Київський політехнічний інститут

На правах рукопису

ЖОЛНАРСЬКИЙ Олександр Анатолійович

УДК 681.513

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ МЕТОДІВ МОДЕЛЮВАННЯ І КЛАСТЕР-АНАЛІЗУ НЕЧІТКИХ ОБ'ЄКТІВ ПРИРОДНОГО СЕРЕДОВИЩА

05.13.01 - керування в технічних системах

А В Т О Р Е Ф Е Р А Т

дисертації на одержання наукового ступеня
кандидата технічних наук

КИЇВ - 1993

Роботе виконана в Київському політехнічному інституті на кафедрі технічної кібернетики

Науковий керівник – член-кореспондент АН України, доктор технічних наук, професор ІВАХНЕНКО О.Г.

Офіційні опоненти – доктор технічних наук, професор ЗАЙЧЕНКО Ю.П.
– кандидат технічних наук АКІШИН Б.О.

Ведуче підприємство – Інститут проблем моделювання в енергетиці АН України, м. Київ

Захист дисертації відбудеться "20" грудня 1993 року в "15" годин на засіданні спеціалізованої ради К 068.14.01 в Київському політехнічному інституті за адресою: 252056, Київ – 56, проспект Перемоги, 37, 503 ауд. 22 корпус

З дисертацією можно ознайомитись в бібліотеці Київського політехнічного інституту.

Автореферат розіслан "___"_____ 1993 року.

Учений секретар
спеціалізованої ради ШУЛЬГА Ю.І.

А Н О Т А Ц І Я

Ціллю дисертаційної роботи є розроблення нових параметричних та непараметричних методів моделювання та кластер-аналізу об'єктів з нечіткими функціонально-структурними зв'язками.

Для досягнення поставленої цілі в роботі вирішені наступні задачі:

- виконан аналіз природних систем як об'єкта керування та дана загальна характеристика основних методів моделювання таких систем;

- розроблен алгоритм структурної ідентифікації в класі гармонічних функцій, що задані за допомогою бінарного коду, для одномірного та двумірного випадків;

- модифікован метод інструментальних змінних для ідентифікації оцінок коефіцієнтів в параметричних алгоритмах метода групового врахування аргументів;

- запропонован ієрархічний (агломеративний) алгоритм кластеризації многомірних об'єктів та доведена його збіжність;

- розроблен непараметричний алгоритм комплексування аналогів для прогноза багатомірних випадкових процесів та повторюющихся подій;

- розроблена концепція автоматизованої системи наукових досліджень для задач моніторінга нечітких об'єктів природного середовища;

- створен та відлажен комплекс програм моделювання та кластер-аналізу нечітких об'єктів природного середовища.

Автор захищає:

1. Алгоритм структурної ідентифікації в класі гармонічних функцій, що задані у вигляді бінарного коду, для одномірного та двумірного випадків.

2. Модифікований метод інструментальних змінних для ідентифікації оцінок коефіцієнтів в параметричних алгоритмах метода групового врахування аргументів.

3. Ієрархічний (агломеративний) алгоритм кластеризації багатомірних об'єктів та доведення його збіжності.

4. Непараметричний алгоритм комплексування аналогів для прогноза багатомірних випадкових процесів та повторюющихся подій.

5. Автоматизовану систему наукових досліджень для задач моніторінга нечітких об'єктів природного середовища.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність проблеми. У зв'язку з інтенсифікацією діяльності людини в навколишньому природному середовищі особливе значення надається задачі моделювання та прогнозу стану екологічних об'єктів (повітря, вода, грунт), що має ціль прийняття керувальницьких рішень на різних масштабних рівнях (глобальному, регіональному, локальному). Специфіка дослідження природних систем придає неабияке значення розробленню методів математичного моделювання та прогнозу за допомогою перебору моделей-кандидатів за деяким ансамблем зовнішніх критеріїв (індуктивний підхід). Необхідність розробки методів, що засновані на індуктивному підході зумовлена тим, що вони подають зручний та підходящий апарат моделювання та прогнозу характеристик екологічних об'єктів, які відносяться до класу складних, погано формалізуємих систем при присутності розмитих (погано означених) залежностей, недостатній вивченості деяких закономірностей і т.д. Так як екологічні об'єкти відносяться до класу нечітких (розмитих), недоозначених систем, то відповідно закону адекватності Ст.Біра, математичний апарат моделювання таких систем повинен бути також нечітким. Адекватним математичним апаратом при описуванні нечітких об'єктів є патерн-аналіз або його розвиток кластер-аналіз. Тобто, для того щоб врахувати вимогу принципу адекватності при моделюванні та прогнозі стану екологічних об'єктів, необхідно застосувати перебір кластеризацій або дискретизацій вибірки даних спостережень про стан досліджуваного об'єкту. Перебір кластеризацій по зовнішньому критерію має тіж логічні підстави і тіж цілі, що й перебір рівнянь моделей по алгоритмам методу групового врахування аргументів (МГВА).

У зв'язку з викладенним вище, набуває актуальності подальше дослідження і розробка ефективних методів моделювання та кластер-аналізу, заснованих на ідеях теорії самоорганізації, і призначених

для моделювання та прогнозу нечітких об'єктів природного середовища.

Методи дослідження. В роботі використовувались методи теорії самоорганізації математичних моделей, теорії класифікації багатомірних об'єктів та зниження розмірності, основні положення теорії систем.

Наукова новизна. При використанні формального апарату поставлена та розв'язана задача переходу від бінарних даних до неперервних, для чого розроблено алгоритм гармонічної реконструкції вибірки даних X, що дозволяє точно відновити невідому неперервну функцію, задану у вигляді бінарного коду. Розроблено модифікований метод інструментальних змінних для ідентифікації оцінок коефіцієнтів лінійних регресійних моделей, який удосконалений за допомогою зменшення розміру зсуву, досягаємого лінейною апроксимацією вибірки даних. Запропоновано ієрархічний алгоритм кластер-аналізу, заснований на використанні порогових значень, а також приведено доведення збіжності агломеративного алгоритму за кінцеве число кроків до правильної кластеризації. Приведено непараметричні переборні алгоритми прогнозу випадкових багатомірних процесів та подій, що повторюються, в яких відсутня необхідність оцінки параметрів. Розроблена структура автоматизованої системи наукових досліджень об'єктів природного середовища, заснована на ідеях нової інформаційної технології, що дозволяє скоротити затрати часу на обробку даних та прийняття рішення.

Практична цінність роботи заключається в розробці алгоритмів індуктивного моделювання та прогнозу нечітких об'єктів природного середовища, що надають можливість значно скоротити затрати часу на обробку даних та прийняття рішення по керуванню такими об'єктами. Приведена модель автоматизованої системи моніторинга об'єктів навколишнього природного середовища, до складу якої входить весь комплекс розглянутих в роботі методів, що надає можливість зв'язати всі етапи експериментального дослідження від постановки задачі до прийняття рішення в єдиний ланцюг.

Реалізація роботи. Розроблені алгоритми індуктивного моделювання та прогнозу реалізовані у вигляді проблемно-орієнтованого пакету прикладних програмних модулей (ППМ) "Методи моделювання та кластер-аналізу нечітких об'єктів". ППМ впроваджено в міжгалузевому інженерному центрі "Планета" (м.Луганск) з економічним ефектом 39005 карбованців. Джерелом економії виступає скорочення затрат на науково-дослідні роботи при розробці алгоритмічного та програмного забезпечення автоматизованих систем моніторинга об'єктів навколишнього середовища.

Апробація роботи. Основні результати роботи доповідалися та обговорювались на: конференції "Бази знань та експерні системи в АСНД" (м.Севастополь, 1990 р.); I та II Всесоюзних школах-семінарах молодих вчених та спеціалістів "Сучасний стан теорії та розробки програмного забезпечення систем керування з ЕОМ"(Самарканд, 1990 р., 1991 р.); школі-семінарі "Навчання розпізнаванню образів та кластеризація" (Київ, 1991 р.); IY Всесоюзній школі-семінарі "Статистичний і дискретний аналіз даних та експертне оцінювання"(Одеса, 1991 р.).

Публікації. По матеріалам дисертації опубліковано 9 друкованих робіт.

Структура та обсяг роботи. Дисертаційна робота складається з вступу, чьотирьох глав, висновку, списку літератури та додатку. Загальний обсяг роботи 150сторінок, включаючий основний текст, 13 малюнків, 10 таблиць, список літератури зі 120 найменувань, додаток, в якому приведені документи, підтверджуючі впровадження, а також лістінги програм, які входять до складу ППМ.

ЗМІСТ ДИСЕРТАЦІЙНОЇ РОБОТИ

В роботі викладається актуальність теми, ціль та задачі досліджень, наукова новизна, методика досліджень, практична значимість та реалізація роботи, зв'язок з тематикой інститута.

На основі проведеного порівняльного аналізу дедуктивного та індуктивного подходів до моделювання складних природних систем

встановлена необхідність та коректність розробки індуктивних методів моделювання та кластер-аналізу для рішення складних задач в складі АСНД екологічних систем.

Нові непараметричні алгоритми методу групового врахування аргументів (МГВА) розроблені як для неперервних, так і для бінарних вхідних змінних. Бінарні змінні незручні як для прогнозу процесів за допомогою поліноміальних алгоритмів МГВА, так і для комплексування аналогів, де замість формул зважуваного підсумування приходиться застосовувати процедури голосування. В роботі приведені постановка та рішення задачі переходу від бінарних признаків до неперервних (задача дебінаризації) за тим, щоб далі для задач розпізнавання чи прогнозу застосовувати вже добре розроблені алгоритми. Тому задача дебінаризації вирішується за допомогою гармонічного тренду

$$f(t) = \sum_{i=1}^{n} a_i \sin w_i t + b_i \cos w_i t + \rho_0 , \qquad (1)$$

де n – кількість частот (кількість гармонічних складаючих); a_i, b_i – лінейно входячі в (I) коефіцієнти моделі; w_i – частота; ρ_0 – вільний член (для квазістаціонарного процесу $\rho_0 = 0$). Задача полягає в визначенні параметрів моделі (I) по значенням невідомої функції f(t) в точках де f(t) = 0 та одному значенню функції f(t) (наприклад, останньому по часу при t = T). Метод рішення задачі включає застосування методу найменьших квадратів для оцінки коефіцієнтів ряду та методу повного перебору складу множини гармонічних складаючих по критерію зміщеного контролю CV. Підстановка даних в (1) надає можливість отримати систему алгебраічних рівнянь виду Y = QX для визначення оцінок коефіцієнтів ряду, де Q – вектор невідомих параметрів a_i, b_i; Y – вектор вихідних величин; X – матриця, яка складається із значень $\sin w_i t$, $\cos w_i t$. Рішення системи знаходиться за допомогою методу найменьших квадратів. Інформація про одне неперервне значення процесу необхідна, щоб уникнути тривіального рішення.

Якщо застосовувати повний перебір гармонічних функцій тільки по методу найменьших квадратів, то у зв'язку з малим числом точок, де функція f(t) відома, оцінка коефіцієнтів a_i, b_i буде значно неточною. Для збільшення точності оцінки в роботі застосовується критерій змінного контролю

$$CV = \frac{1}{N-1}\sum_{i=1}^{N-1}\left[f(t_i) - \hat{f}(t_i)\right]^2 \longrightarrow \min ,$$

де i - номер змінної строки вибірки даних, f(t) - фактичне значення функції в останній точці вибірки, $\hat{f}(t)$ - значення функції в останній точці, розраховане по оцінюваному гармонічному ряду, N - число строк вибірки. Для контролю неможливо застосовувати задане неперервне значення функції f(t) , так як в цьому випадку функція буде відновлюватись тільки по нульовим ординатам, що приведе до тривіального рішення.

Для оцінки ступеня згоди апроксимуючої моделі з відновленою функцією використовується похибка відновлення

$$\delta^2 = \frac{\sum_{i=1}^{Nmax}\left[f(t_i) - \hat{f}(t_i)\right]^2}{\sum_{i=1}^{Nmax}\left[f(t_i) - \bar{f}\right]^2} \leqslant 1,0 ,$$

де $\bar{f}$- середнє значення функції f(t), Nmax - визначається довжиною інтервала T. В роботі рішення задачі реінтерполяції представлено на прикладі числового моделювання сонячної активності по середньорічним значенням чисел Вольфа. При цьому похибка відновлення становила $\delta^2 \approx 0.256$, що свідчить про гарну якість відновлення.

Але в тих випадках, коли показник обумовленості матриці нормальних рівнянь X^TX занадто великий, чи при значних дисперсіях завади, а також коли закон розподілу завади відрізняється від нормального, похибка ідентифікації оцінок коефіцієнтів в параметричних алгоритмах МГВА буде зростати. Це пов'язано з

використанням в цих алгоритмах як основної процедури ідентифікації методу найменших квадратів. В дисертації за допомогою числового моделювання показано, що в тих випадках, коли умови для використання МНК не сприятливі, слід переходити до використання метода інструментальних змінних (МІЗ). Якщо вихідна вибірка даних не представляє собою реалізацію випадкового процесу, то пропонується скористуватися процедурою трансформації вихідних даних: строки об'єднаної матриці (XY) розміщуються зверху до низу по мірі збільшення значення $l_{ij} = \frac{1 - K_{ij}}{2}$, де K_{ij} - коефіцієнт кореляції між строками. Нехай вихідна матриця X та вектор Y мають вигляд

X_{11}	X_{12}	X_{13}	Y_1
X_{21}	X_{22}	X_{23}	Y_2
X_{31}	X_{32}	X_{33}	Y_3
X_{41}	X_{42}	X_{43}	Y_4
X_{51}	X_{52}	X_{53}	Y_5
X_{61}	X_{62}	X_{63}	Y_6

Припустимо, що в приведеній структурі строки вже розміщені по мірі зростання коефіцієнта l_{ij}. Тоді для визначення коефіцієнтів Q_1 для моделі виду $Y(t,Q) = X^T(t)Q + W(t)$ пропонується виконати наступні заходи:

I. Робимо зсув L = 1

$$y_5 y_5 = \hat{Q}_1 x_{51} y_6 + \hat{Q}_2 x_{52} y_6 + \hat{Q}_3 x_{53} y_6$$
$$y_5 y_4 = \hat{Q}_1 x_{41} y_5 + \hat{Q}_2 x_{42} y_5 + \hat{Q}_3 x_{43} y_5$$
$$y_4 y_3 = \hat{Q}_1 x_{31} y_4 + \hat{Q}_2 x_{32} y_4 + \hat{Q}_3 x_{33} y_4$$
$$y_3 y_2 = \hat{Q}_1 x_{21} y_3 + \hat{Q}_2 x_{22} y_3 + \hat{Q}_3 x_{23} y_3$$
$$y_2 y_1 = \hat{Q}_1 x_{11} y_2 + \hat{Q}_2 x_{12} y_2 + \hat{Q}_3 x_{13} y_2.$$

Отримані рівняння сумуються

$$\sum_{i=2}^{6} y_i y_{i-1} = \hat{Q}_1 \sum_{i=2}^{6} x_{i-1_1} y_i + \hat{Q}_2 \sum_{i=2}^{6} x_{i-1_2} y_i + \hat{Q}_3 \sum_{i=2}^{6} x_{i-1_3} y_i. \quad (2)$$

2. Робимо зсув L = 2

$$y_6 y_4 = \hat{Q}_1 x_{41} y_6 + \hat{Q}_2 x_{42} y_6 + \hat{Q}_3 x_{43} y_6$$
$$y_5 y_3 = \hat{Q}_1 x_{31} y_5 + \hat{Q}_2 x_{32} y_5 + \hat{Q}_3 x_{33} y_5$$
$$y_4 y_2 = \hat{Q}_1 x_{21} y_4 + \hat{Q}_2 x_{22} y_4 + \hat{Q}_3 x_{23} y_4$$
$$y_3 y_1 = \hat{Q}_1 x_{11} y_3 + \hat{Q}_2 x_{12} y_3 + \hat{Q}_3 x_{13} y_3$$

Отримані рівняння також сумуються

$$\sum_{i=3}^{6} y_i y_{i-2} = \hat{Q}_1 \sum_{i=3}^{6} x_{i-2_1} y_i + \hat{Q}_2 \sum_{i=3}^{6} x_{i-2_2} y_i + \hat{Q}_3 \sum_{i=3}^{6} x_{i-2_3} y_i. \quad (3)$$

Аналогічно (2) та (3) формується рівняння для L = 3

$$\sum_{i=4}^{6} y_i y_{i-3} = \hat{Q}_1 \sum_{i=4}^{6} x_{i-3_1} y_i + \hat{Q}_2 \sum_{i=4}^{6} x_{i-3_2} y_i + \hat{Q}_3 \sum_{i=4}^{6} x_{i-3_3} y_i. \quad (4)$$

Рівняння (2), (3) та (4) предсатвляють систему рівнянь, з яких будь-яким із знайомих методів можуть бути отримані коефіцієнти Q_i ($i = \overline{1,3}$).

Вплив завади на результат ідентифікації параметрів можно ще зменьшити, якщо застосувати так зване ослаблення вихідної вибірки даних. Тобто, після "перемішки" об'єднана матриця (XY) розширюється за допомогою лінейної інтерполяції між строками. Додаткові строки генеруються за формулою

$$Z_{n-\Delta\lambda} = (1 - \Delta\lambda)Z_n - \lambda Z_{n-1},$$

де $\Delta\lambda = 0{,}1; 0{,}2; \ldots; 1$; Z_n, Z_{n-1} - два сусідні значення матриці (XY). Далі для отримання таким чином даних використовується

модифікований МІП.

Недолік всіх алгоритмів МГВА полягає в оцінці параметрів за допомогою регресійного аналізу, який має ряд обмежень: завади повинні діяти тільки на вихідні змінні чи рівномірно на всі регресори; набір регресорів повинен бути повним, так як невраховані регресори діють як додаткове збільшення завади. Оцінки коефіцієнтів отримуються зміщеними як при присутності завади в даних, так і при неповному числі вихідних змінних. Істина модель є найбільш простою із всіх незміщених моделей, які отримуються при переборі в повній відсутності завад (при точних даних чи при безмежно довгій вибірці). В роботі приведені нові непараметричні алгоритми МГВА, в яких відсутня необхідність оцінки параметрів.

Алгоритм прогнозування випадкових многомірних процесів. Нехай задана вибірка даних X, яка містить N строк (точок) вимірювання M характеристичних змінних через рівні інтервали часу (шаги). Кожна строка вибірки (характеристичний вектор) відповідає деякій точці в просторі характеристичних змінних (ознак). Для останньої по часу спостереження строки B вибірки X треба знайти аналоги, тобто найближчі в змісті деякої міри близості до неї точки A1, A2, ..., AF ($F \leqslant N$). Припустимо, що хід процесу після вихідної точки (тобто шуканий прогноз) близький до ходу процесів, які мали місце після аналогів. Тому прогноз визначається зваженим сумуванням чи комплексуванням прогнозів F аналогів по рівнянням сплайнів.

Для того, щоб прогноз процесу був по властивості точним, треба за допомогою перебору варіантів вирішити наступні задачі оптимізації параметрів алгоритму:

- вибір оптимального числа комплексуємих аналогів $F = F_{opt}$;
- вибір оптитмального набору ознак $m = m_{opt}$;
- вибір допустимої ширини шага вимірювання змінних $n = n_{max}$;

В якості критерія вибору оптимального рішення доцільно використовувати повний критерій змінного контролю. Вибирається той варіант рішення, який дає більш глибокий мінімум цього критерію $CV \longrightarrow min$.

Алгоритм прогнозування повторюючих випадкових подій. Для рішення задачі прогнозування повторюючих випадкових подій

необхідно крім матриці X задати також матрицю векторів вихідних величин Y . Для цієї задачі важливо, щоб були корельовані стовбці матриці X та вибірки Y. Кореляція між строками вибірки звичайно відсутня.

Для прогнозування подій вирішуються наступні задачі:

- вибір оптимального числа комплексируємих аналогів $F = F_{opt}$;
- вибір оптитмального набора ознак $m = m_{opt}$;
- вибір оптимального вектору функції цілі $Y = Y_{opt}$;

Ширина патерна не підлягає зміні та рівна одній строці шуканої вибірки. Тому доцільно виконати перебір компонентів вектору функції цілі Y.

Робота алгоритмів прогнозування за допомогою комплексування аналогів перевірена на задачі короткочасного прогнозу кількості кисню в водах Київського водосховища.

Прогноз по аналогам буде неефективним, якщо попередньо не виконана кластеризація вибірки даних на змінюючи один одного характерні етапи розвитку багатомірного процесу.В роботі запропоновано агломеративний алгоритм (А-алгоритм) кластер-аналізу, який базується на ідеях методу динамічних згущень, однак на кожному рівні ієрархії використовуються не первичні точки, а їх оцінки (представники) - центри ваги утворених кластерів.

Зміст А-алгоритму. Початок. Нехай задана послідовність порогових значень $d_1 \leqslant d_2 \leqslant \ldots \leqslant d_{r_n} = d$; початковий набір середніх $e^0 = (e_1^0, \ldots, e_{N_0}^0)$, $e_i^0 = x_i$, $i \in [1, N_0]$; початкове розбивання $S^0 = (S_1^0, \ldots, S_{N_0}^0)$, де $S_p^0 = (x_p)$, $\bigcup_{p=1}^{N_0} x_p = X$, $S_p^0 \cap S_{p'}^0 = 0$, $p \neq p'$.

Шаг I. Визначим розбивання $S^1 = (S_1^1, \ldots, S_{N_1}^1)$, яке породжується набором e^0, по принципу знаходження точки x_j, найближчої в радіусі d_1 до центру e_1^0, т.б.:

$$S_1^1 = \begin{cases} x_j,\ x_j \in X, \text{ при } x_j = \underset{x_k \in X}{\operatorname{argmin}} d(x_k, e_1^0), x_k = e_1^0, \\ d(x_k, e_1^0) \leqslant d_1,\ x_j \neq 0, \\ x_1,\ \text{при } x_j = 0. \end{cases}$$

$x_p,\ x_j \in X,\ x_j = \underset{x_k \in X}{\operatorname{argmin}}\, d(x_k, e_p^0),\ x_k' \neq e_p^0,$

$$S_1^r = \begin{cases} d(x_k, e_p^0) \leqslant d_1,\ x^p = X - \bigcup_{p'=1}^{p-1} S_{p'},\ x_j \neq 0, \\ x_p, \quad \text{при } x_j = 0, \quad p \in [2, N_1]. \end{cases}$$

Змінні наступного рівня ієрархії знаходимо за правилом:

$$e_p^1 = y_p^1 = \begin{cases} \dfrac{x_p + x_j}{2}, & \text{якщо } x_j \neq 0 \text{ і } S_p = (x_p, x_j); \\ x_p, & \text{якщо } x_j = 0 \text{ і } S_p = (x_p), \ p \in [1, N_1]. \end{cases}$$

Основний цикл. Шаг r. На r-ій ітерації (r>1) знаходимо послідовно за рекурентними формулами:

розбивання $S^r = (S_1^r, \ldots S_{N_r}^r)$

$$S_1^r = \begin{cases} S_1^{r-1} \cup S_{j_1}^{r-1}, \text{ при } y_{j_1}^{r-1} = \underset{y_k^{r-1} \in Y}{\operatorname{argmin}}\, d(y_k^{r-1}, e_1^{r-1}), \\ d(y_{j_1}^{r-1}, e_1^{r-1}) \leqslant d_r,\ y_k^{r-1} \neq e_1^{r-1},\ y_{j_k}^{r-1} \neq 0, \\ S_1^{r-1}, \quad \text{при } y_{j_1}^{r-1} = 0; \end{cases}$$

$$S_p^r = \begin{cases} S_p^{r-1} \cup S_{j_p}^{r-1}, \text{ при } y_{j_p}^{r-1} = \underset{y_k^{r-1} \in Y^p}{\operatorname{argmin}}\, d(y_k^{r-1}, e_p^{r-1}), \\ y_p^{r-1} \neq e_p^{r-1}, \quad Y^p = Y / \bigcup_{p'=1}^{p-1} S_{p'},\ y_{j_p}^{r-1} \neq 0, \\ S_p^{r-1}, \quad \text{при } y_{j_p}^{r-1} = 0; \end{cases}$$

змінні r-го рівня ієрархії (середні значення $e_p^r = y_p^r$)

$$y_p^r = \begin{cases} \dfrac{n_p y_p^{r-1} + n_{j_p} y_{j_p}^{r-1}}{n_p + n_{j_p}}, \text{ при } y_p^{r-1} \neq y_{j_p}^{r-1}, \\ y_p^{r-1}, \quad \text{при } y_p^{r-1} = y_{j_p}^{r-1}, \ p \in [1, N_r], \end{cases}$$

де n_p, n_{j_p} – відповідно число точек в класах S_p^{r-1} та $S_{j_p}^{r-1}$.

При виконанні умов $S^r = S^{r-1}$ та $d_r = d_{r_n}$ допустимо, що $S^r = S$ і алгоритм закінчує роботу, інакше – переходимо до наступної ітерації $r+1$.

Лема. Нехай для даного кластера внутрішньокластерна відстань d дорівнює пороговому значенню d_{r_n}. Тоді при реалізації А-алгоритма сума відхилень від середньої точки e^* кластеру $d^r = \sum_{j=1}^{N_r} \|e^* - y_j^r\|$ строго зменшується : $d^r > d^{r+1}$ $(r=1,\ldots,r_n-1)$ 1, починаючи з деякого номера r_n, дорівнює нулю, при цьому $y_{r_n}^j$ єдиний елемент, рівний e^*.

Слідством доведеної леми являється теорема, що встановлює достатні умови збіжності до незміщеного розбивання $S^* = (S_1^*, \ldots, S_N^*)$.

Теорема. Нехай вибірка X допускає правильну кластеризацію $S^* = (S_1^*, \ldots, S_N^*)$ відносно множини $e(S^*) = (e_1(S^*), \ldots, e_N(S^*))$. де $e_j(S^*)$ – середній вектор кластеру S_j^*. Причому максимальне порогове значення d_{r_n} знаходиться в межах $d_{max} < d_{r_n} < D_{min}$. Тоді А-алгоритм за кінцеву кількість кроків r_n збігається до незміщеного розбивання S^*, а $y_{r_n}^j = e_j(S)$, $j \in [1,N]$ являються оцінками середніх, незалежно від вибору $d_0 \leqslant \ldots \leqslant d_{r_n}$.

Доведення засновано на збіжності оцінок $y_{r_n}^j$ до $e_j(S^*)$ в кожному класі. При цьому порогове значення d_{r_n} забезпечує неможливість створення кластерів із точок, що належать різним кластерам. Так як для всіх $x_{ij} \in S_j^*$ має місце $\|x_{ij} - e_j(S^*)\| \leqslant d_{r_n} < \|x_{ij} - e_k(S^*)\|$, $k \neq j$, то разбивання співпадає з мінімальним дистанційним разбиванням, і також являється незміщеним.

Якщо вибірка X не допускає правильної кластеризации, то А-алгоритм (аналогічно методам динамічних сгущень), збігається до кластеризації $S' = (S_1', \ldots, S_N')$, відносно якої задані "ядра" (центри ваги кластерів) являються найбільш предс авленими.

В даній евристичній постановці А-алгоритм може бути використаний для кластер-аналізу стану навколишнього середовища та природних ресурсів.

Системний підхід до організації досліджень в навколишньому природному середовищу потребує присутності в процесі моделювання у дослідника моделей об'єктів, отриманих в результаті виконання процедури ідентифікації. Цілі дослідження об'єктів обумовлюють означену ієрархію складності рішень, приймаємих на основі ідентифікуючих моделей. Тобто, і моделі, в свою чергу, повинні бути збудовані за ієрархічним принципом, рівні якого конструюються у відповідності до шкали складності, достатньої для прийняття відповідного рішення.

Використовуючи ієрархічний підхід теорії систем можно зробити висновок про те, що експертна система, яка служить для вибору адекватної математичної мови моделювання, повинна включати в себя деякий банк математичних моделей, маючих різну ступінь нечіткості. Це дозволить, в залежності від цілей дослідження, використовувати адекватний об'єкту математичний апарат. Важливим моментом при організації банку моделей є те, що моделі, які використовуються для описування нечітких, недоозначених об'єктів, можуть бути застосовані також для об'єктів, у яких неозначеність відсутня.

В Додатку представлено пакет програмних модулів (ППМ) "Индуктивні методи моделювання та кластер-аналізу нечітких об'єктів". Даний ППМ може складати основу ієрархічного банку моделей експертної системи екологічного моніторінгу навколишнього середовища.

Підвищення вимог до достовірності оцінок параметрів моделей об'єктів навколишнього середовища (наприклад, полів забруднення), швидкодія їх знаходження та точність идентифікації характеристик з однієї стороны, та перебороння труднощів реєстрації аномальних станів, що обусловлені необхідністю прийняття рішень в умовах неозначенності, з іншої сторони потребують підвищення ефективності автоматизованних систем наукових досліджень (АСНД). АСНД відіграє процес збору та обробки інформації про стан навколишнього

середовища на основі математичних моделей досліджуваного явища та спостережувальних за допомогою вимірювальних приладів окремих його якостей.

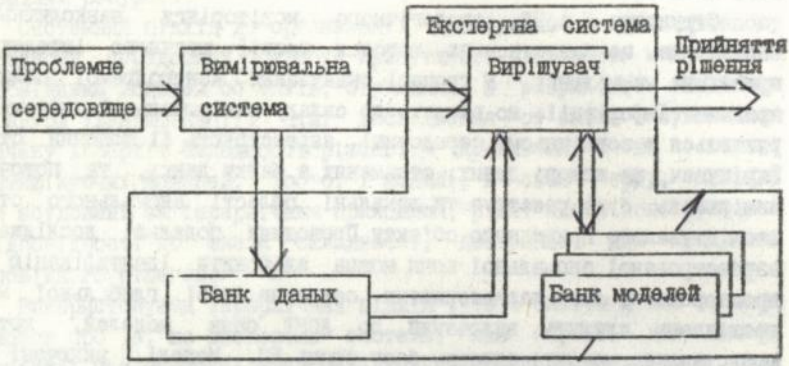
Структура АСНД екологічного моніторінгу навколишнього середовища, що базується на методах теорії штучного інтелекту, приведена на малюнку. В процесі сканування контролюємої області приймачі інформації, що входять до складу вимірювальної системи та рухаються в зовнішньому середовищі, зареєструють її фоновий стан. Вирішувач, на основі даних, отриманих з банку даних, та поточних вимірювань, буде реєструвати локальні області аномального стану досліджуваного природного об'єкту Проводячи подальші дослідження зареєстрованої аномальної зони можна здійснити ідентифікацію її просторово-часових характеристик. осягненню цієї глобальної мети досліджень служить включений до АСНД банк моделей, котрий представляє, по суті справи, базу знань ЕС. Моделі, включені до банку, розміщені за ієрархічним принципом в залежності від ступеня недооз.ченості об'єкту (таблиця). Кожна з моделей, що входить до банку моделей, повинна забезпечувати реалізацію в реальному часі оперативного рішення задач реєстрації з адаптацією до змін стану зовнішнього серед.овища за інформацією, що поступає з вимірювальної системи, а також повинна давати можливість отримати прогноз стану досліджуваного об'єкту на заданий інтервал час.у.

ОСНОВНІ РЕЗУЛЬТАТИ РОБОТИ.

Основним результатом диссертаційної р.боти являється створення нових та модифікація діючих параметричних і непараметричних методів моделювання та кластер-аналізу нечітких об'єктів природного середовища.

Наукові та практичні результати досліджень можна сформулювати у вигляді слідуючих висновків:

І. Пр.ведено аналіз дедуктивного та індукт.ного підходів до моделювання природних об'єктів, який показав, що в області екологічних задач системного аналізу, де об'єкти та їх модел'



Малюнок. Структура АСНД екологичного моніторингу навколишнього середовища

недооозначенi, найбiльша точнiсть розпiзнавання чи прогнозу досягається тодi, коли ступiнь нечiткостi моделi адекватна ступенi нечiткостi об'єкта. Адекватним математичним апаратом для описання нечiтких об'єктiв являється кластер-аналiз. Нечiткi переборнi методи моделювання дозволяють оптимизирувати ступiнь нечiткостi мови описання об'єкта у вигляді кластерiзацiї чи дискретизацiї вибiрки даних за вибраними зовнiшнiми критерiями.

2. В роботi поставлена i вирiшена задача переходу вiд бiнарних даних до неперервних, для чого розроблено алгоритм гармонiчної ребiнаризацiї вибiрки даних X. Запропонований алгоритм дозволяє достатньо точно вiдновити невiдому неперервну функцiю, задану у вигляді бiнарного коду. Алгоритм может бути розширений на випадок моделювання двумiрних полигармонiчних полiв. За допомогою запропонованого алгоритму здiйснено прогнозування сонячної активностi за середньорiчними значеннями чисел Вольфа.

3. Виконано порiвняння методiв оцiнки коефiцiентiв лiнiйної регресiйної моделi за МНК та методом iнструментальних змiнних (МIЗ). За допомогою численого моделювання показано, що похибка iдентифiкацiї оцiнок коефiцiентiв лiнiйної регресiї знаходиться в залежностi вiд величини показника обумовленостi матрицi нормальних рiвнянь X^TX. Показано, що МIЗ має перевагу при значних дисперсiях завад, а також у випадках, коли закон розподiлу завади вiдрiзняється вiд нормального. Запропоновано модифiкований МIЗ, котрий удосконалений за допомогою зменшення величини зсувiв, що досягається лiнiйною апроксимацiєю вибiрки даних.

4. Запропоновано iєрархiчний (агломеративний) алгоритм кластер-аналiзу, заснований на використаннi порогових значень. Приведено доведення збiжностi А-алгоритма за скiнчене число шагiв до правильної кластерiзацiї.

5. Представленi новi непараметричнi переборнi алгоритми, в котрих вiдсутня необхiдність оцiнки параметрiв. За допомогою алгоритма комплексування аналогiв вирiшена прикладна задача прогнозу мiсткостi кисню у водах Київського водосховища.

6. Методи iндуктивного моделювання i кластер-аналiзу розглянуто з позицiй системного пiдходу. Данi концептуальнi основи застосування методiв iндуктивного моделювання та кластер-аналiзу в

загальній задачі побудови системи моніторинга навколишнього природного середовища. Приведена структура автоматизованої системи наукових досліджень об'єктів природного середовища, яка заснована на ідеях нової інформаційної технології.

Основний зміст дисертації опублікований в слідуючих роботах:

I. Бідюк П.І., Жолнарський О.А. Огляд методів ідентифікації великих механічних конструкцій //Адаптивні САК .-1991.-№ 19.-С.25-31.

2. Жолнарський О.О., Литвинов Е.М. Рішення задачі екологічного моніторингу навколишнього середовища з позицій нової інформаційної технології //Доповіді АН України.-1991.- № 7.-С.169-172.

3. Жолнарський О.А., Жолнарська Т.С. Моделювання просторових даних в задачі обробки географічної інформації //Адаптивні САК. -1992.-№ 20.-С.37-42.

4. Івахненко О.Г., Жолнарський О.А. Оцінка коефіцієнтів поліномів в параметричних алгоритмах МГВА за покращеним методом інструментальних змінних //Автоматика.-1992.-№ 3.- С.25-33.

5. Івахненко Г.О., Жолнарський О.А., Івахненко Н.О. Синтез непараметричних моделей прогнозування випадкових багатомірних процесів та явищ //Електронне моделювання.-1992.-т.14.-№3.-С.45-50

6. Івахнечко О.Г., Жолнарський О.А., Пасько В.П., Дослідження правильності функціювання ієрархічного алгоритму кластер-аналізу //Статистичний та дискретний аналіз даних та експертне оцінювання: Тез. допов. наук. конф. (Одеса, 1991). Одеса, 1991.-С.191.

7. Івахненко О.Г., Жолнарський О.А., Мюлер І.А. Алгоритм гармонічної реінарізації вибірки даних //Автоматика.-1992. - №6.-С.26-35.

8. Ковальчук П.І., Жолнарський О.А. Збіжність ієрархічних алгоритмів кластер-аналізу // Доповіді АН України.-1991.-№11. - С.165-168.

9. Краскевич В.Б., Жолнарський О.А., Анурєєва А.Ю. Автоматизована система наукових досліджень для вирішення задач екологічного моніторінгу гидросфери //Сучасний стан теорії та розробок програмного забезпечення СК в ЕОМ: Тез. допов. наук. конф. (Самарканд, 1990). - М.: Гонти-6, 1990. - С.90-91.

Формат 60×90 1/16. Папір книжково-журнальний. Ум. друк. арк. 1,125.
Друк офсетний. Зам. № 3—218. Тираж 100 прим. 1993 р.
Друк, Укоопостачмашу. 252033, Київ, вул. І. Еренбурга, 5.