіх етапів їх побудови та використання.

Ступінь успішності застосування цього методу моделювання і формування статистичного висновку залежить від уміння коректно сформулювати постановку задачі, вибрати змінні процесу, які достатньою мірою характеризують його динаміку або статику, зібрати статистичні дані та використати їх для навчання мережі, а також коректно сформувати результат (висновок) за допомогою побудованої мережі.

Оскільки БМ – це допоміжний інструмент при прийнятті рішень, то виникає питання про його ефективність і як вона змінюється у часі. Невизначеності, притаманні окремим змінним і групам змінних БМ, стають

ключовими факторами впливу на рішення, що приймаються за участю БМ.

Ефективність мережі Байєса

Розглянемо БМ, яка складається з n вузлів (рисунок 6.4). Кореневий вузол X_1 зображує змінну, щодо якої формулюється гіпотеза, а вузли без нащадків зображають $r+1$ інформаційну змінну: $X_{n-r},...,X_{n-1},X_n$. Решта вузлів – проміжні, вони допомагають передавати інформацію (свідчення) від інформаційних змінних до головної, для якої формулюється гіпотеза.

Структура мережі вважається встановленою, якщо визначено число вузлів і зв'язки між ними. Надалі необхідно визначити умовні ймовірності, які встановлюють кількісний рівень зв'язків між вузлами, тобто визначають функціональну структуру мережі. Отримання нової інформації від інформаційних змінних дає можливість оцінювати і оновлювати розподіл ймовірностей для основної змінної X_1.

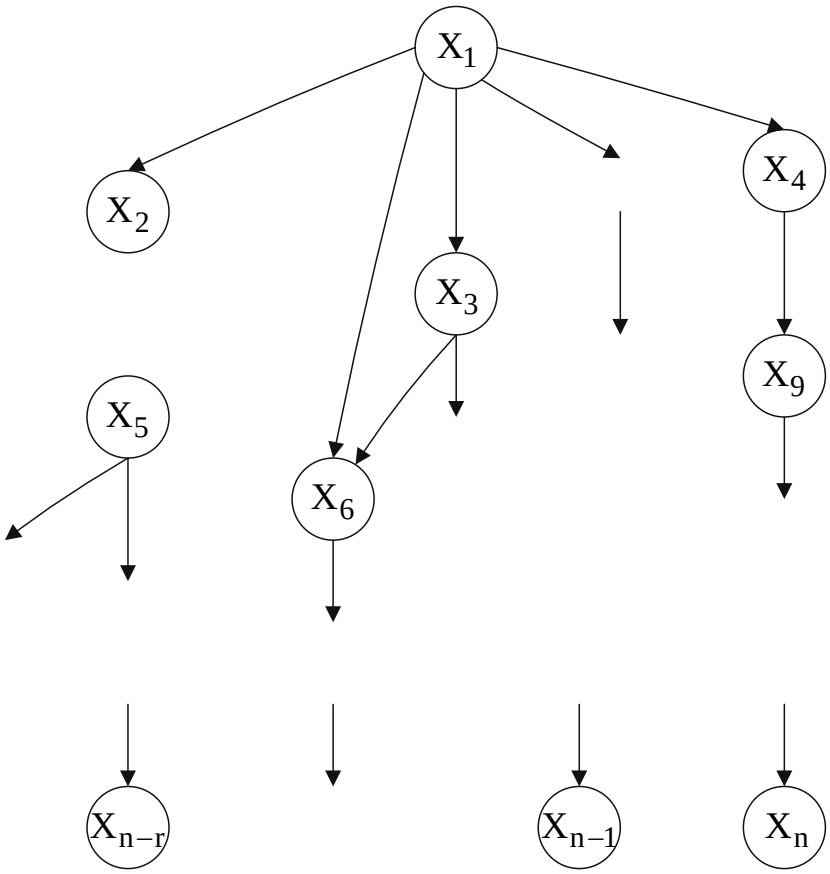


Рисунок 6.4 – Узагальнена структура мережі Байєса з однією змінною, щодо якої формулюється гіпотеза, і множиною інформаційних змінних

Припустимо, що випадкова змінна X_k набуває значення на множині $\{X_{k_1},...,X_{k_m}\}$. Позначимо через $H(X_k) = -\sum_{x_k} .p(x_k)\ ln\ p(x_k)$, тобто додавання тут виконується по всіх значеннях $\{X_{k_1},...,X_{k_m}\}$. Ефективність БМ визначається здатністю використовувати інформацію, що надходить у

процесі функціонування модельованого об'єкта, для оновлення розподілу основної змінної. Очевидно, що така здатність БМ зумовлюється коректністю функціональної специфікації, яка визначає ступінь впливу інформаційних змінних на основну. Таким чином, за міру ефективності можна вибрати міру цього впливу, яка може бути встановлена за допомогою взаємної інформації

$$I(X_1; X_n,..., X_{n-r}) = H(X_n,..., X_{n-r}) - H(X_n,..., X_{n-r} \mid X_1). \quad (6.3)$$

Для конкретного значення $p(x_i \mid y_i)$ випадкових змінних X і Y взаємна інформація $I(X;Y)$ є увігнутою функцією від $p(x_i)$. Аналогічний результат можна отримати і для БМ. Розглянемо першу складову в правій частині рівняння:

$$H(X_n,..., X_{n-r}) = -\sum_{x_n,...,x_{n-r}} p(x_n,..., x_{n-r}) \ln p(x_n,..., x_{n-r}), \quad (6.4)$$

$$p(x_n,..., x_{n-r}) = \sum_{x_n,...,x_{n-r}} p(x_n,..., x_2, x_1) =$$

$$= \sum_{x_n,...,x_{n-r}} p(x_n \mid ba(x_n)) \times p(x_{n-1} \mid ba(x_{n-1})) \times ... \times p(x_2 \mid ba(x_2)) \times p(x_1), \quad (6.5)$$

де $ba(X_i)$ – множина батьківських вузлів для вузла X_i.

Вираз (6.6) записано з урахуванням умовної незалежності змінних, характерною для структури БМ. Оскільки умовні ймовірності $p(x_i \mid ba(x_i))$ являють собою постійні значення, то $p(x_n,..., x_{n-r})$ є лінійною функцією від $p(x_i)$. Таким чином, ентропія $H(X_n,..., X_{n-r})$, яка є увігнутою функцією від $p(x_n,..., x_{n-r})$, буде також увігнутою функцією від $p(x_i)$.

Для другого члена в правій частині (6.3) можна записати

$$H(X_n,...,X_{n-r}) = -\sum_{x_i} p(x_1) \sum_{x_n,...,x_{n-r}} p(x_n,...,x_{n-r} \mid x_1) \ln p(x_n,...,x_{n-r} \mid x_1), \quad (6.6)$$

Тепер скористаємось умовним розподілом

$$p(x_n,..., x_{n-r} \mid x_1) = \frac{p(x_n,..., x_{n-r}, x_1)}{p(x_1)} = \frac{\sum_{x_2,...,x_{n-r}-1} p(x_n,..., x_2, x_1)}{p(x_1)} =$$

$$= \frac{\sum_{x_n,...,x_{n-r-1}} p(x_n \mid ba(x_n)) \times p(x_{n-1} \mid ba(x_{n-1})) \times ... \times p(x_2 \mid ba(x_2)) \times p(x_1)}{p(x_1)} =$$

$$= \sum_{x_n,...,x_{n-r}} p(x_n \mid ba(x_n)) \times p(x_{n-1} \mid ba(x_{n-1})) \times ... \times p(x_2 \mid ba(x_2)).$$

Таким чином, після визначення умовних ймовірностей $p(x_i \mid ba(x_i))$ величина $p(x_n,...,x_{n-r} \mid x_1)$ залишається фіксованою, а ентропія $H(X_n,...,X_{n-r} \mid X_1)$ – лінійна функція $p(x_i)$. Оскільки $I(X_1, X_n,..., X_{n-r})$ – це різниця між увігнутою функцією від $p(x_1)$ і лінійною функцією від $p(x_x)$, то вона буде увігнутою функцією від $p(x_1)$.

Якщо змінна, щодо якої формулюється гіпотеза, може мати h альтернатив, тобто $X_1 = \{x_{1_1},...,x_{1_h}\}$, то розподіл імовірностей $p(X_1)$ визначається числами $\{p(x_{1_1}),...,p(x_{1_h})\}$. Множина всіх таких функцій розподілу утворює симплекс вимірності h у просторі R^h, який визначається співвідношеннями

$$0 \le p(x_{1_i}) \le 1,\ i = 1,2,...,h \text{ і } \sum_{i=1}^{h} p(x_{1_i}) = 1.$$

Різниця $I(X_1, X_n,..., X_{n-r})$ – увігнута функція, визначена на цьому симплексі. Отже, існує така зв'язана підмножина S цього симплекса, на якій взаємна інформація є константою, що відповідає глобальному максимуму. Тобто, якщо S – одна точка, то функція $I(X_1, X_n,..., X_{n-r})$ має єдиний глобальний максимум у цій точці.

Апріорний розподіл імовірностей $p(X_1)$ для змінної, щодо якої формулюється гіпотеза, відображує характер поточної ситуації на будь-якому етапі прийняття рішень. Зміна ситуації приводить до зміни параметрів розподілу на розглянутому вище симплексі. Можна зробити такий висновок: якщо $p(X_1)$ належить множині S, то функціонально визначена мережа дає можливість максимально використати всю інформацію, зібрану на поточний момент. У міру того як розподіл $p(X_1)$ зміщується відносно S, відображаучи тим самим еволюцію ситуації, здатність мережі Байєса використовувати нову інформацію зменшується. Для того щоб повернутися до оптимального режиму роботи, необхідно змінити її функціональну специфікацію таким чином, щоб розподіл $p(X_x)$ став частиною множини S або лежав на її межі.

Ця зміна потребує виконання таких дій (однієї або обох):

- змінити множину спостережуваних змінних таким чином, щоб зібрана інформація мала вищий ступінь наближення до явно вираженої ситуації;
- змінити положення проміжних вузлів і тим самим зв'язки між елементами мережі таким чином, щоб розповсюдження свідчень (інформації) по мережі краще відповідало новій поточній ситуації.

Необхідно підкреслити, що увігнутість функції $I(X_1, X_n,..., X_{n-r})$ – це важливий факт, який свідчить про те, що БМ є найбільш ефективною стосовно деякої унікальної ситуації або множини ситуацій, що відповідають унікальній точці або зв'язаній області ймовірнісного симплекса. Також

необхідно підкреслити, що БМ часто має суб'єктивну природу, вона відображує процес прийняття рішення конкретною особою. У зв'язку з цим природно виникають запитання: Чи є цей процес ефективним і внутрішньо консистентним? Чи коректно відображує створений ланцюжок причинних зв'язків дійсний перебіг ситуацій? На ці запитання можна дати стверджувальну відповідь, якщо взаємна інформація $I(X_1, X_n, ..., X_{n-r})$ набуває максимального значення у точці ймовірнісного симплекса, яка наближається до точки, що відповідає поточній ситуації. Якщо ж ці точки розміщуються далеко одна від одної, то суб'єктивний процес створення мережі був неефективним і, можливо, внутрішньо неконсистентним.

Таким чином, графічні моделі у вигляді мереж Байєса являють собою зручний і важливий інструмент аналізу невизначеностей різної природи, зокрема невизначеностей статистичного і структурного типів. Зображення моделі процесу у вигляді графа дає можливість швидко осмислювати ситуації, наглядно подавати взаємодію елементів (змінних) і формувати ймовірнісний висновок щодо вибраних змінних. При розробленні моделі процесу потрібно відрізняти якісні аспекти від кількісних. Це дає можливість зосередитися на першому етапі аналізу процесу (об'єкта), на побудові причинної структури мережі, не беручи до уваги ймовірнісні аспекти взаємодії елементів процесу. При цьому необхідно тільки чітко розуміти причини та можливі наслідки тих чи інших дій. На другому етапі побудови БМ треба визначити умовні ймовірності для зв'язків, тобто побудувати таблиці умовних ймовірностей. Для розв'язання цієї задачі можна скористатися статистичними даними (якщо це можливо) або суб'єктивними знаннями експертів. Очевидно, що обидва способи потребують додаткових досліджень.

6.4 Особливості методу байєсівського оцінювання ймовірностей

Надалі розглядатимемо системи зі скінченним числом дискретних змінних, що потребуватиме тільки простих визначень та елементарних понять з теорії ймовірностей. Пояснення для неперервних змінних подаються окремо, але вони застосовуватимуться в обмеженому обсязі. Для поглибленого ознайомлення з математичним апаратом теорії ймовірностей можна скористатися ґрунтовними посібниками, наприклад [31]. У цьому підрозділі наведено стислий огляд елементарних імовірнісних понять; особливу увагу приділено формуванню байєсового висновку та його зв'язку з психологією людського мислення в умовах невизначеності, що зазвичай не підкреслюється у посібниках.

Також будемо дотримуватися байєсівської інтерпретації ймовірності, згідно з якою ступені довіри до подій визначаються через імовірності й ці величини використовуються для посилення, оновлення чи послаблення ступеня довіри. При такій формалізації пропозиційним висловлюванням (твердження, що має істинне чи хибне значення) на певній мові

присвоюються визначені ступені довіри, які компонуються й обробляються за правилами ймовірнісного числення. Надалі не робитимемо різниці між пропозиційними висловлюваннями та фактичними подіями, відображеними цими висловлюваннями. Наприклад, якщо A означає: «*Василь Коваленко виставлятиме свою кандидатуру в президенти у 2015 році*», то $p(A|K)$ означає власну віру людини в подію, описану як A, при даній сукупності знань K, яка повинна включати уявлення цієї людини про сучасну внутрішню економічну та соціальну політику, відносини з Європейським Союзом, Росією, а також щодо самої особистості. У визначальному ймовірнісному виразі частіше пишуть просто $p(A)$, опускаючи символ K. Одначе коли вхідна інформація може змінюватися, необхідно конкретно визначити припущення, які відповідають за наші переконання і явно становлять умови K (або деякі їхні елементи).

У байєсовій формалізації ступені довіри задовольняють умови основних аксіом імовірнісного числення:

$$0 \le p(A) \le 1, \tag{6.7}$$

$$p\ (\text{істинний вислів}) = 1 \tag{6.8}$$

$p(A \text{ або } B) = p(A) + p(B)$, якщо A і B є взаємовиключальним. (6.9)

Третя аксіома стверджує, що ступінь довіри до будь-якої множини подій є сумою ступенів довіри до його компонент, що не перетинаються. Тому будь-які подія A може бути записана як об'єднання сумісних подій $(A \wedge B)$ і $(A \wedge -B)$, а відповідні їм імовірності мають вигляд

$$p(A) = p(A,B) + p(A,\neg B), \tag{6.10}$$

де $p(A,B)$ – короткий запис виразу $p(A \wedge B)$.

Узагалі, якщо B_i, де $i = 1, 2, ..., n$, – множина взаємовиключальних тверджень, і не перетинаються й об'єднання яких є достовірним (їх називають змінною), тоді $p(A)$ можна обчислити за допомогою $p(A, B_i),\ i = 1, 2, ..., n$ як суму:

$$p(A) = \sum_i p(A, B_i), \tag{6.11}$$

яка відома під назвою «формула повної ймовірності».

Операція додавання ймовірностей по всіх B_i також називається «*маржиналізацією по* B», а результат – імовірність $p(A)$ – називається *безумовною ймовірністю події* A. Наприклад, $p(A)$ = «Результати кидання двох кубиків однакові», може бути обчислена додаванням за сумісними подія* $(A \wedge B_i), i = 1,2,...,6$, де B означає «Результат кидання першого кубика it. Тобто ця ймовірність

$$p(A) = \sum_i p(A, B_i) = 6 \times \frac{1}{36} = \frac{1}{6}. \tag{6.12}$$

Прямим наслідком з (6.2) і (6.4) є те, що загальна істинність

висловлювання і його заперечення повинні дорівнювати одиниці:

$$p(A)+p(\neg A)=1, \tag{6.13}$$

тому що одне з двох тверджень є однозначно істинним.

Елементарними виразами в байєсівській формалізації є визначення умов ймовірності. Наприклад, $p(A|B)$ показує ступінь довіри до A за умови, що B відомо з абсолютною достовірністю. Якщо $p(A|B)=p(A)$, ми говоримо, що A та B *незалежні*, оскільки ступінь довіри A не змінюється з отриманням інформації про істинне значення B. Якщо $p(A|B,C)=p(A|C)$, ми говоримо, що символи A і B *умовно незалежні* при заданому C. Таким чином, якщо відомо C, то ступінь довіри до A не змінюватиметься з визначенням істинного значення B.

На противагу традиційній практиці визначення умовних ймовірностей у термінах сумісних подій за виразом

$$p(A|B)=\frac{p(A,B)}{p(B)}. \tag{6.14}$$

Філософія байєсового методу розглядає умовні відношення як більш фундаментальні, ніж просто в рамках сумісних подій, тобто більш схожі з організацією людських знань. З цієї точки зору B є вказівкою на контекст, систему або фрейм знань, а $A|B$ означає подію A в контексті, що задає B (наприклад, симптом A в контексті захворювання B). Тобто емпіричне знання незмінно виражатиметься в умовно-ймовірнісних твердженнях, тоді як ступінь довіри сумісних подій визначатиметься з цих тверджень через добуток:

$$p(A,B)=p(A|B)p(B), \tag{6.15}$$

що еквівалентно (6.14). Наприклад, отримання оцінки

$$p(A,B_i)=\frac{1}{36}$$

безпосередньо за допомогою (6.6) було б частково неприродним.

Розумовий процес, що лежить в основі такого оцінювання, передбачає незалежність двох результатів, тобто для точності цього оцінювання ймовірність сумісних подій (еквівалентність, B_i) має бути визначена з умовної події (еквівалентність, B_i) через добуток:

$$p(\text{еквівалентність}\,|\,B_i)p(B_i)=p(\text{результом другого кубика є}\ i\,|\,B_i)p(B_i)=$$

$$=\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}.$$

Як показано в (6.12), імовірність будь-якої події A можна визначити приведенням її до будь-якого набору взаємовиключальних подій $B_i, i=1,2,...,n$ і обчисленням суми:

$$p(A)=\sum_i p(A|B_i)p(B_i). \tag{6.16}$$

Це розкладання забезпечує базис для гіпотетичного (або такого, що грунтується на припущеннях) міркування. Воно вказує на те, що ступінь довіри до будь-якої події A є зваженою сумою ступенів довіри до всіх різних варіантів, через які може бути виражене A. Наприклад, якщо необхідно обчислити ймовірність того, що результат X першого кубика буде більшим за результат Y другого, то можна зумовити подію $A: X > Y$ усіма можливими значеннями X та отримати ймовірність

$$p(A) = \sum_{i=1} p(Y < X \mid X = i)p(X = 1) = \sum_{i-1} p(Y < i)\frac{1}{6} = \sum_{i=1}^{6}\sum_{j=1}^{i=1} p(Y = j)\frac{1}{6} = \frac{1}{6}\sum_{i=2}^{6}\frac{i-1}{6} = \frac{5}{12}.$$

Потрібно ще раз підкреслити, що формули типу (6.16) завжди подаються такими, що можуть бути застосовані в деякому ширшому контексті K, який визначає припущення, котрі вважаються загальнозрозумілими (наприклад, чесність кидання кубиків). Насправді рівняння (6.16) є скороченим записом і для твердження

$$p(A \mid K) = \sum_{i} p(A \mid B_i, K)p(B_i \mid K). \qquad (6.18)$$

Це рівняння випливає з того факту, що кожна умовна ймовірність $p(A \mid K)$ сама є справжньою ймовірнісною функцією, а тому воно задовольняє (6.18).

Іншим корисним узагальненням теореми добутку ймовірностей є *ланцюгове правило*. Воно стверджує: якщо існує набір *n* подій $E_1, E_2, ..., E_n$, то ймовірність сумісної події $E_1, E_2, ..., E_n$ можна записати як добуток n умовних імовірностей:

$$p(E_1, E_2, ..., E_n) = p(E_n \mid E_{n-1}, ..., E_2, E_1)...p(E_2 \mid E_1)P(E_1). \qquad (6.19)$$

Цей добуток можна отримати багаторазовим застосуванням виразу (6.9) у будь-якому зручному порядку.

Байєсівський висновок ґрунтується на відомій формулі оберненого перетворення

$$p(H \mid e) = \frac{p(e \mid H)p(H)}{p(e)}, \qquad (6.20)$$

яка стверджує, що ступінь довіри до гіпотези H за отриманою інформацією e можна обчислити множенням попереднього ступеня довіри $P(H)$ на правдоподібність $P(e \mid H)$ того, що e має місце, якщо H істинна.

Вираз $P(e \mid H)$ іноді називають *апостеріорною ймовірністю, а* $P(H)$ – *апріорною ймовірністю*. Знаменник $p(e)$ у (6.20) практично не береться до уваги, тому що він є просто нормуючою константою

$$p(e) = p(e \mid H)p(H) + p(e \mid \neg H)p(\neg H),$$

яка може бути обчислена з урахуванням того, що $P(H \mid e)$ і $P(\neg H \mid e)$ у сумі становлять одиницю.

Оскільки формально (6.20) має бути виключеним з визначення умовної ймовірності як тавтологічне висловлювання

$$p(A|B)=\frac{p(A,B)}{p(B)} \text{ і } p(B|A)=\frac{p(A,B)}{p(A)}, \quad (6.21)$$

то байєсівський суб'єктивізм розглядає (6.20) як нормативне правило для коригування ступенів довіри. Іншими словами, незважаючи на те що умовні ймовірності можуть розглядатися як чисто математичні конструкції, як у (6.21), байєсівський імовірнісним розглядає їх як прості елементи мови і як точні переклади висловлювання «..., за умови, що я знаю A». Байєсівська ймовірність не є означенням, а рідше взаємовідношенням між висловлюваннями, яке перевіряється емпірично. Воно стверджує, окрім того, що достовірність, яку особа приписує B після визначення A, ніколи не буває нижчою за достовірність, що приписується $A \wedge B$ до визначення A. Також співвідношення між цими двома достовірностями зростатиме пропорційно ступеню неочікуваності $[p(A)]^{-1}$, який особа асоціює з визначенням A.

Суть правила Байєса (див. рівняння (6.20)) легко розкривається при використанні поняття відношення правдоподібності. Поділивши (6.20) на додаткову форму для $p(\neg H|e)$, отримаємо

$$\frac{p(H|e)}{p(\neg H|e)}=\frac{p(e|H)}{p(e|\neg H)}\frac{p(H)}{p(\neg H)}. \quad (6.22)$$

Визначаючи *апріорну нерівність* ймовірностей для гіпотези H як

$$O(H)=\frac{p(H)}{p(\neg H)}=\frac{p(H)}{1-p(H)} \quad (6.23)$$

і відношення правдоподібності як

$$L(e|H)=\frac{p(e|H)}{p(e|\neg H)}, \quad (6.24)$$

знайдемо, що *апостеріорна нерівність*

$$O(H|e)=\frac{p(H|e)}{p(\neg H|e)} \quad (6.25)$$

має вигляд добутку

$$O(H|e)=O(H)L(e|H). \quad (6.26)$$

Таким чином, правило Байєса стверджує, що загальний ступінь довіри до гіпотези H, яка ґрунтується на попередньо отриманих знаннях K і спостережуваних фактах e, має дорівнювати добутку двох множників апріорної нерівності $O(H)$ і відношення правдоподібності $L(e|H)$. Перший множник характеризує *прогнозовану чи очікувану* основу щодо H тільки за основними відомостями, тоді як другий являє собою *ретроспективну або діагностичну* основу гіпотези H за даними фактичних спостережень.

Приклад 6.2. Аналіз якості на виробництві: дискретні дані та дискретні параметри (висновок щодо пропорціональності).

Компанія *А-комп* є виробником плат оперативної пам'яті (оперативні запам'ятовуючі пристрої, або ОЗП) для персональних комп'ютерів.

Виробництво модулів пам'яті сягає сотень тисяч одиниць на місяць.

Позначимо через θ ймовірність виготовлення дефектного модуля пам'яті для серії з n одиниць. Для простоти вважатимемо, що θ може набувати три можливих значення: 0,25 – хороший результат; 0,50 – прийнятний результат; 0 – поганий результат.

Статистика виробничого процесу компанії *А-комп* свідчить про те, що протягом п'яти попередніх років виробництва ОЗП отримувались такі показники якості процесу: протягом 60 % часу вироблялися пристрої з імовірністю дефекту $\theta = \boldsymbol{0,25}$; протягом 30% часу вироблялися пристрої з імовірністю дефекту $\theta = \boldsymbol{0,50}$ і протягом 10% часу з попередніх п’яти років вироблялись пристрої з імовірністю дефекту $\theta = \boldsymbol{0,75}$. Компанія вирішила скористатися цими результатами як апріорними ймовірностями для того, щоб спрогнозувати рівень якості продукції на майбутнє. Значення цих апріорних імовірностей наведо в таблиці 6.2.

Після виробництва 10000 пристроїв оперативної пам'яті компанія вирішила перевірити цю партію, щоб з'ясувати, чи підтримується якість на попередньому рівні.

Виявилося, що з трьох пристроїв пам'яті, випадково обраних для перевірки, два пристрої мають дефекти. Який висновок щодо якості продукції необхідно зробити, тобто яким є апостеріорний розподіл для θ ?

Таблиця 6.2 – Апріорні ймовірності

Значення ймовірності	Якість продукції		
	хороша	прийнятна	погана
Імовірність дефекту θ	0,25	0,50	0,75
Щільність імовірностей для θ $\boldsymbol{(\ p(\theta)\ )}$	0,60	0,30	0,10

У цьому разі дані мають дискретний характер. Припустимо, що вони мають біноміальний розподіл з параметром θ; число успішних подій $\boldsymbol{r = 2}$ (успішними вважаємо події, пов'язані з випуском дефектних пристроїв) з трьох можливих $\boldsymbol{n = 3}$. Таким чином, функція правдоподібності має вигляд

$$\boldsymbol{L(\theta) = C_2^3 \theta^2 (1-\theta),}$$

де $\theta = \boldsymbol{[0,25;\ 0,5;\ 0,75]}$. У цьому разі

$$\boldsymbol{h(\theta \mid r,n) \infty L(r \mid \theta, n) g(\theta).}$$

Обчислимо правдоподібність $\boldsymbol{L(\theta)}$:

– якщо $\theta = \boldsymbol{0,25}$, то $\boldsymbol{L(\theta) = \dfrac{1 \cdot 2 \cdot 3}{1 \cdot 2} (0,25)^2 (1-0,25) = 0,140625;}$

– якщо $\theta = \boldsymbol{0,50}$, то $\boldsymbol{L(\theta) = \dfrac{1 \cdot 2 \cdot 3}{1 \cdot 2} (0,50)^2 (1-0,50) = 0,375000\ ;}$

– якщо $\theta = \boldsymbol{0,75}$, то $\boldsymbol{L(\theta) = \dfrac{1 \cdot 2 \cdot 3}{1 \cdot 2} (0,75)^2 (1-0,75) = 0,41875\ .}$

Апріорні та апостеріорні щільності ймовірності для h дано у таблиці 6.3.

Таблиця 6.3 – Апостеріорні ймовірності

θ	Апріорні ймовірності для θ	Правдоподібність $L(\theta)$	(Апріорні ймовірності)× ×(Правдоподібність)	$h =$ Апостеріорна щільність для θ
0,25	0,60	0,140625	0,0843750	0,35294
0,50	0,30	0,375000	0,1125000	0,47059
0,75	0,10	0,421875	0,0421875	0,17647
Загальна сума:			0,2390625	0,99999

Елементи у стовпчику h знайдено в результаті ділення значень третього стовпчика *((апріорна ймовірність) x (правдоподібність))* на загальну суму для цього (передостаннього) стовпчика з метою нормування остаточного результату до одиниці, тобто

$$p(\theta = 0,25 \mid r,\ n) = h(0,25 \mid 2,3) = \frac{0,60 \cdot 0,140625}{0,6 \cdot 0,140625 + 0,3 \cdot 0,375 + 0,1 \cdot 0,421875} =$$

$$= \frac{0,084375}{0,2390625} = 0,35294 \ ;$$

$$p(\theta = 0,50 \mid r, n) = h(0,50 \mid 2,3) = \frac{0,30 \cdot 0,375}{0,6 \cdot 0,140625 + 0,3 \cdot 0,375 + 0,1 \cdot 0,421875} = 0,47059;$$

$$p(\theta = 0,75 \mid r,\ n) = h(0,75 \mid 2,3) = \frac{0,10 \cdot 0,421875}{0,6 \cdot 0,140625 + 0,3 \cdot 0,375 + 0,1 \cdot 0,421875} =$$

$$= \frac{0,041875}{0,2390625} = 0,17647 \ .$$

З таблиці 6.3 видно, що найбільш імовірним значенням θ, на основі аналізу якості останньої партії пристроїв пам'яті є $\theta = 0,5$, тобто виробництво має «прийнятний» рівень якості готової продукції.

Очевидно, що перед тим як зробити висновок щодо модифікації виробничою процесу з метою підвищення якості продукції, компанія *А-комп* має тестувати більшу кількість блоків пам'яті, ніж $n = 3$; наприклад, $h = 100, 500$ або 1000.

Приклад 6.3. Група студентів складається з трьох підгруп. Для тестування необхідно вибрати студента з однієї із цих трьох підгруп.

Отже, для тестування необхідно вибрати одного студента з групи, яка складається з трьох підгруп. Задача полягає у тому, щоб визначити: *з якої підгрупи вибрали студента для тестування*.

Спочатку розглянемо, яку апріорну інформацію можна використати. Припустимо, що в минулому році всі три підгрупи навчалися з тими ж викладачами, що і поточного року; екзамен був таким же, а результати тестування студентів у певний момент часу були такими, як наведено в таблиці 6.4. З цієї таблиці видно що оцінки студентів мають нормальний

розподіл із середніми значеннями m_i і дисперсією $S_i^2 = 225,\ i = 1,\ 2,\ 3$.

Таблиця 6.4 – Моделі тогорічних оцінок студентів

Модель	Середнє	Параметри розподілу	Апріорна ймовірність p{M}
M_1	$\theta_1 = 74$	N(74, 225)	0,25
M_2	$\theta_2 = 81$	N(81, 225)	0,50
M_3	$\theta_3 = 68$	N(68, 225)	0,25

Для того щоб допомогти передбачити оцінки, які будуть отримані в поточному році, необхідно присвоїти апріорні ймовірності оцінкам поточного року на основі інформації про торішні результати та поточної інформації про студентів. Випадково виберемо студента A, який має хороші поточні оцінки. Також відомо, що кращою серед трьох підгруп є друга підгрупа, яка характеризується моделлю M_2. На основі наявної інформації ми присвоюємо апріорні ймовірності трьом моделям таким чином, як показано в четвертому стовпчику таблиці 6.4.

Правдоподібність (щільність імовірностей) для нормального розподілу визначається за виразом

$$L(x \mid \theta_i) = \frac{1}{15\sqrt{2\pi}} exp\left\{(-0,5)\left(\frac{x-\theta_i}{15}\right)^2\right\},\ i = 1,\ 2,\ 3.$$

Оскільки у поточному році студент A набрав 76 % від максимальної оцінки, то для обчислення функцій правдоподібності скористаємось оцінкою $x = 0,76$. Знайдемо z-оцінки (нормовані або стандартизовані випадкові змінні) за виразом

$$z_i = \frac{x-\theta_1}{15}.$$

Оскільки $z_i \approx N(0,1)$ зі стандартною нормальною щільністю розподілу

$$\Phi(z_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_1^2},$$

то значення правдоподібності можна визначити за допомогою таблиці площ стандартного нормального розподілу. Відповідні результати наведено в таблиці 6.5.

З таблиці 6.5 видно, що модель M_2 має найвищу апостеріорну ймовірність з трьох моделей. Таким чином, робимо висновок, що випадково вибраний для тестування студент навчається у третій групі.

Зазначимо, що в даному прикладі ми скористалися тільки одним експериментальним значенням.

Таблиця 6.5 – Апостеріорні ймовірності для моделей оцінок студентів

Модель	z_i	$\Phi(z_i)$	Правдопо-дібні $\Phi(z_i)/15$	Апріорні ймовірно-сті	(Правдоподібні) х × (апріорні)	Апосте-ріорні
M_1	0,133	0,395	0,02633	0,25	0,0065825	0,264
M_2	-0,333	0,377	0,02513	0,50	0,0125650	0,504
M_3	0,533	0,346	0,02307	0,25	0,0057675	0,231
Загальна сума					0,0249150	0,999

Можна було б поставити задачу визначення підгрупи для кількох студентів, які належать до однієї підгрупи. Це збільшило б обсяг експериментальних даних. Більше того, оскільки вибірка складалася тільки з одного значення (оцінка студента A), то це мало несуттєвий вплив на зміну апріорних імовірностей, тобто апріорні та апостеріорні ймовірності майже однакові. Якби вибірка даних була більшою, то апріорні ймовірності не мали б такого значного впливу на результат.

6.5 Запитання і вправи

1) Поясніть суть теореми Байєса.

2) Поясніть зручність використання моделей процесів різної природи у вигляді напрямлених графів. Які типи змінних можуть бути використані для побудови моделей у вигляді графів?

3) Дайте формальне визначення байєсової мережі. Як Ви розумієте вираз «J-спільний розподіл ймовірностей змінних $V = \{X_1, X_2, ..., X_n\}$»?

4) Наведіть приклад формування ймовірнісного висновку на основі простої форми теореми Байєса

$$p(D \mid S) = \frac{p(D)p(S \mid D)}{p(S)}.$$

Сформулюйте послідовність дій, необхідну для формування байєсового висновку.

5) Поясніть різницю між суб'єктивними і об'єктивними ймовірностями. Як Ви розумієте термін «правдоподібність»?

6) Наведіть приклад мережі Байєса, яка містить батьківські та дитячі змінні. Поясніть на цьому прикладі суть матриці умовних ймовірностей.

7 SOFT COMPUTING

7.1 Методи оброблення нечіткої інформації

Методи теорії нечітких множин поряд із нейронними мережами та методами еволюційного моделювання належать до парадигми „Soft Computing". Таку назву для вказаних технологій визначив професор Каліфорнійського університету Лотфі Заде (Lotfi A. Zadeh) [35, 36], який і дав початковий поштовх аналізу нечіткої і неповної інформації, опублікувавши у 1965 році статтю „Fuzzy Sets" у восьмому номері журналу „Information and Control".

Нехай Ω – універсальна множина, що описує предметну область, елемент $x \in \Omega$. Підмножина є $A \subset \Omega$ набором пар $A = \{(x, \mu_A(x))\}$, де

$$\mu_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

***Означення* 7.1.** Нечітка підмножина $A \subset \Omega$ є набором пар $A = \{(x, \mu_A(x))\}$", де $x \in \Omega$ і $\mu_A : \Omega \to [0;1]$ – функції належності, які є суб'єктивною мірою відповідності елемента x нечіткій підмножині A.

Означення 7.2. Опускаючи деякі граничні випадки, висотою нечіткої множини будемо вважати $h = \max\limits_{x \in \Omega} \mu_A(x)$.

Означення 7.3. Нечітку множину називають нормальною, якщо $h = 1$, в іншому випадку – субнормальною.

***Означення* 7.4.** Нечітка множина називається унімодальною, якщо $\mu_A(x) = 1$ лише для одного $x \in \Omega$.

Означення 7.5. Носієм нечіткої множини A є звичайна множина $\sup p(A) = \{ x \mid x \in \Omega \mu_A(x) > 0 \}$.

Нечіткі множини і відповідні функції належності можуть мати дискретну і кусково-неперервну форми запису. Дискретна форма запису найчастіше є такою:

$$A = \left\{ \frac{\mu_A(x_1)}{x_1}, \frac{\mu_A(x_2)}{x_2}, ..., \frac{\mu_A(x_n)}{x_n} \right\},$$

де $\mu_A(x_i), i = 1,n$ – значення функції належності елемента $x_i \in \Omega$ нечіткій множині A. Неперервна форма запису у загальному випадку є такою:

$$\mu_A(x) = \begin{cases} \mu_{A_1}, x \in A_1, \\ \mu_{A_2}, x \notin A_2, \\ ... \\ \mu_{A_m}, x \notin A_m, \end{cases}$$

де $A = A_1 \cup A_2 \cup ... \cup A_m$, $A_i \cap A_j = \Theta$, $i \neq j$, μ_{A_i} – неперервна функція належності елемента $x \in \Omega$, *м*ножині A_i, $i = \overline{1,m}$. Такі неперервні функції належності можуть бути трикутними (параметри a і c), трапецієподібними різного вигляду (у загальному випадку п'ять параметрів – $(\underline{m}, \overline{m}, \alpha, \beta, h)$, дзвоноподібними (з двома параметрами a і b), гауссівськими (з двома параметрами m і σ) (рисунок 7.1) та ін.

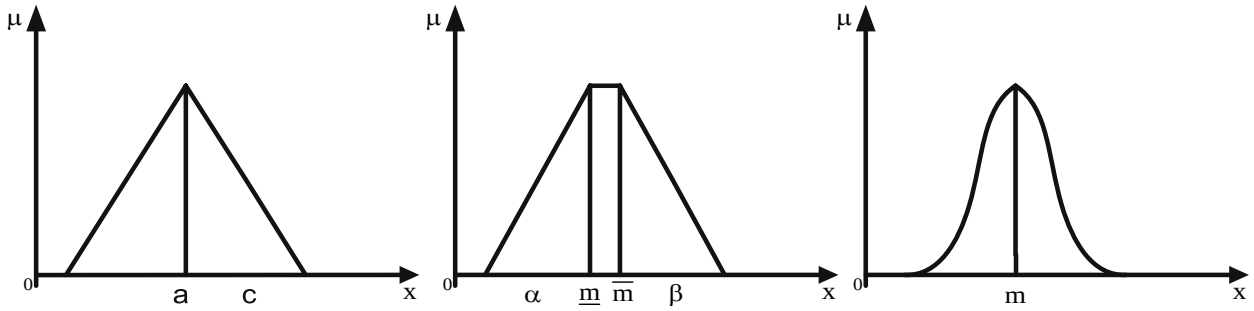


Рисунок 7.1 – Найбільш поширені функції належності

Нечіткі множини мають такі властивості:

1. Якщо $\mu_A(x) = 0 \forall x \in \Omega$, то A – порожня.

2. Якщо $\mu_A(x) = \mu_b(x) \forall x \in \Omega$, A і B – еквівалентні нечіткі множини.

3. Якщо $\mu_A(x) \leq \mu_B(x) \forall x \in \Omega$, то $A \subseteq B$, де $\mu_A(x) = \mu_b(x) \forall x \in \Omega$.

Операції над нечіткими множинами такі:

1. Доповненням $\overline{A}$ нечіткої множини A називається нечітка множина з функцією належності

$$\mu_{\overline{A}}(x) = 1 - \mu_A(x) \forall x \in \Omega.$$

2. Перетином $A \cap B$ нечітких множин A і B називається нечітка множина з функцією належності

$$\mu_{A \cap B}(x) = \mu_A(x) \wedge \mu_B(x) \forall x \in \Omega.$$

3. Об'єднанням $A \cup B$ нечітких множин називається нечітка множина з функцією належності

$$\mu_{A \cup B}(x) = \mu_A(x) \vee \mu_B(x) \forall x \in \Omega.$$

Означення 7.6. Нечітким числом називається опукла нормальна нечітка множина з кусково-неперервною функцією належності, що задана на множині дійсних чисел.

Принцип узагальнення Заде [38, 39]. Якщо $u = f(x_1, x_2, ..., x_n)$ – функція від n незалежних змінних і аргументи $x_1, x_2, ..., x_n$ задані нечіткими числами $\hat{x}_1, \hat{x}_2, ..., \hat{x}_n$ відповідно, то значенням функції $\tilde{u} = f(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n)$ є нечітке дійсне число $\tilde{u}$ з функцією належності

$$\mu_{\tilde{u}}(u^*) = \sup_{\substack{u^* = f(x_1^*, x_2^*, ..., x_n^*) \\ x_i^* \in \sup p(x_i), i = \overline{1,n}}} \min_{i=1,n} \mu_{\hat{x_i}}(x_i^*). \qquad (7.1)$$

Згідно з цим принципом можна знайти функцію належності нечіткого числа, що відповідає значенню чіткої функції від нечітких аргументів. Його реалізація здійснюється за таким *алгоритмом:*

Крок 1. Зафіксувати значення $u = u^*$.

Крок 2. Знайти всі набори $(x^*_{1j}, x^*_{2j}, ..., x^*_{nj})$, $j = \overline{1, k}$, що задовольняють умови

$$u^* = f(x^*_{1j}, x^*_{2j}, ..., x^*_{nj}) \quad i \quad x^*_{ij} \in \sup p(x_i^{\wedge}), \; i = \overline{1,n}.$$

Крок 3. Міру належності елемента u нечіткому числу u обчислити за формулою

$$\mu_{\tilde{u}}(u^*) = \max_{j=\overline{1,k}} \min_{j=\overline{1,n}} \mu_{x_i^{\wedge}}(x^*_{ij}). \qquad (7.2)$$

Крок 4. Перевірити умову «Чи вибрані всі елементи u*?».* Якщо так, то перейти на крок 5. Інакше – зафіксувати нове значення u і перейти на крок 2.

Крок 5. Закінчення алгоритму.

Означення 7.7. Лінгвістичною змінною називають п'ятірку $< I, T, \Omega, S, P >$, де I – ідентифікатор лінгвістичної змінної; T – терм-множина, яка є сукупністю найменувань нечітких змінних, кожна з яких визначена в Ω; S – синтаксична процедура, що дозволяє генерувати нові терми; P – семантична процедура, яка призначена для перетворення значень лінгвістичної змінної у нечіткі змінні.

Означення 7.8. Множиною рівня α, або α-перерізом нечіткої множини $A \subseteq \Omega$ називають чітку множину $A_\alpha = \{x \in \Omega \mid \mu_A(x) \geq \alpha\}$, $a \in [0;1]$.

Приклад 7.1. Нехай $\Omega = \{x_1, x_2, x_3, x_4\}$, A – нечітка множина, для якої $\mu_A(x_1) = 0{,}1$; $\mu_A(x_2) = 0{,}4$; $\mu_A(x_3) = 0$; $\mu_A(x_4) = 0{,}7$. Тоді A можна записати так:

$$A = \left\{ \frac{0{,}1}{x_1}; \frac{0{,}4}{x_2}; \frac{0}{x_3}; \frac{0{,}7}{x_4} \right\},$$

або $A = \{0{,}1 / x_1 + 0{,}4 / x_2 + 0 / x_3 + 0{,}7 / x_4\}$. Знак „+" тут змістовно означає об'єднання.

Приклад 7.2. Нехай $\Omega = \{0, 1, 2, ..., 240\}$. Тоді нечітка множина A = {людина високого зросту (в см)} може бути подана так:

$$A = \{0/0 + 0/1 + ... + 0{,}001/151 + ... + 0{,}7/180 + ... + 1/120\}.$$

Приклад 7.3. Нечіткі числа $\tilde{x}_1$ і $\tilde{x}_2$ задані такими трапецієподібними функціями належності (рисунки 7.2, 7.3):

$$\mu_{\tilde{x}_2}(x)=\begin{cases}0, \text{ якщо } x<4 \text{ або } x>7,\\ x-4, \text{ якщо } x\in[4;5],\\ 1, \text{ якщо } x\in(5,6),\\ -x+7, \text{ якщо } x\in[6;7].\end{cases}$$

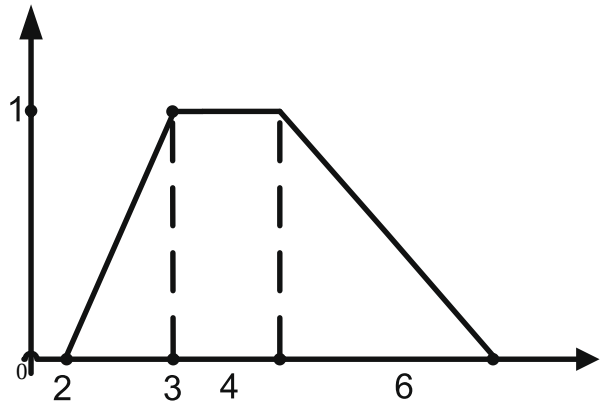


Рисунок 7.2 - Трапецієподібна функція належності для числа $\tilde{x}_1$

$$\mu_{\tilde{x}_2}(x)=\begin{cases}0, \text{ якщо } x<2 \text{ або } x>6,\\ x-2, \text{ якщо } x\in[2;3],\\ 1, \text{ якщо } x\in(3,4),\\ -\frac{1}{2}x+3, \text{ якщо } x\in[4;6].\end{cases}$$

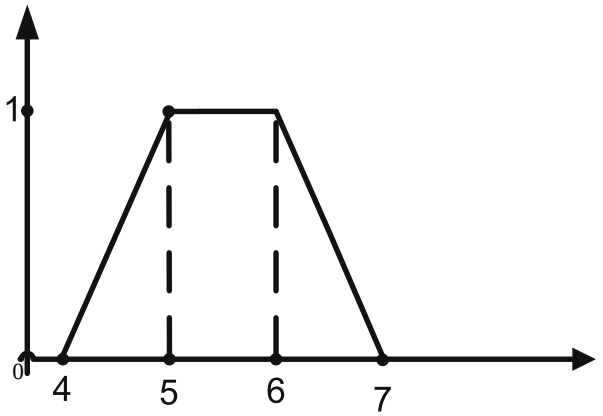


Рисунок 7.3 - Трапецієподібна функція належності для числа $\tilde{x}_2$

Знайдемо нечітке число $\tilde{y}=\tilde{x}_1\cdot\tilde{x}_2$, використавши принцип узагальнення. Задамо нечіткі аргументи на чотирьох точках $\{2,3,4,6\}$ для $\tilde{x}_1$ і $\{4,5,6,7\}$ для $\tilde{x}_2$. Тоді $\tilde{x}_1=\frac{0}{2}+\frac{1}{3}+\frac{1}{4}+\frac{0}{6}$ і $\tilde{x}_2=\frac{0}{4}+\frac{1}{5}+\frac{1}{6}+\frac{0}{7}$. Результати операцій множення зведемо в таблицю 7.1.

Таблиця 7.1 – Результати реалізації принципу узагальнення

$y^*=x_1^*\cdot x_2^*$	8	10	12		14	15	16	18	20	21	24		28	30	36	42
x_1^*	2	2	2	3	2	3	4	3	4	3	4	6	4	6	2	2
x_2^*	4	5	6	4	7	3		6	5	7	6	4	7	5	4	5
$\mu_{\tilde{x}_1}(x_1^*)$	0	0	0	1	0	1	1	1	1	1	1	0	1	0	0	0
$\mu_{\tilde{x}_2}(x_2^*)$	0	1	1	0	0	1	0	1	1	0	1	0	0	1	0	1
$\min(\mu_{\tilde{x}_1}(x_1^*),\mu_{x_2}(x_2^*))$	0	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0
$\mu_{\tilde{y}}(y^*)$	0	0		0	0	1	0	1	1	0		1	0	0	0	0

Результуюча нечітка множина задана в першому і останньому рядках таблиці. Перший рядок містить елементи універсальної множини, останній – міру їх належності до значення виразу $\tilde{x}_1\cdot\tilde{x}_2$. Одержимо такий результат: $\frac{0}{8}+\frac{1}{15}+\frac{1}{24}+\frac{0}{42}$.

Припустимо, що тип функції належності $\tilde{y}$ буде такий, як і аргументів

$\tilde{x}_1$ *і* $\tilde{x}_2$ – тобто трапецієподібний. У цьому випадку функцію належності задають виразом

$$\mu_{\tilde{x}_2}(x) = \begin{cases} 0, \text{ якщо } x < 8 \text{ або } y > 42, \\ \frac{1}{7}x - \frac{8}{7}, \text{ якщо } x \in [8; 15], \\ 1, \text{ якщо } x \in (15, 24), \\ -\frac{1}{18}x + \frac{7}{3}, \text{ якщо } x \in [24; 42]. \end{cases}$$

Результати розрахунків показано на рисунку 7.4.

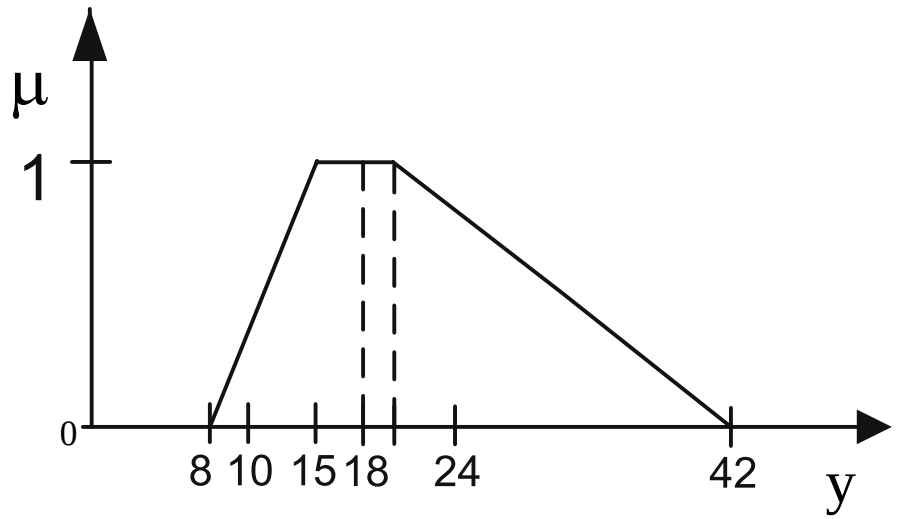


Рисунок 7.4 – Функція належності нечіткого добутку

***Приклад* 7.4.** (Означення лінгвістичної змінної). Нехай експерти класифікують автомобілі за значенням максимальної швидкості руху за допомогою понять „Мала швидкість", „Середня швидкість", „Велика швидкість". При цьому максимальне значення максимальної швидкості – 300 км/г. Формалізація такого опису здійснюється за допомогою лінгвістичної змінної $\lambda =< I, T, \Omega, S, P >$, де I – максимальна швидкість автомобіля; T – {„Мала швидкість", „Середня швидкість", „Велика швидкість"}; Ω – [140; 300]; S – процедура утворення нових термів за допомогою логічних зв'язок $\&, \neg, \vee$ і модифікаторів типу „дуже", „значно",..., наприклад, „не дуже велика швидкість"; P – процедура задання на Ω – [140; 300] нечітких підмножин $A_1 =$ "Мала швидкість", $A_2 =$ „Середня швидкість", $A_3 =$„Велика швидкість".

7.2 Нечіткі відношення і нечітке логічне виведення

Відомо, що булеве логічне виведення базується на таких тавтологіях:
- модус поненс: $(A \& (A \to B)) \to B;$

- модус толленс: $((A \to B) \& B) \to A;$

- силогізм: $((A \to B) \& (B \to C)) \to (A \to C);$

- контрапозиція: $(A \to B) \to (B \to A).$

***Означення* 7.9.** Нечітким логічним виведенням називається одержання висновку у вигляді нечіткої множини, що відповідає поточним значенням входів з використанням нечіткої бази знань і нечітких операцій.

Композиційне правило виведення Заде. Якщо відоме нечітке відношення $\tilde{R}$ між вхідною X і вихідною Y змінними, то при нечіткому значенні вхідної змінної $x = \tilde{A}$ нечітке значення вихідної змінної визначається так: $x = \tilde{A} \circ \tilde{R}$, де $\circ$ – знак максимінної композиції.

***Означення* 7.10.** Максимінною композицією нечітких відношень $\tilde{A}$ і $\tilde{B}$, заданих на $X \times Z$ *і* $Z \times Y$, називається відношення $\tilde{G} = \tilde{A} \circ \tilde{B}$ на множині $X \times Y$ з функцією належності

$$\mu_{\tilde{G}}(x,y) = \sup \min \left\{ \mu_{\tilde{A}}(x,z), \mu_{\tilde{B}}(x,z) \right\}, (x,y) \in X \times Y, (x,z) \in X \times Z, (z,y) \in Z \times Y.$$

У випадку скінченних множин X, Y, Z матрицю нечіткого відношення $\tilde{G} = \tilde{A} \circ \tilde{B}$ одержують як максимінний добуток матриць $\tilde{A}$ і $\tilde{B}$. Ця операція виконується як звичайний добуток матриць, при цьому операція поетапного добутку замінена на знаходження мінімуму, а знаходження суми – на знаходження максимуму.

***Приклад* 7.5.** Нехай задане нечітке правило: якщо $x = \tilde{A}$, то $y = \tilde{B}''$ з нечіткими множинами:

$$\tilde{A} = \frac{0,1}{1} + \frac{0,2}{2} + \frac{0,4}{4} + \frac{0,8}{8}, \quad \tilde{B} = \frac{0,3}{10} + \frac{0,5}{20} + \frac{0,8}{30}.$$

Визначити значення вихідної змінної y, якщо

$$x = \tilde{C} = \frac{0,2}{1} + \frac{0,6}{2} + \frac{0,7}{4} + \frac{0,3}{8}.$$

Розрахуємо нечітке відношення, що відповідає правилу: якщо $x = \tilde{A}$, то $y = \tilde{B}''$, застосовуючи операцію знаходження мінімуму:

$$\tilde{R} = \begin{bmatrix} 0,1 & 0,1 & 0,1 \\ 0,2 & 0,2 & 0,2 \\ 0,3 & 0,3 & 0,4 \\ 0,3 & 0,5 & 0,8 \end{bmatrix}.$$

Далі за формулою $y = \tilde{C} \circ \tilde{R}$ розрахуємо нечітке значення вихідної змінної:

$$y = 0,3 / 10 + 0,4 / 20 + 0,4 / 30.$$

Нечітке логічне виведення елементною базою має сукупність нечітких предикатних правил:

$$\theta_1 : \textit{Якщо } x \in A_1, \textit{ то } y \in B_1;$$

$$\theta_2 : \textit{Якщо } x \in A_2, \textit{ то } y \in B_2;$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\theta_n : \textit{Якщо } x \in A_n, \textit{ то } y \in B_n,$$

де x – вхідна змінна; y – змінна виведення; $A_i, B_i, \; i = \overline{1,n}$ – функції належності.

У загальному випадку логічне виведення здійснюється в чотири етапи. На першому етапі будують функції належності, які визначають міру істинності кожної передумови кожного правила. Далі здійснюється логічне виведення, яке полягає у тому, що, виходячи із значень істинності для передумов правила, обчислюють висновок кожного правила. На третьому етапі здійснюють композицію усіх нечітких підмножин, які відповідають кожній змінній виведення. На останньому етані здійснюють дефазифікацію нечіткого набору виведень у чітке число.

Відомо декілька алгоритмів логічного виведення, розглянемо їх детально.

Алгоритм Мамдані (Mamdani) [40].

Для спрощення запису алгоритму припустимо, що базу знань складають два нечітких правила вигляду

$$P_1 : \textit{ якщо } x \in A_1, i \; y \in B_1 \textit{ то } Z \in C_1;$$

$$P_2 : \textit{ якщо } x \in A_2, i \; y \in B_2 \textit{ то } Z \in C_2.$$

Крок 1. Знаходимо міри істинності $A_1(x), A_2(x_0), B_1(y_0), B_2(y_0)$ (рисунок 7.5).

Крок 2. Знаходимо рівні „відтинання" для передумов кожного з правил

$$\alpha_1 = A_1(x_0) \wedge B_1(y_0),$$

$$\alpha_2 = A_2(x_0) \wedge B_2(y_0).$$

Крок 3. Знаходимо функції належності

$$C_1(Z) = (\alpha_1 \wedge C_1(Z)),$$

$$C_2(Z) = (\alpha_2 \wedge C_2(Z)).$$

Крок 4. Виконуємо об'єднання знайдених функцій і знаходимо результуючу нечітку множину для вихідної змінної з функцією належності:

$$\mu_\theta(Z) = C(Z) = C_1(z) \vee C_2(Z) = (\alpha_1 \wedge C_1(Z)) \vee (\alpha_2 \wedge C\ (Z)).$$

Крок 5. Виконуємо дефазифікацію, наприклад за методом центра ваги, і знаходимо чітке значення

$$Z_0 = \frac{\int_{\underline{z}}^{\bar{z}} z\mu_\theta(z)dz}{\int_{\underline{z}}^{\bar{z}} \mu_\theta(z)dz},$$

де інтервал $\underline{z}, \bar{z}$ є носієм функції належності.

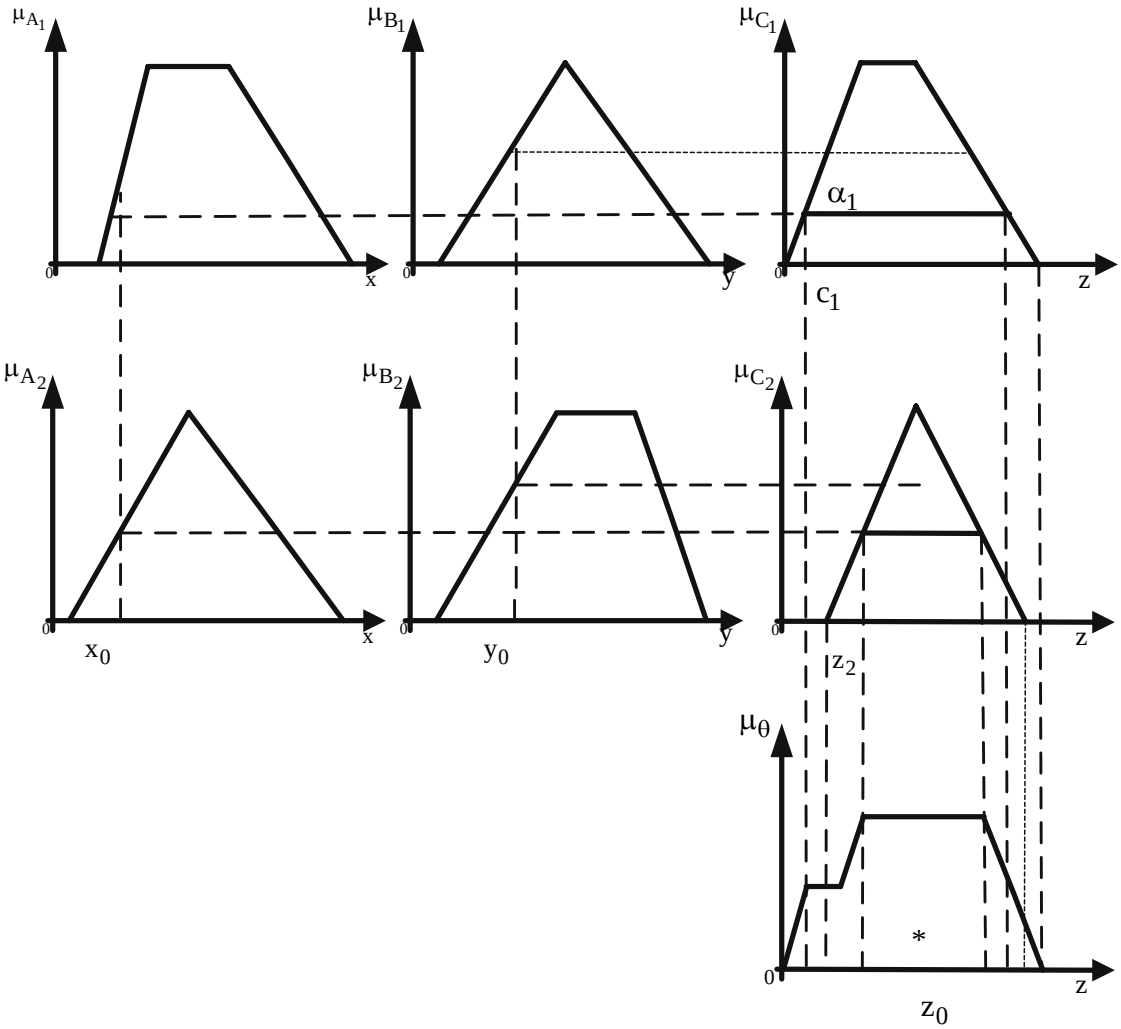


Рисунок 7.5 – Реалізація алгоритму Мамдані

Алгоритм Цукамото (Tsukamoto) [37].

Вихідні передумови аналогічні алгоритму Мамдані за умови, що функції $C_1(z)$ і $C_2(z)$ є монотонними.

Крок 1. Знаходимо міри істинності $A_1(x_0), A_2(x_0), B_1(y_0), B_2(y_0)$ (рисунок 7.6).

Крок 2. Знаходимо рівні "відтинання" α_1 і α_2 і через розв'язання рівнянь $\alpha_1 = C_1(z_1)$, $a_2 = C_2(z_2)$ – чіткі значення $(z_1$ і $z_2)$ для кожного із вихідних правил.

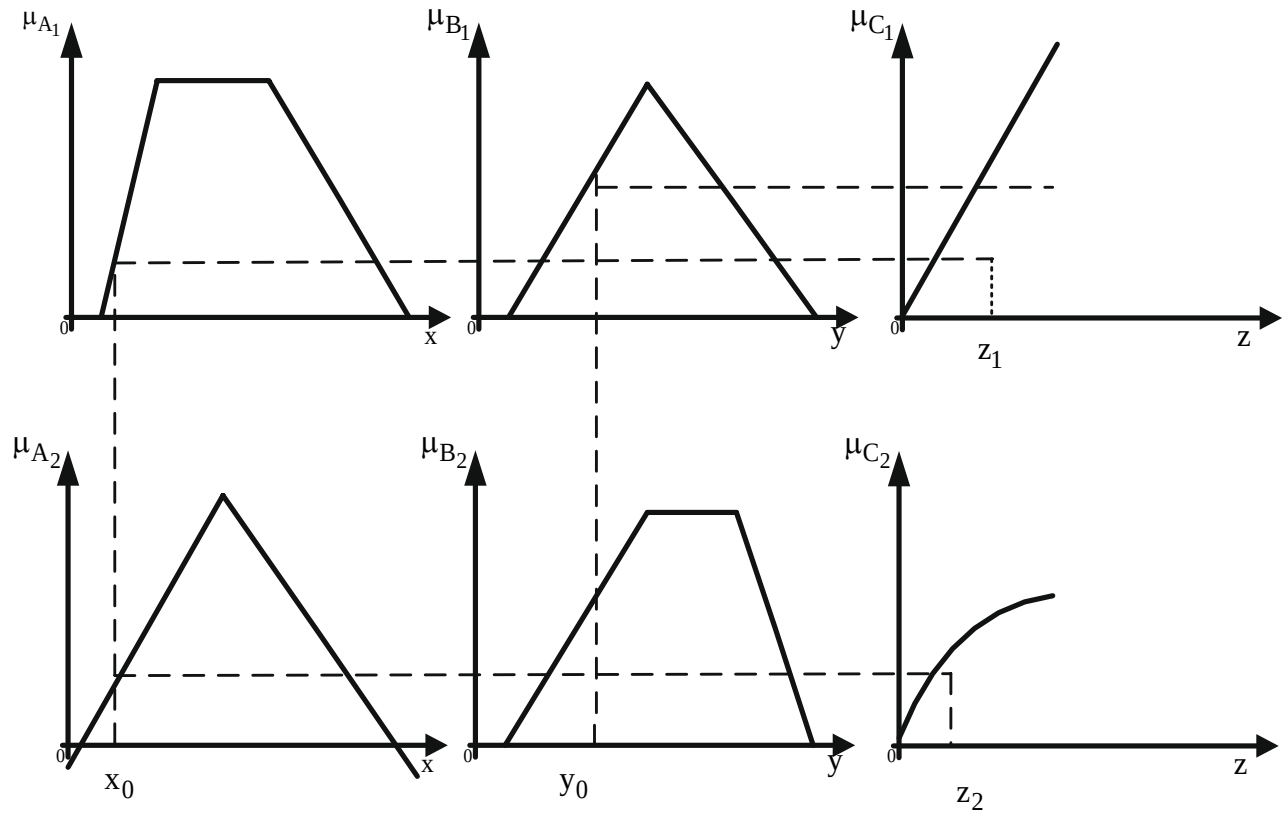


Рисунок 7.6 – Реалізація алгоритму Цукамото

***Крок* 3.** Визначаємо чітке значення вихідної змінної (як зважене середнє z_1 і z_2): $z_0 = \dfrac{\alpha_1 z_1 + \alpha_2 z_2}{\alpha_1 + \alpha_2}$ у загальному випадку

$$z_0 = \sum_{i=1}^{n} \alpha_i z_i \Big/ \sum_{i=1}^{n} \alpha_i .$$

Алгоритм Сугено і Такажі (Sugetw і Takagi) [41, 42].

Використовується набір правил у такій формі:

$$\prod_1 \ : якщо \ \ x \in A_1 \ \ і \ \ y \in B_i, \ то \ \ Z_1 = a_1 x + b_1 y,$$

$$\prod_2 \ : якщо \ \ x \in A_2 \ \ і \ \ y \in B_2, \ то \ \ Z_2 = a_2 x + b_2 y.$$

Крок 1. Знаходимо міри істинності $A_1(x_0), A_2(x_0), B_1(y_0), B_2(y_0)$ (рисунок 7.7).

Крок 2. Розраховуємо

$$\alpha_1 = A_1(x_0) \wedge B_1(y_0),$$

$$\alpha_2 = A_2(x_0) \wedge B_2(y_0),$$

а також

$$z_1^* = a_1 x_0 + \wedge b_1 y_0,$$

$$z_2^* = a_2 x_0 \wedge b_2 y_0.$$

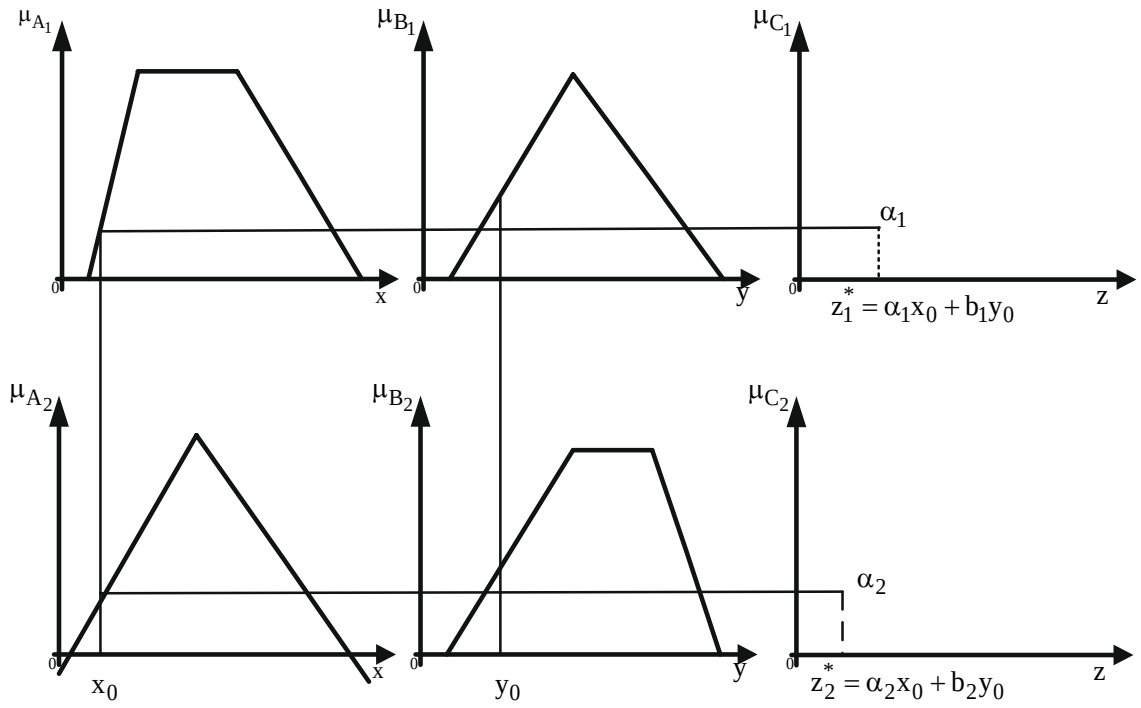


Рисунок 7.7 – Реалізація алгоритму Сугено і Такажі

Крок 3. Знаходимо чітке значення

$$z_0 = \frac{\alpha_1 z_1^* + \alpha_2 z_2^*}{\alpha_1 + \alpha_2}.$$

Алгоритм Ларсена (Larsen) [43].

***Крок* 1.** Знаходимо міри істинності $A_1(x_0), A_2(x_0), B_1(y_0), B_2(y_0)$ (рисунок 7.8).

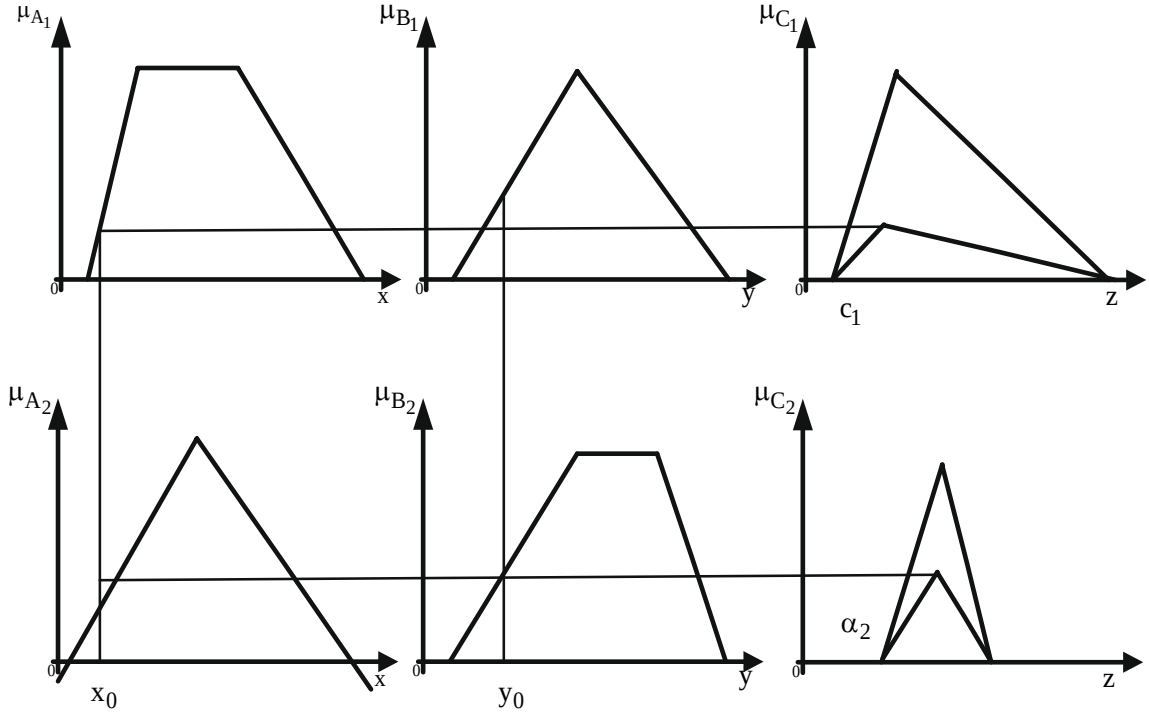


Рисунок 7.8 – Реалізація алгоритму Ларсена

Крок 2. Розраховуємо значення

$$\alpha_1 = A_1(x_0) \wedge B_1(y_0),$$
$$\alpha_2 = A_2(x_0) \wedge B_2(y_0),$$
$$\alpha_1 C_1(z) \text{ і } \alpha_2 C_2(z).$$

Крок 3. Знаходимо результуючу нечітку підмножину з функцією належності

$$\mu_{\sum}(z) = C(z) = (\alpha_1 C_1(z)) \vee (\alpha_2 C_2(z))$$

(у загальному випадку $\mu_{\sum}(z) = \bigvee_{i=1}^{n} (a_i C_i(z))$.

Крок 4. Знаходимо чітке значення.

Спрощений алгоритм.

Початкові правила є такими:

$\prod_1$: *якщо* $x \in A_1$ *і* $y \in B_i$, *то* $Z_1 = C_1$,

$\prod_2$: *якщо* $x \in A_2$ *і* $y \in B_2$, *то* $Z_2 = C_2$ *де* $C_i, i = \overline{1,2}$, *– чіткі числа*.

Крок 1. Знаходимо міри істинності $A_1(x_0), A_2(x_0), B_1(y_0), B_2(y_0)$ (рисунок 7.9).

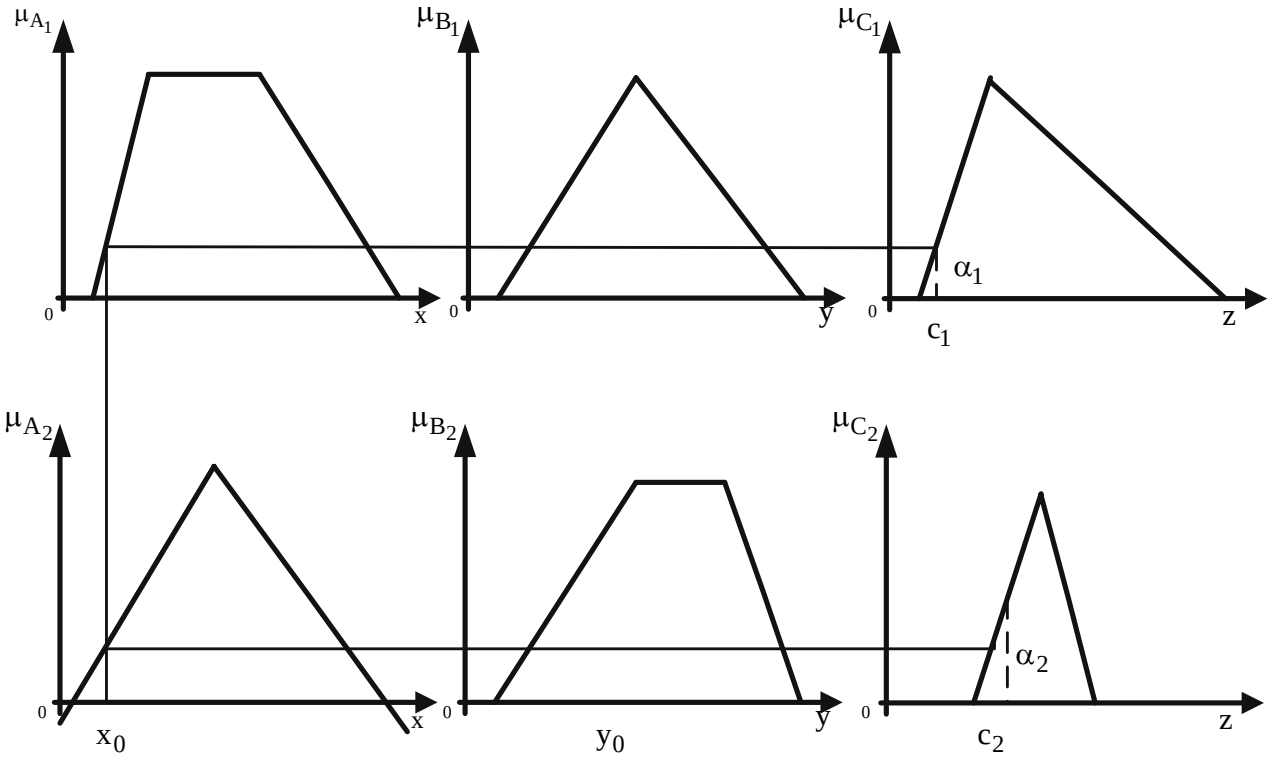


Рисунок 7.9 – Реалізація спрощеного алгоритму

Крок 2. Розраховуємо числа

$$\alpha_1 = A_1(x_0) \wedge B_1(y_0), \alpha_2 = A_2(x_0) \wedge B_2(y_0).$$

Крок 3. Знаходимо чіткі числа $Z_0 = \dfrac{\alpha_1 C_1 + \alpha_2 C_2}{\alpha_1 + \alpha_2}$.

7.3 Аналіз нечітких експертних заключень

Проілюструємо особливості аналізу нечітких експертних висновків на прикладі розрахунку можливого фінансування деякого проекту. Відомо, що у фінансуванні проекту беруть участь чотири організації, про які відомо таке:

Організація A – абсолютно надійна і стабільна, сума фінансування становить 100 од.

Організація B – стабільна, можливе фінансування проекту в сумі від 70 до 140 од., причому є більша впевненість у тому, що буде надано від 100 до 120 од.

Організація C – стабільна, але ненадійна, найбільша впевненість є в тому, що фінансування буде надано в сумі від 100 до 200 од., але можлива і повна його відсутність.

Організація $Д$ – ненадійна і нестабільна, напевне проект не профінансує, а якщо профінансує, то в розмірі 20 – 30 од., із зменшенням упевненості в наданні коштів із ростом суми.

Необхідно встановити найбільш можливу загальну суму фінансування, найменш імовірну і т.п.

Різні джерела фінансування подамо за допомогою нечітких величин, які будемо інтерпретувати як нечіткі інтервали, задані п'ятіркою елементів $(\underline{m}, \overline{m}, \alpha, \beta, h)$ (порядок слідування елементів важливий), де $\underline{m}$ – ліве модальне значення; $\overline{m}$ – праве модальне значення; α – лівий коефіцієнт скошеності; β – правий коефіцієнт скошеності; h – висота.

Нечітку величину зобразимо за допомогою функції належності (див. рисунок 7.1). Зазначимо, що нечіткою величиною $M_i \oplus M_j$, де M_i і M_j – два трапецієподібних нечітких інтервали, є також трапецієподібний інтервал $(\underline{m}, \overline{m}, \alpha, \beta, h)$,

де

$$h = \min(h_i, h_j),\ \alpha = h\left(\frac{\alpha_i}{h_i} + \frac{\alpha_j}{h_j}\right),\ \beta = h\left(\frac{\beta_i}{h_i} + \frac{\beta_j}{h_j}\right),$$

$$\underline{m} = \underline{m}_i + \underline{m}_j - \alpha_i - \alpha_j + \alpha,\ \overline{m} = \overline{m}_i + \overline{m}_j + \beta_i + \beta_j - \beta.$$

Нечіткі величини, що відповідають умові задачі, подамо так:

$$A = (100, 100, 0, 0, 1);\ B = (100, 120, 30, 20, 1);$$

$$C = C_1 \cup C_2 = (100, 200, 0, 0, 0{,}8) \cup (0, 0, 0, 0, 0{,}2);$$

$$D = D_1 \cup D_2 = (0, 0, 0, 0, 0{,}8) \cup (20, 20, 0, 10, 0{,}2).$$

Відповідні графіки функцій належності мають такий вигляд (рисунок 7.10):

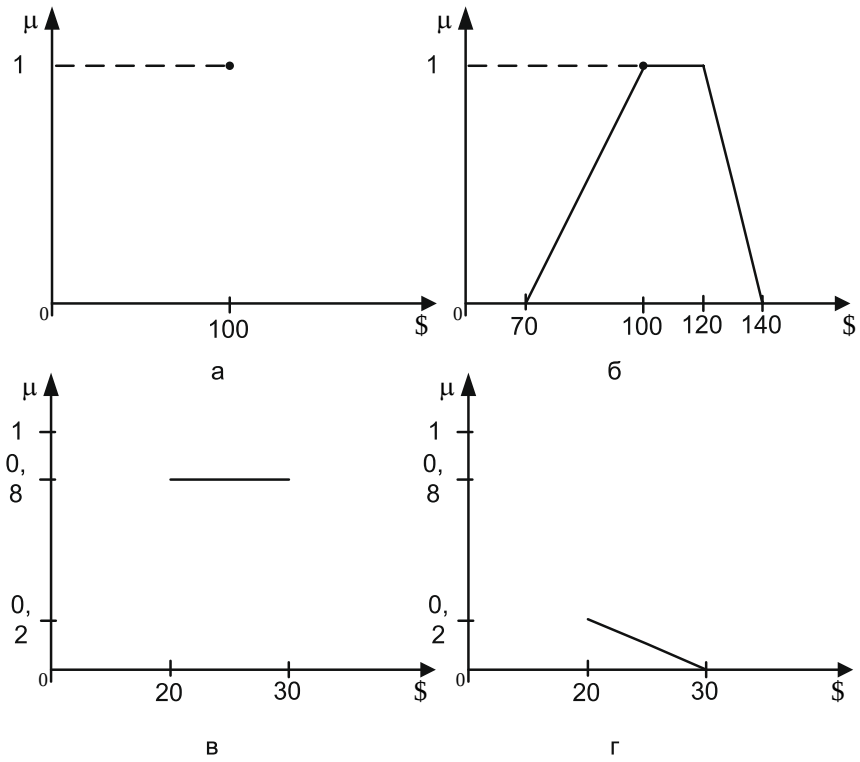


Рисунок 7.10 – Функції належності щодо можливості фінансування проекту

Можливі чотири варіанти фінансування:

$$S_1 = A \oplus B \oplus C_1 \oplus D_1,\ S_2 = A \oplus B \oplus C_1 \oplus D_2,$$

$$S_3 = A \oplus B \oplus C_2 \oplus D_1,\ S_4 = A \oplus B \oplus C_2 \oplus D_2.$$

Розрахувавши за наведеними вище формулами значення S_1, S_2, S_3, S_4, будуємо графік, причому внутрішні лінії витираємо. Наприклад, на рисунку 7.11 зображено графік суми

$$S = (294, 424, 24, 16, 0{,}8) \cup (296, 456, 614, 0, 2) \cup$$

$$\cup (176, 236, 6, 4, 0, 2) \cup (196, 256, 6, 14, 0, 2).$$

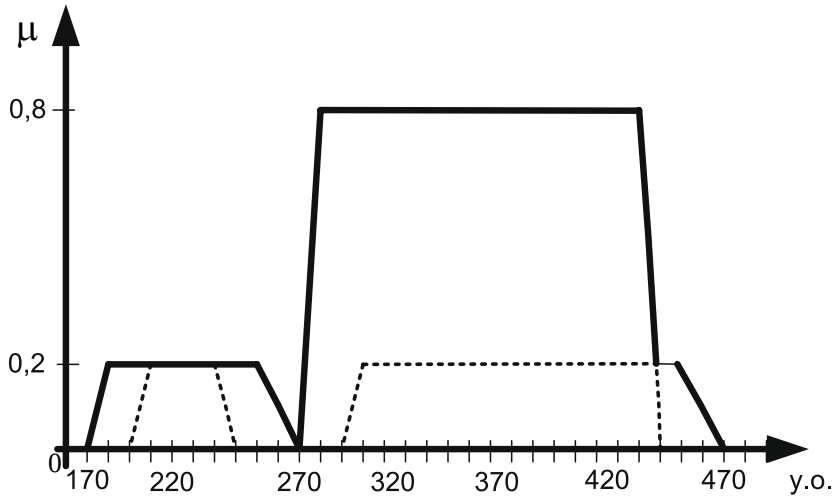


Рисунок 7.11 – Результуючий графік

Відповідно до одержаного результату область найбільш можливого фінансування знаходиться в діапазоні 294 – 424 од.; перевищення суми в 424 од. можливе, але впевненість у цьому зменшується із ростом суми; невелика впевненість існує у тому, що надходження не становитимуть більше 176 – 256 од. (рівень 0,2); у будь-якому випадку вони не можуть опуститись нижче 170 од. і піднятись вище 470 од.

7.4 Мурашині алгоритми

Розглянемо метаевристику, яка називається оптимізацією мурашиних колоній (Ant Colony Optimization (ACO)). Вона і була запропонована проф. Марко Доріго (Marco Dorieo, 44) і його колегами на початку 90-х років минулого століття як метод розв'язання складних комбінаторних оптимізаційних задач.

Алгоритм ACO належить до предметної області, яку називають інтелектом рою (swarm intelligence) і в якій вивчаються алгоритми, розроблені внаслідок спостережень над поведінкою роїв комах, зграй птахів, гуртів тварин тощо. Визначальною характеристикою такої поведінки, що певним чином реалізується в алгоритмах, є кооперація індивідів через самоорганізацію без будь-якого централізованого управління ззовні.

ACO-метаевристика виникла у результаті спостережень за колоніями мурах. Одним з перших вчених, який відзначив певну соціальну поведінку мурах, був французький ентомолог П'єр-Пауль Грассе (Pierre-Paul Grasse). У 40-х роках минулого століття він спостерігав поведінку термітів, які були здатні реагувати на „важливі стимули" і потім передавати цю інформацію іншим комахам.

Грассе назвав це явище стігмерджентність (stigmergy) і визначив, що від інших видів комунікацій його відрізняють такі властивості:

- фізична, несимвольна сутність інформації, яка реалізується через комунікацію комах і відповідає змінам стану навколишнього середовища;
- локальна природа такої інформації, оскільки вона досяжна лише для комах, які перебувають у безпосередній близькості до джерела інформації.

Приклади стігмерджентності спостерігаються у колоніях мурах. Більшість їх типів, подорожуючи до й від джерел їжі, викладають субстанцію, яку називають феромоном (phero-mone). Інші мурахи здатні відчувати феромон і його присутність визначає вибір шляху, причому того, де концентрація феромону є найбільшою. Проведені експерименти, коли джерело їжі та групу мурах розділяли два мости, переконують у цьому висновку. Якщо довжини мостів були однаковими, то по них рухалась приблизно однакова кількість мурах, якщо ж довжина одного мосту була меншою, то через деякий час щільність мурах на ньому ставала більшою. Розглянемо модель, яка пояснює цей експеримент.

Припустимо, що після моменту часу t з початку експерименту m_1 мурах використовують для руху перший міст і m_2 – другий міст. Відповідно $m = m_1 + m_2$. Тоді ймовірність p_1 для $(m+1)$-ї мурахи вибрати перший міст обчислюється так:

$$p_{1_{(m+1)}} = \frac{(m_1 + k)^h}{(m_1 + k)^h + (m_2 + k)^h}, \tag{7.3}$$

де параметри k і h визначаються з експериментальних даних. Ймовірність, що та ж $(m+1)$-та мураха обере другий міст, є такою:

$$p_{2_{(m+1)}} = 1 + p_{1_{(m+1)}}. \tag{7.4}$$

Експериментальні дослідження засвідчили, що модель (7.3) – (7.4) при використанні реальних даних була адекватною при $k \approx 20$ і $h \approx 2$.

Ця базова модель, яка пояснює поведінку мурах, може бути застосовною для розв'язання оптимізаційних задач. Використаємо аналогії з природними мурахами, тоді головні характеристики стігмерджентності можуть бути розширеними на випадок штучних агентів при:

- асоціації змінних стану зі станами різних задач;
- агентам буде дозволено лише локальний доступ до цих змінних.

Іншим важливим аспектом, який може бути використаний штучними мурахами, є поєднання автокаталітичного механізму та оцінки розв'язків. При (неявній) оцінці розв'язків помічаємо факт, що найкоротші шляхи (які відповідають найменшій ціні розв'язків у випадку штучних мурах) закінчуються раніше, ніж довші, і тому на них накопичується більше феромону. Неявна оцінка розв'язків у поєднанні з автокаталізом може бути справді ефективною: коротший шлях – швидше відкладається феромон – більше мурах використовують коротший шлях. При відповідному застосуванні це є потужний механізм в оптимізаційних алгоритмах, що базуються на використанні поведінки популяцій (наприклад, в еволюційних алгоритмах автокаталіз використовується при селекційно-репродукційних механізмах).

Стігмерджентність разом із непрямою оцінкою розв'язків та автокаталітичною поведінкою дозволяють сформувати алгоритм функціонування мурашиної колонії (АФМ). Базова ідея АФМ є близькою до біологічної ідеї. І природні, і штучні колонії мурах є популяціями індивідів, які працюють разом для досягнення кінцевої мети. Колонія є популяцією простих, незалежних, асинхронних агентів, які кооперуються для знаходження кращих розв'язків задачі. У випадку природних мурах – це задача знаходження їжі, для штучних мурах – знаходження кращих розв'язків оптимізаційних задач. Проста мураха (і природна, і штучна) здатна знайти розв'язок такої задачі, але лише кооперація між багатьма індивідами через стігмерджентність дозволяє знаходити кращі розв'язки. Штучні мурахи живуть у віртуальному світі, тому вони лише модифікують

числові значення (штучний феромон), що асоційовані з різними станами задачі. Послідовність значень феромону, що асоційована з станами задачі, називається слідом штучного феромону. В АФМ він є єдиним комунікаційним фактором для мурах. Механізм, аналогічний випаровуванню фізичного феромону в природних колоніях мурах, дозволяє штучним мурахам забувати історію і фокусуватись на нових перспективних напрямках пошуку.

Подібно до природних штучні мурахи створюють розв'язки послідовно, рухаючись від одного стану задачі до іншого. Є декілька відмінностей між реальними та штучними мурахами:

- штучні мурахи живуть у дискретному світі – вони рухаються послідовно через скінченну множину станів задачі;

- зміна концентрації феромону для природних і штучних мурах здійснюється неоднаково. Інколи зміна концентрації феромону здійснюється лише деякими штучними мурахами і часто лише після одержання розв'язку;

- передбачається використання механізмів, яких немає у природі.

АФМ був формалізований як метаевристика комбінаторної оптимізації М. Доріго і відтоді активно використовується для розв'язання задач комбінаторної оптимізації (ЗКО). Перший крок застосування АФМ для розв'язання ЗКО полягає у визначенні адекватної моделі. Вона використовуватиметься для визначення центральної компоненти АФМ, тобто моделі концентрації феромону.

Моделлю ЗКО називається модель $P = (S, \Omega, f)$, яка складається:

– з простору пошуку S, який визначається як скінченна множина змінних дискретних розв'язків, і множини обмежень Ω;

– цільової функції $f : S \to R_0^+$, яку необхідно мінімізувати.

Простір пошуку S визначається так: дана множина дискретних змінних $X_i, i = \overline{1,n}$ зі значеннями $\vartheta_i^j \in D_i = \left\{ \upsilon_i^1, ..., \upsilon_i^{|D_i|} \right\}$. Те, що змінна X_i має значення ϑ_i^j, позначимо $X_i = \nu_i^j$. Розв'язок $s \in S$, у якому кожна складова змінна має значення, яке задовольняє усі обмеження множини Ω, є допустимим розв'язком ЗКО. Якщо множина Ω є порожньою, P називається моделлю задачі без обмежень, інакше – з обмеженнями. Розв'язок $s^* \in S$ називається глобальним оптимумом тоді і тільки тоді, коли $f(s^* \le f(s)) \ \forall s \in S$. Множину всіх глобальних оптимальних розв'язків позначимо $S^* \subseteq S$. Розв'язання ЗКО потребує знаходження хоча б одного $s^* \in S^*$.

Модель ЗКО використовується для одержання моделі феромону з використанням АФМ. Спочатку ініціалізована змінна розв'язку $X_i = \nu_i^j$

називається компонентом розв'язку і позначається C_{ij}. Множина всіх можливих компонентів розв'язку – C, множина всіх параметрів сліду феромону – T. Значення параметра сліду феромону T_{ij} позначимо τ_{ij} і назвемо значенням феромону. Це значення використовується і модифікується під час пошуку алгоритмом АФМ, що дозволяє моделювати ймовірнісний розподіл різних компонентів розв'язку.

В АФМ штучні мурахи будують розв'язок ЗКО, подорожуючи конструктивним графом $G_c(V, E)$. Повнозв'язний граф складається з множини вершин V і множини ребер E.

Множина компонент С може бути асоційованою з множиною вершин або з множиною ребер. Мурахи рухаються від вершини до вершини вздовж ребер графа, будуючи часткові розв'язки. Вони також залишають певну кількість феромону на компонентах: або у вершинах, або на ребрах, які вони проходять. Кількість феромону $\Delta\tau$ залежить від якості знайдених розв'язків. Наступні мурахи використовують інформацію про феромон як вказівку на більш перспективні області простору пошуку.

Метаевристика АФМ містить ініціалізацію і цикл з трьох алгоритмічних компонент. Проста ітерація циклу складається з побудованих усіма мурахами розв'язків, їх покращання з використанням локального алгоритму пошуку і модифікації значення феромону. Пояснимо ці три алгоритмічні компоненти більш детально.

АФМ-*метаевристика:* Ініціалізація параметрів, сліду феромону.

Доки не виконана умова зупинки, необхідно:

– сконструювати розв'язки (СР);

– застосувати локальний пошук (ЗЛП);

– модифікувати концентрацію феромону (МКФ);

– кінець циклу.

Розглянемо елементи метаевристики.

Конструювання розв'язків. Множина з m штучних мурах конструює розв'язки з елементів скінченної множини допустимих компонентів розв'язку $C = \{c_{ij}\}, i = \overline{1,n},\ j = \overline{1,|D_i|}$. Конструювання розв'язку починається з порожнього часткового розв'язку $s^p = \otimes$. Далі на кожному кроці частковий розв'язок s^p розширюється шляхом додавання випадкового компоненту розв'язку з множини ймовірних сусідів $N(S^p) \subseteq C$. Процес конструювання розв'язків може розглядатись як шлях на конструктивному графі $G_C = (V, E)$. Такий шлях у G_C однозначно визначається механізмом конструювання розв'язку, за допомогою якого формується множина $N(S^p)$ по відношенню до частинного розв'язку S^p.

Вибір компоненти розв'язку з $N(S^p)$ виконується випадково на кожному кроці конструювання. Точні правила для випадкового вибору

компонент розв'язку є різними у різних варіантах АФМ. Одне з найбільш відомих правил таке:

$$p\left(c_{ij} \middle/ s^p\right) = \frac{\tau_{ij}^{\alpha} \cdot \eta(c_{ij})^{\beta}}{\sum_{c_{ij} \in N(S^p)} \tau_{ij}^{\alpha} \cdot \eta(c_{ij})^{\beta'}}, \quad (7.5)$$

де τ_{ij} – значення концентрації феромону, що відповідає комопоненті c_{ij}; $\eta(*)$ – функція, за допомогою якої призначається на кожному кроці конструювання евристичне значення кожному випадковому компоненту розв'язку $c_{ij} \in N(S^p)$. Значення, яке повертає ця функція у загальному випадку називають евристичною інформацію, α і β – додатні параметри, значення яких визначаються відносною важливістю концентрації феромону та евристичної інформації. Рівняння (7.5) є узагальненням (7.3) і свідчить про подібність АФМ до біологічних механізмів.

Застосування локального пошуку. Якщо деякі розв'язки вже одержані, то перед зміною концентрації феромону необхідно виконати певні додаткові дії. Їх часто називають демон-акціями і вони є спеціальними процедурами, оскільки не виконуються простими мурахами. Найчастіше демон-акція полягає у застосуванні локального пошуку для побудови розв'язків, тобто у формуванні множини локально оптимізованих розв'язків для визначення того, що потрібно змінити концентрацію феромону.

Модифікація концентрації феромону. Метою зміни концентрації феромону є збільшення значення феромону, що асоціюється з перспективними розв'язками, і зменшення – інакше. Зазвичай це досягається через зменшення усіх значень концентрації феромону внаслідок його випаровування і збільшення значень концентрації феромону, що відповідають множині перспективних розв'язків S_{upd}:

$$\tau_{ij} \leftarrow (1-\rho)\tau_{ij} + \rho \sum_{s \in S_{upd} | c_{ij} \in s} F(s), \quad (7.6)$$

де S_{upd} – множина розв'язків, які модифікуються; $\rho \in (0;1]$ – коефіцієнт випаровування; $F : S \to R_0^+$ – функція, для якої $f(s) < f(s') \Rightarrow F(s) \geq F(s'), \forall s \neq s \in S$. $F(*)$ називається фітнес-функцією.

Випаровування феромону необхідне для уникнення надто швидкої збіжності алгоритму. Така операція є певною формою „забування" і сприяє появі нових областей у просторі пошуку. Інші АФМ алгоритми відрізняються зміною концентрації феромону.

Модифікації правила (7.6) одержують, виконуючи різні специфікації S_{upd}, які в багатьох випадках є підмножиною $S_{iter} \cup \{s_{bc}\}$, де S_{iter} – множина розв'язків, яка одержана на поточній ітерації; S_{bs} – найкращий

розв'язок, одержаний, починаючи з першої ітерації. Добре відомим прикладом є AS (Ant System)-правило модифікації, де

$$S_{upd} \leftarrow S_{iter}. \tag{7.7}$$

Досить часто на практиці використовують IB-правило (iteration best)

$$S_{upd} \leftarrow arg \max_{s \in S_{iter}} F(s). \tag{7.8}$$

Правилом IB вводиться більш строге зміщення для знаходження перспективного розв'язку, ніж AS-правилом. Хоча це збільшує швидкість його пошуку, але одночасно збільшується і ймовірність передчасної збіжності. Сильніше зміщення вводиться BS-правилом, де використовується найкраще рішення з першої ітерації S_{bs}. У цьому випадку S_{upd} є множиною $\{S_{bs}\}$. На практиці АФМ алгоритм, який використовує IB- або BS-правила модифікації і додатково включає механізм запобігання передчасній збіжності, має кращі результати, ніж при використанні AS-правила.

Головні модифікації АФМ. У науковій літературі запропоновано декілька варіантів АФМ. Розглянемо три з них: AS (Ant System) – першу реалізацію АФМ алгоритму, MM AS (Max-Min Ant System) і ACS (Ant Colony System) разом з їх короткими додатками. Для ілюстрації відмінностей між ними використаємо приклад задачі комівояжера.

AS – перший АФМ *-алгоритм.* Його головною особливістю є те, що концентрація феромону змінюється після того, як усі мурахи здійснять повний тур. Для ребра, що з'єднує вершини i і j, концентрація феромону змінюється за законом

$$\tau_{ij} \leftarrow (1-\rho)\tau_{ij} + \sum_{k=1}^{m} \Delta\tau_{ij}^{k}, \tag{7.9}$$

де ρ – коефіцієнт випаровування; m – кількість мурах; $\Delta\tau_{ij}^{k}$ – кількість феромону на одиницю довжини, відкладеного на ребрі (i, j) *к*-ю мурахою:

$$\Delta\tau_{ij}^{k} = \begin{cases} 0, \text{ в іншому випадку} \\ \dfrac{Q}{L_k}, \text{ якщо } k-\text{та мураха використовує ребро } (i,j) \text{ у своєму маршруті;} \end{cases}$$

Q – константа; L_k – довжина маршруту **к**-ї мурахи.

Будуючи розв'язки в AS-алгоритмі, мурахи рухаються по конструктивному графу і з певною ймовірністю приймають рішення у кожній вершині. Ймовірність переходу p_{ij}^{k} k-ї мурахи з точки i до точки j є такою:

$$p_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha} \cdot \eta_{ij}^{\beta}}{\sum_{l \in A_k} \tau_{ij}^{\alpha} \cdot \eta_{ij}^{\beta}}, \text{ якщо } j \in A_k, \\ 0, \text{ інакше }, \end{cases} \tag{7.10}$$

де A_k – список міст, які k-та мураха ще не відвідала; α і β – параметри, які визначають відносну важливість концентрації феромону та евристичної інформації, заданої виразом

$$\eta_{ij} = \frac{1}{d_{ij}}, \tag{7.11}$$

де d_{ij} – довжина шляху (i, j).

Деякі реалізації AS-алгоритму застосовувались для різних задач ЗКО. Найбільш відомим є його застосування для розв'язання ЗК (задачі комівояжера), квадратичної задачі про призначення, задач теорії розкладів, транспортної задачі та інших.

7.5 Основи генетичного програмування

Генетичне програмування, запропоноване Л. Крамером (L. Cramer) у 1985 році і далі розвинене Дж. Козою (J. Koza) [45], вирішило проблему фіксованої довжини розв'язків, створивши нелінійні сутності з різними розмірами і формами. Алфавіт, використаний для створення цих сутностей, був також різноманітним, утворюючи багатогранну систему подання. Проте створеним індивідам не вистачало простого автономного генома, який функціонував би одночасно як геном і феном. ГП є простими реплікаторами і адаптуються завдяки їхнім власним властивостям. Нелінійні сутності (дерева аналізу) ГП нагадують використання молекул білка в своєму алфавіті та їх складне ієрархічне подання. Тому ГП-сутності здатні до демонстрації широкої функціональності. Але ці сутності дуже складні для відтворення з модифікацією, оскільки генетичні модифікації виконуються безпосередньо на дереві аналізу. Як наслідок більшість модифікацій генерують структурні невідповідності. Для порівняння варто зауважити, що в природі вираз будь-якого гена білка завжди має місце в дійсній структурі білка (у природі немає такої сутності, як структурно некоректний білок).

Отже, у ГП генетичні оператори діють безпосередньо на дереві аналізу, що з першого погляду здається вигідним, але дуже обмежує цю технологію (неможливо змусити апельсинове дерево продукувати плоди манго, тільки прищепивши і обрізавши гілки). До того ж підґрунтя застосування генетичних операторів, доступних у ГП, дуже обмежене, оскільки більшість з них приводила б до некоректних дерев аналізу і тому в ГП використовується майже ексклюзивна спеціальна рекомбінація, яка діє на рівні дерев аналізу. У цьому ГП-специфічному кросовері відібрані гілки обмінюються між батьківськими деревами аналізу для створення нової популяції. Ідея полягає в обміні менших, математично стислих блоків для того, щоб розвивати складніші ієрархічні розв'язки, складені з менших блоків.

Оператор мутації у ГП також відрізняється від точкових природних мутацій для того, щоб гарантувати створення синтаксично правильних програм. Оператор мутації вибирає вузол у дереві аналізу і замінює гілку внизу цього вузла випадково створеною гілкою. Повна форма дерева незначно змінюється у результаті такої мутації.

Перестановка – третій оператор, який використовується у ГП, і подібно до рекомбінації і мутації є значно обмеженим: вибирають два структурно еквівалентних вузли (два термінали або дві функції з однаковим числом аргументів) і їх позиції обмінюються. У цьому разі повна форма дерева залишається незмінною.

Незважаючи на те що Дж. Коза описав ці три оператори як основні оператори ГП, кросовер – практично єдиний генетичний оператор, який використовується у більшості релізацій ГП. Не дивно, що в ГП великі популяції дерев аналізу використовуються з метою створення усіх складових блоків з перевіркою початкової популяції для того, щоб ґарантувати знаходження розв'язку, тільки переміщуючи ці початкові блоки.

Нарешті, завдяки дуальній функції дерев аналізу (геномів і феномів) у ГА і ГП неможливий простий, рудиментарний вираз: в усіх випадках повне дерево аналізу є розв'язком.

Програмування генетичних виразів (ПГВ). Програмування генетичних виразів є природним розвитком ГА і ГП.

ПГВ використовує такий же вигляд діаграми подання, як у ГП, але сутності, створені ПГВ (дерева виразів), є виразами геномів. Тому з ПГВ другий еволюційний поріг – поріг фенотипу – був перетнутий, забезпечуючи нові і ефективні рішення для еволюційних обчислень.

Отже, особливістю ПГВ є використання хромосом, здатних до подання будь-якого дерева виразів. Для цього була створена нова мова (Karva), щоб читати і „добувати" інформацію з ПГВ хромосом. До того ж структура хромосом проектувалася для того, щоб дозволити створення множини генів, кожен з яких кодує піддерево виразу. Гени структурно організовані на початку і в кінці, і це є та структурна і функціональна організація генів ПГБ, яка завжди ґарантує створення валідних програм, незважаючи на те, скільки або як глибоко модифікуються хромосоми.

Головних гравців у ПГВ лише два: хромосоми і дерева виразів (ДВ), останні – вираз генетичної інформації, що кодується в хромосомах. Як і в природі, процес декодування інформації називається трансляцією, і ця трансляція очевидно передбачає свого роду код і набір правил. Генетичний код дуже простий: він є взаємовідношенням між символами хромосоми і функцій або терміналів. Правила також прості: вони визначають просторову організацію функцій і терміналів у ДБ і типи взаємодії між під-ДВ.

У ПГВ є дві мови: мова генів і мова ДВ. Знаючи послідовність або структуру однієї, знаємо це і іншої. У природі, незважаючи на можливість

зробити висновок про послідовність білків, поданої послідовністю генів і навпаки, ми практично нічого не знаємо про правила, які визначають тривимірну структуру білків. Але в ПГВ завдяки простим правилам, які визначають структуру ДВ і їх взаємодії, можливо зробити висновок про фенотип, поданий послідовністю генів, і навпаки. Ця двомовна і недвозначна система називається Karva.

Геном. У ПГВ геном або хромосома складається з лінійного символьного рядка фіксованої довжини, складеного з одного або більше генів. Незважаючи на їхню фіксовану довжину, ми переконаємось у тому, що у ПГБ хромосоми кодують ДВ різних розмірів і форм.

Відкриті для читання фрейми і гени. Структурну організацію генів ПГВ краще зрозуміти в термінах фреймів, відкритих для читання (ФБЧ). У біології ФБЧ або кодована послідовність генів починається зі „стартового" кодону, продовжується кодонами амінокислоти і закінчується заключним кодоном. Проте ген є більшим за відповідний ФБЧ, з послідовностями вгору від стартового кодону і послідовностями вниз від кодону зупинки. Незважаючи на те, що в ПГВ стартова сторінка, вона завжди – перша позиція гена, точка зупинки не завжди збігаються з останньою позицією гена. Зазвичай гени ПГВ мають незакодовані регіони вниз від точки зупинки. Поки що ми не розглядатимемо ці незакодовані регіони, оскільки це не перешкоджає створенню виразів.

Розглянемо, наприклад, алгебраїчний вираз

$$\frac{ab}{c} + \sqrt{d - e}\,. \tag{7.12}$$

Він може бути також поданий діаграмою (рисунок 7.12), де Q є функцією квадратного кореня.

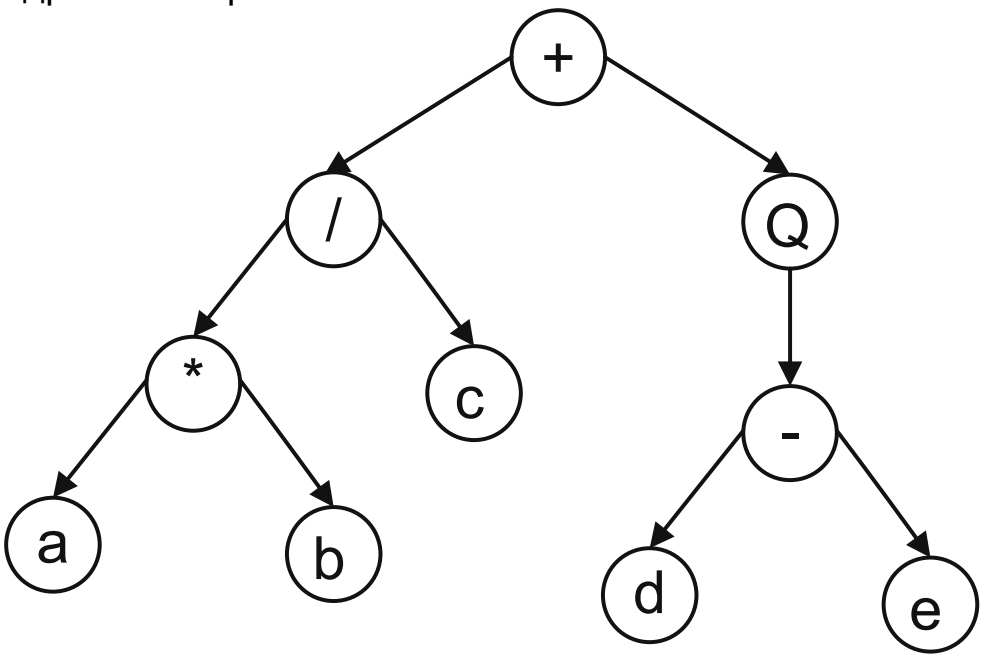


Рисунок 7.12 – Організація генів алгебраїчного виразу $\frac{ab}{c} + \sqrt{d - e}$

Цей вигляд діаграмних подань – фактично фенотип хромосом ПГВ, з якого легко записати генотип, як показано нижче:

$$\begin{array}{cccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ + & / & Q & * & c & - & a & b & d & e. \end{array} \quad (7.13)$$

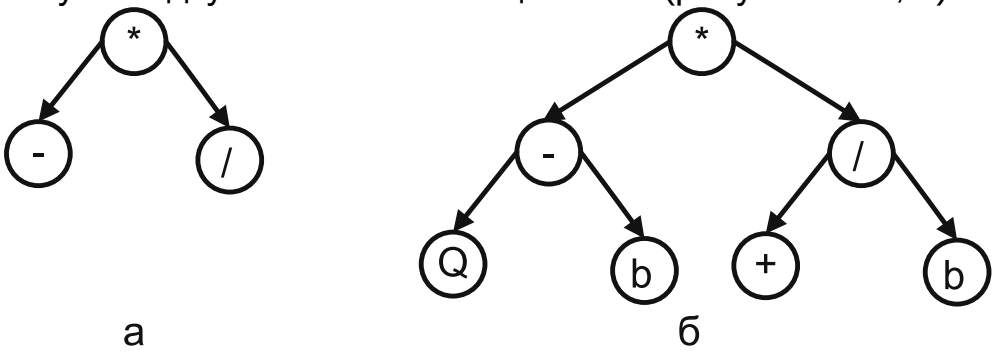
Його одержують внаслідок читання ДВ зліва направо і знизу догори (точно так, як ми читаємо сторінку тексту). Вираз (7.13) є ФБЧ, який починається з „+" (позиція 0) і закінчується в „є" (позиція 9). К. Ферейра назвала ці ФВЧ **К**-виразами (виходячи із Karva-нотацій).

Розглянемо інший ФВЧ і наступний **К**-вираз:

$$\begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 & 1 \\ + & - & / & Q & b & + & b & + & a & a & a & b. \end{array} \quad (7.14)$$

Він також простий. Щоб правильно подати ФВЧ, ми маємо дотримуватися правил управління просторовим розподілом функцій і терміналів. По-перше, початок гена відповідає кореню ДВ, формуємо цей вузол у першій лінії. По-друге, залежно від кількості аргументів кожного елемента (функції, можливо, мають різну кількість аргументів, тоді як термінали мають нуль операндів) у наступній лінії розміщені багато вузлів, які є аргументами функцій попередньої лінії. По-третє, зліва направо заповнюємо вузли у тому ж порядку елементами генів. По-четверте, процес повторюється доти, доки лінія, що містить тільки термінали, не буде сформована. Оскільки для **К**-виразу (9.19) корінь ДВ – символ в позиції 0, то отримаємо

Функція добутку має два аргументи, тому наступна лінія матиме два вузли, у даному випадку символи в позиціях 1 і 2 (рисунок 7.12, а).



а б

Рисунок 7.13 – Структура фреймів, які відкриті для читання

Віднімання і ділення – функції двох аргументів, тому в наступній лінії розміщені ще чотири вузли. У цьому разі це символи у позиціях 3, 4, 5 і 6 (рисунок 7.13, б).

Таким чином, ми маємо дві різні функції у третій лінії: одна – функція одного аргументу (Q), інша – функція двох аргументів (+). Тому ще три вузли потрібно побудувати в наступній лінії. У даному випадку вони

заповнені елементами в позиціях 7, 8 і 9 (рисунок 7.14, а).

У цій новій лінії, не зважаючи на те, що є три вузли, тільки один з них – функція (+). Тому відповідні вузли розміщені нижче цієї функції і заповнені наступними елементами в ФВЧ (позиції 10 і 11). Отримуємо рисунок 7.14, б.

У цьому випадку цим кроком ДВ було цілком сформовано, оскільки остання лінія містить тільки вузли з терміналами. Ми бачимо, що завдяки структурній організації генів ПГВ остання лінія усього ДВ містить виключно термінали. Це дає підстави стверджувати, що ДВ синтаксично правильні.

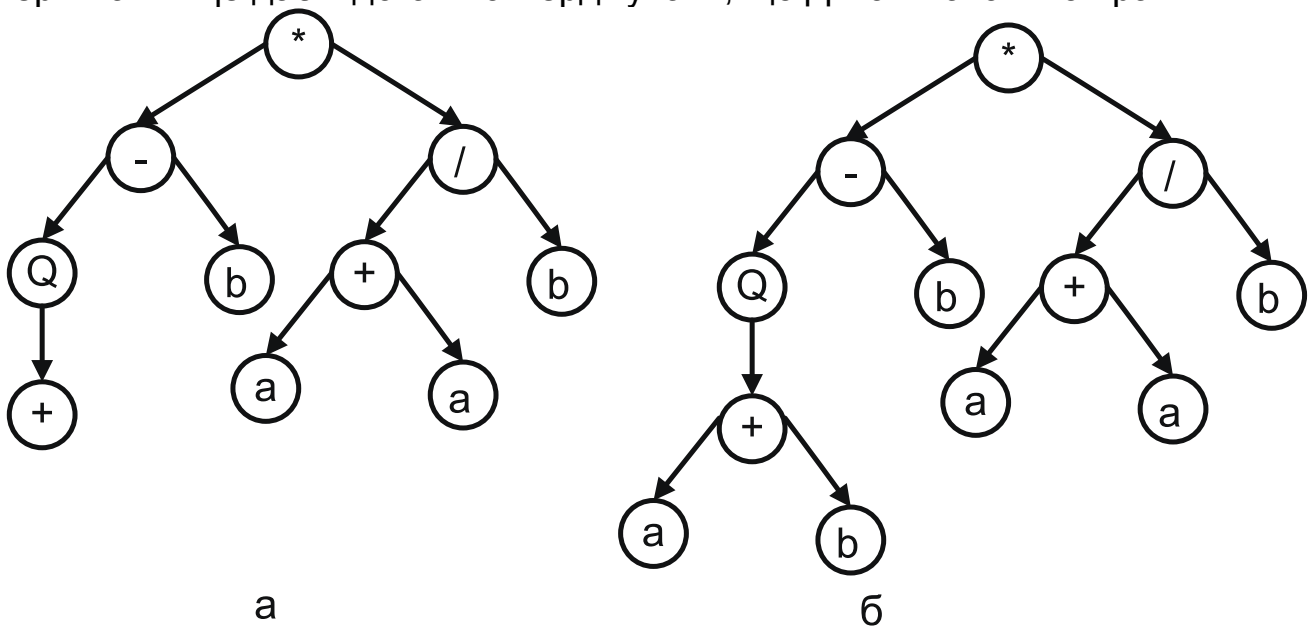


Рисунок 7.14 – Частково термінальна та повністю термінальна структура фреймів

Розглядаючи структуру ФВЧ, важко або навіть неможливо визначити переваги такого подання, окрім, можливо, його простоти і елегантності. Проте, коли ФБЧ проаналізувати в контексті генів, переваги цього подання стають очевидними. Як вже було вказано, хромосоми ПГВ мають фіксовану довжину і утворюються одним або декількома генами однакової довжини. Тому довжина гена також фіксована. Тоді в ПГБ змінюється не довжина генів, яка є постійною, а довжина ФБЧ. Дійсно, довжина ФВЧ є такою, що дорівнює довжині гена. У першому випадку точка зупинки збігається з кінцем гена, а в останньому – точка зупинки розміщена, де завгодно вгорі від кінця гена.

Виконаємо аналіз структурної організації генів ПГВ для того, щоб зрозуміти, як вони незмінно кодуються для синтаксично правильних програм і чому вони допускають застосування будь-якого генетичного оператора без обмежень.

Гени ПГВ. Гени ПГВ складаються з „голови" і „хвоста". „Голова" містить символи, які подають як функції, так і термінали, тоді як „хвіст" містить тільки

термінали. Для кожної задачі довжина „голови" h вибирається, тоді як довжина „хвоста" t – це функція від h і кількості аргументів функції з найбільшою кількістю аргументів n і обчислюється так:

$$t = h(n-1)+1 \ . \tag{7.15}$$

Розглянемо ген, для якого набір функцій є таким: $F = \{Q, *, /, -, +\}$ і набір терміналів – $T = \{a, b\}$. У цьому випадку $n = 2$ і якщо ми вибрали $h = 15$, то $t = 16$. Тому довжина гена g становить 15 + 16 = 31. Нижче показано один такий ген (,,хвіст" зображено жирним):

$$\begin{matrix} 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0 \\ /\ a\ Q\ /\ b\ *\ a\ b\ /\ Q\ a\ *\ b\ *\ -\ a\ b\ a\ b\ a\ a\ b\ a\ b\ b\ a\ b\ b\ b\ b\ a. \end{matrix} \tag{7.16}$$

Він кодується ДВ (рисунок 7.14, а). У цому випадку ФВЧ закінчується в позиції 7, тоді як ген закінчується в позиції 30.

Припустимо, що мутація відбувається у позиції 2, і змінимо Q на +. Тоді отримаємо такий ген:

$$\begin{matrix} 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0 \\ /\ a\ +\ /\ b\ *\ a\ b\ /\ Q\ a\ *\ b\ *\ -\ a\ b\ a\ b\ a\ a\ b\ a\ b\ b\ a\ b\ b\ b\ b\ a. \end{matrix} \tag{7.17}$$

Його подання буде таким, як показано на рисуноку 7.14, в. У цьому випадку точка зупинки зміщується на 10 позицій вправо.

Очевидно, що може трапитись і протилежне, і ФБЧ скоротиться. Наприклад, розглянемо знову ген (7.16) і припустимо, що мутація відбувалася у п'ятій позиції, замінюючи де „*" на „b":

$$\begin{matrix} 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0 \\ /\ a\ Q\ /\ b\ b\ a\ b\ /\ Q\ a\ *\ b\ *\ -\ a\ b\ a\ b\ a\ a\ b\ a\ b\ b\ a\ b\ b\ b\ b\ a. \end{matrix} \tag{7.18}$$

Його вираз приводить до такого ДВ (рисунок 7.14, б). У цьому випадку ФБЧ закінчується у пятій позиції, скорочуючи батьківське ДВ у двох вузлах.

Отже, незважаючи на його фіксовану довжину, кожен ген має потенціал для кодування ДВ різних розмірів і форм, будучи найпростіше складеним тільки з одного вузла (коли перший елемент гена – термінал) і найскладніше складеним з такої кількості вузлів, яка дорівнює довжині гена (коли всі елементи „голови" є функцією з максимальним числом аргументів).

Як видно з наведених прикладів, будь-яка модифікація, зроблена в геномі, незважаючи на її глибину, дає у результаті структурно правильну ДВ.

Єдина річ, з якою потрібно бути обережним, полягає у непорушності структурної організації генів при визначенні границі між „головою" і „хвостом". Не можна також дозволяти символам подавати функції у „хвості.

Ці питання розглядаються нижче, де і буде виконано аналіз механізмів та ефекту застосування різних генетичних операторів.

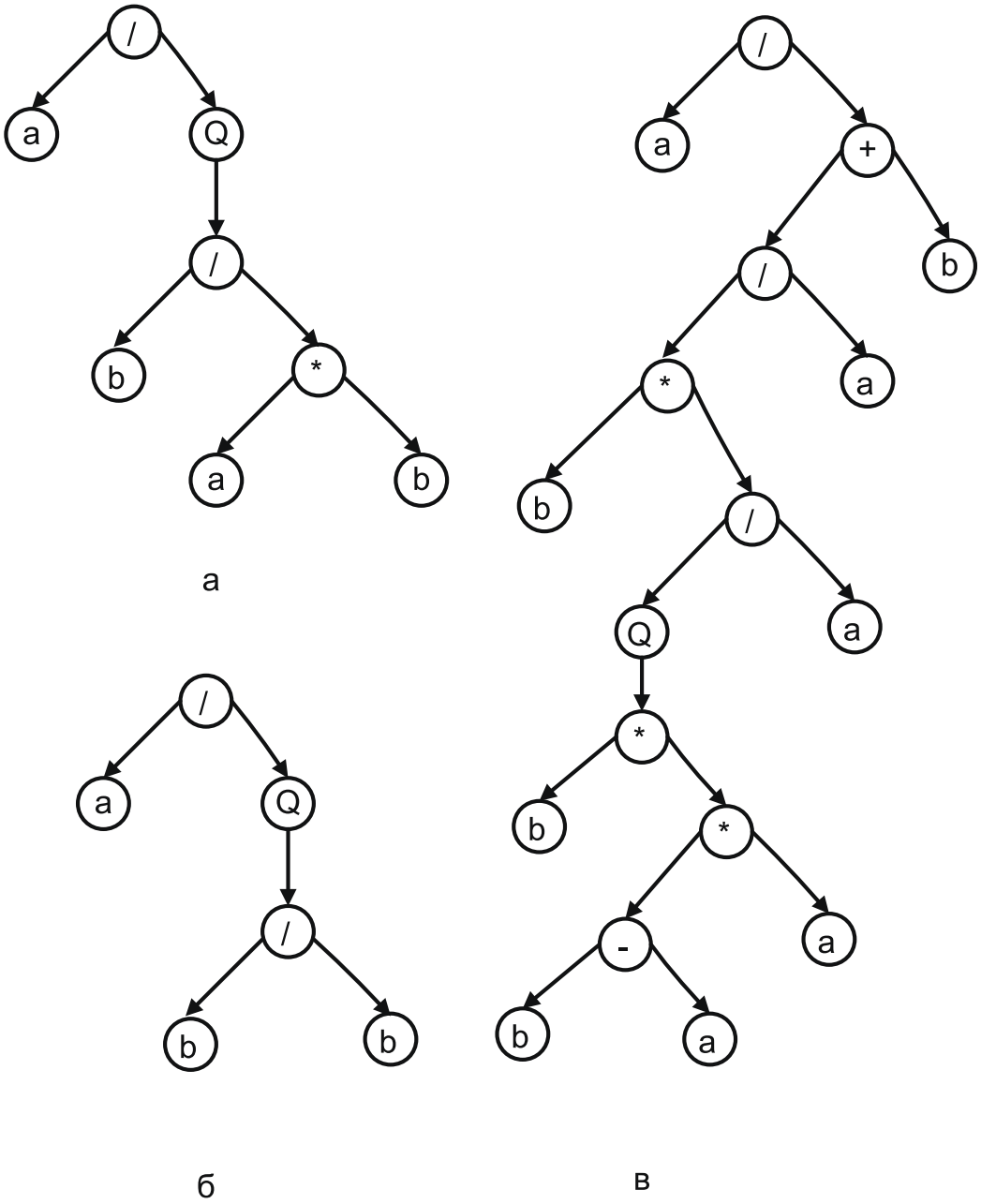


Рисунок 7.14 – Структура мутації генів

Генетичні оператори і еволюція. Генетичні оператори – ядро генетичиих алгоритмів, два з яких є загальними для усіх еволюційних систем: селекції і репродукції. Фактично вони можуть тільки спричиняти генетичний дрейф, роблячи популяції меншими і менш різноманітними з часом дoти, доки всі індивіди не стануть однаковими (7.19). Отже, основою усіх еволюційних систем є модифікація, або точніше – генетичні оператори, які є базою варіант. У різних алгоритмах модифікації виконуються по-різному. Наприклад, у ГА зазвичай використовуються мутація, рекомбінація; ГП базується на майже ексклюзивній

ГП-специфічній рекомбінації; а в ПГВ мають місце мутація, рекомбінація, переміщення.

Покоління № 0
012345678901201234567890 12
+- /a aaaaa aa/ /+*aaa aaaaaa -[0]= 10.64033
/-/a//aaaaa aa+*+a/+a aaaaaa-[1] =16.2117
*+a-+a aaaaa aa – - -*III a* aaaaaa- [2]=13.81953
+a*A a aaaaa aa**+a*aa aaaaaa- [3]=18.32701
*-+a/ – aaaaa aa /aa +a/a aaaaaa- [4]=11.13926
+*//a / aaaaa aa- – - aa-a aaaaaa- [5]= 13.88255
--*a aaaaa aa/ – a // /a aaaaaa- [6]=7.777691
/++a-*aaaaa aa/ + a*+-a aaaaaa- [7]= 13.14786
// +*aaaaaaa aa*+ -/ – -a aaaaaa- [8]= 7.713599
-**+- /aaaaa aa*/ /a a /a aaaaaa- [9]=8.73985

Покоління № 1
012345678901201234567890 12
*+a- +aaaaaaaa – - -/ / / aaaaaaa -[0]=13.81953
/- /a // aaaaaaa +*+a/ +aaaaaaa -[1]= 16.2117
* -+a/ – aaaaaaa / aa+a / aaaaaaa- [2]= 11.13926
+*//a / aaaaaaa – - – aa – aaaaaaa- [3]= 13.88255
+a*/- aaaaaaaa ** +a*a aaaaaaa- [4]= 18 32701
-**+- / aaaaaaa */ / aa/ aaaaaaa- [5]= 8.73985 (7.19)
-**+-/ aaaaaaa */ / aa/ aaaaaaa- [6] =8.73985
//+*aa aaaaaaa *+- / – - aaaaaaa- [7]=7.713599
/++a-*aaaaaaa /+a*+- aaaaaaa- [8] =13.14786
/- / a/ /aaaaaaa +*+a/+ aaaaaaa- [9]=16.2117

Покоління № 8
012345678901201234567890 12
/ -/ a // aaaaa aa+*+a/+ aaaaaaa -[0]= 16.2117
/-/a//aaaaa aa+*+a/+ aaaaaaa- [1] =16.2117
/-/a // aaaaa aa++a/+ aaaaaaa- [2]=16.2117*
/- / a // aaaaa aa+*+a/+ aaaaaaa- [3]=16.2117
/- / a // aaaaa aa+*+a/+ aaaaaaa- [4]=16.2117
/- / a // aaaaa aa+*+a/+ aaaaaaa- [5]=16.2117
/- / a // aaaaa aa+*+a/+ aaaaaaa- [6]=16.2117
/-/ a // aaaaa aa++a/+ aaaaaaa- [7] =16.2117*
/- / a // aaaaa aa+*+a/+ aaaaaaa- [8]=16.2117
/-/ a // aaaaa aa++a/+ aaaaaaa- [9]=16.2117*

За винятком ГП, яке серйозно обмежене засобами генетичної модифікації, у ГА і ПГВ можна здійснювати послідовність генетичних операторів, здатних спричиняти генетичну диверсифікацію, оскільки хромосоми обох алгоритмів легко дозволяють свою імплементацію. Фактично декілька генетичних операторів виконуються у ПГВ, проливаючи

світло на динаміку еволюційних систем, але, що важливо, вони дозволяють передбачити необхідний ступінь генетичної диверсифікації, щоб відбулася еволюція. Мутація сама собою (безумовно найголовніший оператор) здатна дивувати. Проте використання мутації та інших генетичних операторів не тільки дозволяє ефективну еволюцію, але і дублювання будівельних блоків, їх циркуляцію у генетичному пулі, створення послідовностей, що повторюються, і т.д., роблячи результати дійсно цікавими.

7.6 Контрольні запитання

1) Що є об'єктом вивчення у теорії нечітких множин?
2) У чому полягають відмінності теорії нечітких множин від теорії ймовірностей?
3) Дайте означення нечіткої множини і поясніть її суть.
4) Які існують типи подання для запису нечітких множин і відповідних функцій належності?
5) Які Ви знаєте типи функцій належності? У чому полягають їхні відмінності?
6) Які властивості мають нечіткі множини?
7) Наведіть приклади операцій над нечіткими множинами.
8) Для розв'язання яких задач може бути використано принцип узагальнення Заде?
9) Опишіть алгоритм узагальнення Заде.
10) Дайте означення лінгвістичної змінної і наведіть приклади.
11) Наведіть приклад застосування алгоритму узагальнення Заде.
12) Для розв'язання яких задач використовується композиційне правило виведення Заде?
13) Наведіть приклад використання композиційного правила Заде.
14) Опишіть алгоритм логічного виведення Мамдані та наведіть приклад.
15) Опишіть алгоритм логічного виведення Ларсена та наведіть приклад.
16) Опишіть алгоритм логічного виведення Цукамото та наведіть приклад.
17) Опишіть алгоритм логічного виведення Сугено та наведіть приклад.
18) Наведіть приклад розв'язання задачі з аналізом експертних нечітких заключень.

Бібліографічний список

1. Бокс, Дж. Анализ временных рядов, прогноз и управление [Текст] : в 20 т. / Дж. Бокс, Г. М. Дженкинс. – М. : Мир, 1974. –Т. 12. – 604 с.
2. Аляев, Ю. А. Дискретная математика и математическая логика [Текст] : учебник / Ю. А. Аляев, С. Ф. Тюрин. – М. : Финансы и статистика, 2006. – 368 с.
3. Box, G.E. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models [Text]/ G. E. Box, D. A. Pierce// Journal of the American Statistical Association. 1970. – № 65. – P. 1509 - 1526.
4. Эконометрия [Текст] / В. И. Услов, Н. М. Ибрагимов, Л. П. Талышева, А.А. Цыплаков. – Новосибирск : СО РАН, 2005. – 744 с.
5. Атникова, Т. А. Введение в эконометрический анализ панельных данных [Текст] / Т. А. Атникова // Экономический журнал ВШЭ. – 2006. – № 3. – С. 492–519.
6. Durbin, J.Testing for serial correlation in least-squares regression [Text] / J. Durbin, G. S. Watson // Biometrika. – 1951. – № 38. – P. 159–178.
7. Чернова, Н. И. Теория вероятностей [Текст] : учеб. пособие / Н. И. Чернова. – Новосибирск: Новосиб. гос. ун-т. 2007. – 160 с.
8. Якушев, А. И. Взаимозаменяемость, стандартизация и технические измерения [Текст] / А. И. Якушев, Л. Н. Воронцов, Н. М. Федотов. – 6-е изд., перераб. и доп. – М. : Машиностроение, 1986. – 352 с.
9. Романовский, П. И. Ряды Фурье. Теория поля. Аналитические и специальные функции. Преобразования Лапласа [Текст] / П. И. Романовский. – М. : Наука, 1980. – 336 с.
10. Айвазян, С. А.Прикладная статистика. Основы эконометрики [Текст] : учебник для вузов: в 2 т. / С. А. Айвазян. – 2-е изд., испр. – М. : ЮНИТИ- ДАНА, 2001. – Т. 2. – 432 с.
11. Линник, Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений [Текст] / Ю. В. Линник. – 2-е изд. – М. : Физматгиз 1962. – 349 с.
12. Степашко, В. С. Теоретичні аспекти МГУА як методу індуктивного моделювання [Текст] / В. С. Степашко// Управляющие системы и машины. – 2003. – № 2. – С. 31 – 44.
13. Ивахненко, А. Г. Метод групового урахування аргументів – конкурент методу стохастичної апроксимації [Текст] / А. Г. Ивахненко // Автоматика. – 1968. – № 3. – С. 58 – 72.
14. Гмурман, В. Е. Теория вероятностей и математическая статистика [Текст] : учеб. пособие для вузов / В. Е. Гмурман. – 10-е изд., стереотип. – М. : Высш. шк., 2004. – 479 с.
15. Елисеева, И. И. Общая теория статистики [Текст] : учебник / И. И. Елисеева, М. М. Юзбашев / под ред. И. И. Елисеевой: – 4-е изд., перераб. и доп. – М. : Финансы и статистика, 2002. – 480 с.
16. Эконометрия [Текст] / В. И. Суслов, Н. М. Ибрагимов, Л. П. Талышева, А. А. Цыплаков. – Новосибирск: СО РАН, 2005. – 744 с.
17. Крамер, Г. Математичні методи статистики [Текст] : пер. англ./ Г. Крамер. – М. : 1948. – 648 с.

18. Магнус, Я. Р. Эконометрика. Начальный курс [Текст] / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – М. : Дело, 2007. – 504 с.
19. Абраменко, А. В. Лекции по курсу «Методы вычислений и программирование» [Электронный ресурс] / А. В. Абраменко. – Режим доступа: ftp://freesky.no-ip.org/incoming/freesky/213/Theory/NL_LSFit.pdf. – 2011 г.
20. Сараев, П. В. Нелинейный метод наименьших квадратов и блочные рекуррентно-итерационные процедуры в обучении нейронных сетей [Текст] / П. В. Сараев // Управление большими системами. – 2010. – Вып. 30. – М. : ИПУ РАН. – С. 24 – 34.
21. Самыловский, А. И. Учебно-методический комплекс по учебным дисциплинам «Теория вероятностей» и «Математическая статистика» для социологов [Текст] : учеб.-метод. пособие / А. И. Самыловский. – 3-е изд., перераб. и доп. – М. : МГУ, 2010. – 60 с.
22. Akaike, H.A. New look at the statistical model identification [Text] / H. A. Akaike // IEEE Transactions on Automatic Control. – 1974. – Т. 19. – P. 716 – 723.
23. Burnham, K.P. Model selection and multimodel inference: a practical information-theoretic approach [Text] / K. P. Burnham, D. R. Anderson – Springer, 2002. – 488 p.
24. Ветров, Д. П. Байесовские методы машинного обучения [Текст] : учеб. пособие по спецкурсу / Д. П. Ветров, Д. А. Кропотов. – М.: МГУ, 2007. – 67 с.
25. Fisher R. A., On a distribution yielding the error functions of several well known statistics [Text] / R. A. Fisher. Proc. Intern. Math. Congr. Toronto". – 1928. – V. 2. – P. 805–813.
26. Венчиков, А. И. Основные приемы статистической обработки результатов наблюдений в области физиологии [Текст] : монография / А. И. Венчиков, В. А. Венчиков. –2-е изд., перераб. и доп. – М. : Медицина, 1974. – 152 с.
27. Cowell, F. A. How much inequality can we explain? [Text] / F. A. Cowell, S. P. Jenkins // A methodology and an application to the United States, Economic Journal. – 1995. – № 105 (429). – P. 421 – 30.
28. Тимошик, Н. П. Застосування методу групового врахування аргументів для системи автоматичної побудови моделей мережевого трафіку систем виявлення атак [Текст] / Н. П. Тимошик // Вісник Національного університету "Львівська політехніка". Автоматика, вимірювання та керування. – 2009. – № 639. – С. 133–141.
29. Ивахненко, А. Г. Справочник по типовым программам моделирования [Текст] / А. Г. Ивахненко, Ю. В. Коппа, В. С. Степашко. – К. : Техніка, 1980. – 184 с.
30. Gabor, D. A universal nonlinear filter, predictor and simulator which optimizes itself by a learning process. [Text] / D. Gabor, W. R. Wilby, R. A. Woodcock // Proc. Inst. Electr. Engrs. – 1961. – Vol. 108., part. B, № 40. – P. 85–98.
31. Успенский, В. А. Теорема Гёделя о неполноте [Текст] / В. А. Успенский. – М. : Наука, 1982. – 110 с.
32. Сосинский, А. Б. Теорема Геделя. Нормальные формы, симметричность, минимальность и красота в математике и в природе

[Электронный ресурс] / А. Б. Сосинский. – Режим доступа: http://www.mathnet.ru/php/presentation.phtml?option_lang=rus&presentid= 7343. – 28 июля 2013 г.

33. Міщенко, Н. М. Байєсові мережі як засіб моделювання причинно-наслідкових зв'язків у системах з неозначеністю [Текст] / Н. М. Міщенко // Проблемы программирования. – 2002.– № 4. – С. 125–131.

34. Pearls, J. Bayesian inference methods [Text] / J. Pearls. – Encyclopedia of Artificial Intelligence, Sec. Ed., Vol. 1, A Wiley-Interscience Publicationю. – New York, 1992. – P. 89–98.

35. Заде, Л. Понятие лингвистической переменной и его применение к принятию приближенных решений [Текст] / Л. Заде. – М. : Мир, 1976. – 165 с.

36. Заде, Л. А. Роль мягких вычислений и нечеткой логики в понимании, конструировании и развитии информационных интеллектуальных систем [Текст] / Л. А. Заде// Новости искусственного интеллекта. – № 2, 3. – 2001. – С. 7–11.

37. Круглов, В. В. Нечёткая логика и искусственные нейронные сети [Текст] : моногр. / В. В. Круглов, М. И. Дли, Р. Ю Голунов. – М. : Физматлит, 2001. – С. 384.

38. Пегат, А. Нечеткое моделирование и управление / пер. с англ. [Текст] / А. Пегат. – М. : БИНОМ, 2009. – 798 с.

39. Ягер, Р. Р. Нечеткие множества и теория возможностей. Последние достижения [Текст] / Р. Р. Ягер / под ред. Р. Ягера. – М. : Радио и связь, 1986. – 408 с.

40. Дьяконов, В. Алгоритмы нечёткого вывода: алгоритм Мамдани и алгоритм Сугэно. Математические пакеты расширения MATLAB [Текст] : Спец. справочник / В. Дьяконов, В. Круглов. – СПб. : Питер, 2001. — С. 307–309

41. Рутковская, Д. Нейронные сети, генетические алгоритмы и нечеткие системы [Текст]: / Д. Рутковская, М. Пилиньский, Л. Рутковский / пер. с польск. И. Д. Рудинского. – М. : Горячая линия Телеком, 2006. – 452 с.

42. Демидова, Л. Алгоритмы и системы нечеткого вывода при решении задач диагностики городских инженерных коммуникаций в среде Matlab [Текст] / Л. Демидова, В. В. Кираковский, А. Н. Пылькин. – М. : Радио и связь, Горячая линия Телеком, 2005. – 368 с.

Зміст

Навчальне видання

Розсоха Сергій Володимирович
Туркін Ігор Борисович
Шостак Ігор Володимирович

ФОРМАЛЬНІ МЕТОДИ ІДЕНТИФІКАЦІЇ ТА ПРОГНОЗУВАННЯ ОБ'ЄКТІВ І ПРОЦЕСІВ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

Редактор С. П. Гевло

Зв. план, 2014
Підписано до друку 18.06.2014
Формат 60×84 1/16. Папір офс. № 2. Офс. друк
Ум. друк. арк. 7,5. Обл.-вид. арк. 8,38. Наклад 100 пр.
Замовлення 241. Ціна вільна

Видавець і виготовлювач
Національний аерокосмічний університет ім. М. Є. Жуковського
"Харківський авіаційний інститут"
61070, Харків-70, вул. Чкалова, 17
http://www.khai.edu
Видавничий центр «ХАІ»
61070, Харків-70, вул. Чкалова, 17
izdat@khai.edu

Свідоцтво про внесення суб'єкта видавничої справи
до Державного реєстру видавців, виготовлювачів і розповсюджувачів
видавничої продукції сер. ДК № 391 від 30.03.2001