УДК 004.942

Т. В. Шулькевич[1], І. В. Баклан[1], О. В. Нестеренко[2], Ю. М. Селін[3]
[1]Національний технічний університет України «КПІ ім. Ігоря Сікорського»
Проспект Перемоги, 37, 03056 Київ, Україна
[2]Національна академія управління
вул. Ушинського, 15, 03151 Київ, Україна
[3]Інститут прикладного системного аналізу Національного технічного університету України «КПІ ім. Ігоря Сікорського»
Проспект Перемоги, 37, корп. 35, 03056 Київ, Україна

Математичний апарат інтелектуального аналізу даних для прогнозування нелінійних нестаціонарних процесів

Викладено математичний апарат, який можна застосовувати у задачах аналізу даних різної природи задля прогнозування нелінійних нестаціонарних процесів.

Ключові слова: *нелінійні нестаціонарні процеси, прогнозування, екологічні процеси, економічні процеси.*

Вступ

Сучасний розвиток інформаційних технологій, їхнє розповсюдження в усьому світі та можливості опрацювання значних об'ємів інформації дозволили створювати системи підтримки прийняття рішень на макрорівні з урахуванням комплексності розгляду проблем. Так, наприклад, стала зрозумілою глобальна небезпека екстенсивного шляху розвитку світової економіки, що призводить, у тому числі, до деградації довкілля і, як наслідок, до відповідного погіршення якості життя суспільства. 2009 року Нобелевські лауреати Джозеф Стігліц та Амартія Сен оприлюднили доповідь, у якій обґрунтували використання показника якості життя як основного критерію економічного розвитку суспільства замість ВВП [1]. Поняття «якість життя» визначають як узагальнюючу соціально-економічну категорію [2].

Ключова думка доповіді — це необхідність впоратися відразу з двома кризами: економічною і екологічною, які суттєво впливають на якість життя, і, як наслідок, ставлять перед нами питання: чи дає наявна статистика та методи її опрацювання правильні сигнали, що дозволяють приймати потрібні сьогодні ефективні рішення?

Аналіз останніх публікацій [3–6] свідчить виключно про філософський напрямок розвитку поняття якості життя, про якісні оцінки його рівня, але ніяк не спроби кількісно порахувати деякі його складові.

Необхідно зазначити, що до цього часу і екологічна наука, так само як і економічна, вже активно використовували розвинений математичний апарат і належні технології для прогнозування розвитку відповідних процесів, у тому числі й методи інтелектуального аналізу даних.

Інтелектуальні методи аналізу даних за визначенням — це процес пошуку у «сирих» даних раніше не відомих, нетривіальних, практично корисних і доступних інтерпретацій знань, які необхідні для прийняття рішень у різних сферах людської діяльності [7]. Інтелектуальний аналіз даних (ІАД) — термін, що застосовується для опису здобуття знань у базах даних, дослідження даних, обробки зразків даних, очищення та збору даних. Це процес виявлення кореляції, тенденцій, шаблонів, зв'язків і категорій [8]. Інтелектуальний аналіз даних розвивається на базі таких наук як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних тощо.

Процес автоматичного пошуку прихованих закономірностей або взаємозв'язків між змінними в інтелектуальному аналізі даних поділяється на задачі класифікації, моделювання та прогнозування з використанням статистичних і математичних методів. Різноманітні дані в екології, економіці, соціології та інших сферах надходять у вигляді часових рядів. Вони, як правило, є нестаціонарними, оскільки їхні основні характеристики змінюються у часі. Основою для прогнозування служить історична інформація, яка зберігається в інформаційних сховищах у вигляді часових рядів. Якщо можна побудувати математичну модель і знайти шаблони, що адекватно відбивають цю динаміку, є ймовірність, що за їхньою допомогою можна передбачати і поведінку системи в майбутньому. Прогнозування часових послідовностей дозволяє на основі аналізу поведінки часових рядів оцінити майбутні значення прогнозованих змінних.

Але проблема прогнозування нелінійних нестаціонарних часових рядів процесів парадигми якості життя є проблемою міждисциплінарною. Оскільки, не дивлячись на достатньо розвинений математичний апарат аналізу та прогнозування окремих складових, можна констатувати наступне. Математичні апарати є несхожими, тож фахівці-економісти не володіють засобами обробки екологічних даних і навпаки, екологи не використовують математичний апарат прогнозування економічних показників. Тож, можна констатувати відсутність математичного апарату, який можуть спільно використовувати фахівці різних напрямків.

Метою статті є розробка математичного апарату для аналізу та прогнозування нелінійних нестаціонарних процесів різної природи, що складають парадигму якості життя.

Складні задачі на макрорівні, такі як задачі прогнозування якості життя, характеризуються аналізом і прогнозуванням нелінійних нестаціонарних процесів. У зв'язку з цим пропонуються до розгляду деякі методи, які можна застосовувати для використання в подібних задачах. До них можна віднести наступні:

— приховані марківські моделі;

— метод подібних траєкторій;

— лінгвістичне моделювання.

Основна частина

Метод прихованих марківських моделей

Традиційно приховані марківські моделі (ПММ) визначають як трійку

$$\lambda = (A, B, \Pi),$$

де:

1) $A = \{a_{ij}\}$ — матриця ймовірностей переходів зі стану S_j до стану S_i, $a_{ij} = P\left[q_{i+1} = S_j \mid q_i = S_i \right], 1 \le i, j \le N$;

2) $B - \{b_j(k)\}$ — розподіл імовірностей спостережуваних символів у стані j, де $b_j(k) = P\left[v_k \mid q_i = S_j \right]$, $1 \le j \le N$, $1 \le k \le M$ (для безперервного випадку $b_j(k)$), задається як функція розподілу щільності ймовірності);

3) $\Pi = \{\Pi_i\}$ — імовірність кожного початкового стану.

Загальну схему функціонування прихованої марківської моделі зображено на рис. 1.

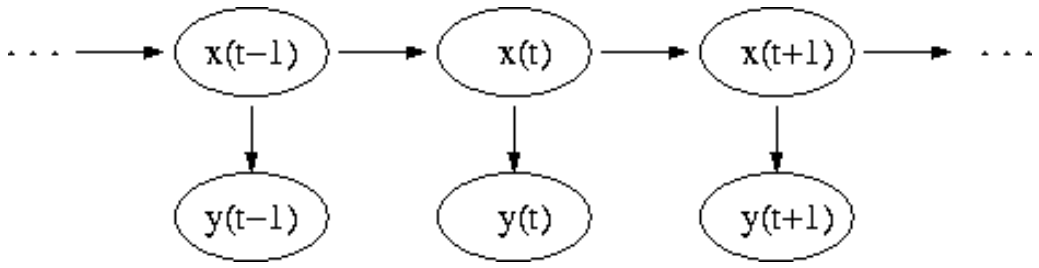

Рис. 1. Загальна схема функціонування прихованої марківської моделі

Основні завдання при застосуванні ПММ до визначення параметрів процесів полягають у наступному. Для використання ПММ при розпізнаванні мови необхідно вирішити три задачі [9].

Задача 1. Якщо задані послідовність спостережень і модель $\lambda = (A, B, \Pi)$, то як ефективно обчислити $P(O \mid \lambda)$ — вірогідність такої послідовності при заданих параметрах моделі?

Задача 2. Якщо задані послідовність спостережень і модель $\lambda = (A, B, \Pi)$, то як визначити відповідну послідовність внутрішніх станів?

Задача 3. Якщо задані послідовність спостережень, то як визначити параметри моделі $\lambda = (A, B, \Pi)$, виходячи з критерію максимізації $P(O \mid \lambda)$?

Очевидно, що розв'язання першої задачі дозволить ефективно розпізнавати елементи з деякого обмеженого набору, маючи готові, вже навчені моделі цих елементів. У цьому випадку наявний елемент дискретизується та після сегментації на певні примітиви, розглядається як набір послідовностей спостережень. Замість безпосередньо графічного сигналу системи зазвичай працюють з якимось набором примітивів (меншим за об'ємом), що характеризує наявний сигнал. Цей етап називається передобробкою сигналу. Потім кожна послідовність

спостережень, що є невідомим словом, перевіряється на відповідність усім наявним у словнику моделям слів. Для цього обчислюється вірогідність генерації такої послідовності спостережень кожній з наявних у словнику моделей. Слово, модель якого з найбільшою вірогідністю генерує таку послідовність спостережень, і є результатом розпізнавання.

Рішення другої задачі дозволяє виявляти «приховану» частину моделі, тобто знаходити її внутрішні стани. Насправді визначити ту послідовність станів, яка мала місце, неможливо, а у випадку з визначанням такої послідовності просто немає, оскільки вважається, що модель мовоутворення людини не носить стохастичний характер. Проте, знаходження оптимальної послідовності дає нам додатковий критерій порівняння при пошуку оптимальної моделі та дозволяє виявляти статистичні характеристики конкретних станів моделі.

Рішення третього завдання дозволяє навчати моделі, тобто обчислювати параметри моделі так, щоб вона найкращим чином описувала так звані «тренувальні» послідовності. Завдання тренування моделей вважається найскладнішим при використанні ПММ у системах визначення параметрів процесу, оскільки невідомо єдиного і універсального способу її рішення, а від результату навчання моделей залежить якість роботи системи розпізнавання.

Наведемо підходи до розв'язання першої задачі ПММ, а саме: ефективне обчислення ймовірності генерації заданої послідовності $P(O_j)$.

Як уже вказувалося, це завдання пов'язане з необхідністю обчислити вірогідність послідовності спостережень при заданих параметрах моделі λ, тобто $P(O \mid \lambda)$. Прямий метод обчислення цієї вірогідності полягає в обчисленні суми вірогідності всіх можливих послідовностей станів. Розглянемо одну з таких можливих послідовностей $O_t, t = 1, T$.

Імовірність послідовності спостережень $P(O \mid Q, \lambda)$ при заданій послідовності станів обчислюється як

$$P(O \mid Q, \lambda) = \prod_{i=q}^{N} P(O_i \mid q_i, \lambda)$$

або

$$P(O \mid Q, \lambda) = b_{q_1}(O_1) * b_{q_2}(O_2) * \ldots * b_{q_T}(O_T).$$

Імовірність такої послідовності станів може бути записана як

$$P(Q \mid \lambda) = \Pi_{q_1} * a_{q_1 q_2} * a_{q_2 q_3} * \ldots * a_{q_{T-1} q_T}.$$

Сумісна вірогідність $P(O|\lambda)$ обчислюється як добуток приведеної вище вірогідності:

$$P(O \mid Q, \lambda) = P(O \mid Q, \lambda) P(Q \mid \lambda).$$

Таким чином, вірогідність $P(O \mid Q, \lambda)$ обчислюється як сума сумісної імовірності за всіма можливими послідовностями станів q:

$$P(O|\lambda) = \sum_{aiiQ} P(O|Q,\lambda)P(Q|\lambda) = \sum_{q_1,q_2,\ldots,q_T} \Pi_{q_1} * b_{q_1}(O_1) * a_{q_1q_2} * b_{q_2}(O_2) * \ldots * a_{q_{T-1}q_T} b_{q_T}(Q_T).$$

Неважко підрахувати, що кількість операцій множення, які необхідні для обчислення цієї суми, дорівнює $(2T-1)N^T$. Тобто, якщо модель має п'ять станів ($N = 5$), і послідовність спостережень має довжину сто ($T = 100$), то кількість арифметичних операцій дорівнює $2 \cdot 100 \cdot 5^{100} \approx 10^{72}$.

Проте існує ефективніший метод обчислення вірогідності P. Він називається процедурою Forward-Backward і полягає в наступному: вводиться змінна, яка визначається як

$$\alpha_i(i) = P(O_1, O_2, ..., O_i, q_i = S_i | \lambda).$$

Назвемо її «прямою» змінною, яка є ймовірністю появи для даної моделі часткової послідовності спостережень $O_1, O_2, ..., O_i$ (до моменту часу t і стану S_i) при заданих параметрах моделі λ. Можна індуктивно визначити цю пряму змінну як:

1) ініціалізація:

$$\alpha_1(i) = \prod_i b_i(O_1), \ 1 \le i \le N;$$

2) індукція:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_1(i) a_{ij} \right] b_i(O_{t+1}), \ 1 \le t \le T-1, \ 1 \le j \le N;$$

3) завершення:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

При цьому крок 1 ініціалізував прямі змінні сумісною ймовірністю станів і початкових спостережень. Індуктивний крок 2 є серцем процедури, він проілюстрований на рис. 2.

Рисунок показує, як стан S_i може бути досягнутий у момент часу $t+1$ з N можливих станів $S_i, 1 \le i \le N$, в яких система могла знаходитись у момент часу t. Оскільки $\alpha_t(i)$ — це спільна ймовірність того, що спостерігалася послідовність $O_1, O_2, ..., O_i$, і у момент часу t система знаходилась у стані S_i, то множення $\alpha_t(i)a_{ij}$ — це спільна ймовірність того, що спостерігалася послідовність $O_1, O_2, ..., O_i$, і в момент

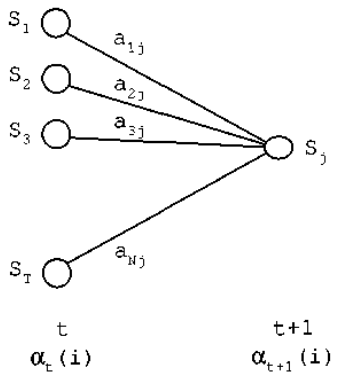

Рис. 2. Індуктивний крок процедури Forward-Backward

часу $t+1$ система знаходилась у стані S_j, до якого перейшла зі стану S_i, в якому знаходилась у момент часу t. Підсумовуючи ці результати за всіма N можливими станами S_i, $1 \le i \le N$, в яких система могла знаходитись у момент часу t, отримаємо вірогідність переходу системи у момент часу $t+1$ до стану S_{ji} з урахуванням усіх можливих передуючих цьому часткових послідовностей спостережень. Тепер, помноживши цю суму на вірогідність $b_j(O_{t+1})$, просуваємося на один крок, отримавши $\alpha_T(i)$.

На завершальному кроці 3 обчислюється шукана вірогідність $P(O \mid \lambda)$ як сума за *i* остаточними значеннями прямих змінних $\alpha_T(i)$, при цьому

$$\alpha_T(i) = P(O_1, O_2, ..., O_T, q_T = S_i \mid \lambda).$$

Таким чином, для обчислення ймовірності $P(O \mid \lambda)$ потрібно вже $O(N^2T)$ обчислених операцій — замість $O(TN^T)$ для прямого методу, так що даний метод набагато швидший. У нашому прикладі для $N = 5$ і $T = 100$ він вимагає всього біля 3000 арифметичних операцій, замість 10^{72} операцій для прямого обчислення.

Тепер введемо «зворотну» змінну $\beta_t(i)$ і визначимо її як

$$\beta_t(i) = P(O_1, O_2, ..., O_T,\ q_t = S_i \mid \lambda),$$

тобто $\beta_t(i)$ — це вірогідність часткової послідовності спостережень від моменту часу t до кінця послідовності при заданому стані S_i у момент часу t і параметрах моделі λ.

$\beta_t(i)$ ми точно також можемо індуктивно обчислити за допомогою наступних процедур:

1) ініціалізації:

$$\beta_T(i) = 1,\ \ 1 \le i \le N;$$

2) індукції:

$$\beta_T(i) = \left[\sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\right] \beta_{t+1}(j),$$
$$t = T-1,\ T-2, ..., 1;\ \ 1 \le i \le N.$$

Очевидно, що для вирішення першого завдання ПММ достатньо обчислення тільки «прямої» змінної, або тільки «зворотної».

Метод «подібних траєкторій»

Однією з проблем прогнозування часової послідовності є те, що існує можливість збільшення вимірів даних за рахунок зменшення періоду дискретизації чи за допомогою інтерполяції; тобто можна отримати більше даних, не додаючи при цьому нової інформації. Це є проблемою, оскільки у дискретизованого сигналу з

надто великою частотою всі найближчі траєкторії будуть знаходитись у часовій послідовності поруч один до одного.

Опишемо алгоритм пошуку найближчих траєкторій за наявності алгоритму пошуку найближчих точок [10]. Нехай знайдено точку x_i, яка знаходиться ближче за k-ту найближчу точку.

Підрахуємо відстані до попередніх точок послідовно від точки x_i ($x_{i-1}, x_{i-2}, ...$) для пошуку найближчого локального мінімуму. Повторимо цю процедуру для точок, що йдуть після x_i ($x_{i+1}, x_{i+2}, ...$). Локальний мінімум позначимо за $x_{\min}$. Це буде найближча точка в даному сегменті траєкторії.

Виключимо інші точки цього сегменту з подальшого розгляду. Для цього підрахуємо відстані до попередніх точок послідовно від точки x_i ($x_{i-1}, x_{i-2}, ...$), поки не досягнемо локального максимуму, або відстань перевищить відстань до k-ї найближчої точки, знайденої раніше. Позначимо цю точку за $x_{\max}$ і виключимо з подальшого розгляду точки, що знаходяться між $x_{\min}$ та $x_{\max}$.

Повторимо попередній крок для точок, що йдуть після $x_{\min}$. Замінимо k-ту найближчу точку точкою $x_{\min}$ і продовжимо пошук найближчих точок.

Дамо графічну інтерпретацію методу подібних «траєкторій». Ідея методу полягає в наступному. Маємо ряд спостережень екологічного процесу, що їх зроблено за якийсь час $\{y(1), y(2), ..., y(n)\}$, графік якого наведено на рис. 3.

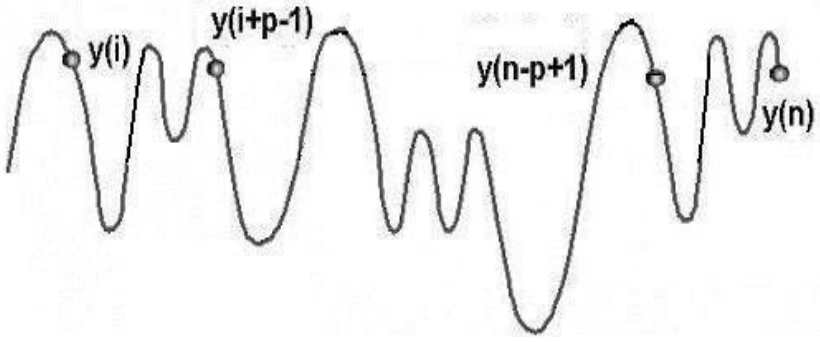

Рис. 3. Динаміка ряду спостережень

Змінна $y(i)$, $i = \overline{1, N}$, тут представлена фізичними значеннями відповідного процесу (наприклад, сила вітру, інтенсивність стоку води, сила підземних поштовхів).

За обраним критерієм обирається ділянка траєкторії, яка є «найближчою» до ділянки, що передує прогнозованій точці. Надалі оцінюється прогноз за формулою $y(n+1) = y(i+p)$, де:

$$I = \min\left\{\sum_{i=1}^{p} | y(j+i-1) - y(n-p+i) |\right\}, \quad J = 1, 2, ..., n-p,$$

$$J = \min_i \ | y(i+j-1) - y(n) |, \ \ i = I, I+1, \ldots, I+p-1.$$

Формалізувати метод можна наступним чином. Нехай ми маємо вектори спостережень:

$$Y_1 = (y_1, y_2, ...y_p)^T, Y_2 = (y_3, y_3, ...y_{p+1})^T, \ldots, Y_K = (y_k, y_{k+1}, ...y_{k+p+1})^T, \ldots,$$
$$Y_N = (y_{n-p+1}, y_{n-p+2}, ...y_n)^T.$$

Знаходимо найближчу точку з умови мінімальної відстані:

$$Y_k = \arg\min_j d(Y_n, Y_j).$$

Є й інші способи пошуку найближчої точки, наприклад, найбільш поширена метрика — квадрат евклідової відстані:

$$d(Y_k, Y_n) = (Y_k - Y_n)^T (Y_k - Y_n).$$

Метод лінгвістичного моделювання

Побудова лінгвістичної моделі. Для досягнення поставленої мети має бути розв'язаною задача знаходження лінгвістичного образу часового ряду, до якого входять:

а) обчислення різницевих рядів вихідного часового ряду;

б) вибір критерію інтервалізації різницевих рядів;

в) інтервалізація певного різницевого ряду згідно обраного критерію;

г) знаходження лінгвістичного ланцюжка для певного різницевого ряду;

д) знаходження матриці переходів для кожної можливої пари символів у лінгвістичному ланцюжку певного різницевого ряду.

Вхідними даними для даної задачі є значення часового ряду.

Вихідними даними для цієї задачі є лінгвістичний образ часового ряду (динамічного процесу), що являє собою:

— множину інтервалів, яку отримано в результаті інтервалізації різницевого ряду певного порядку від часового ряду;

— матрицю переходів (передування), яку побудовано на множині інтервалів (описаній вище) та за часовим рядом.

Указаний лінгвістичний образ будується окремо для різницевих рядів, вхідного часового ряду, різних порядків [11–13]. Таким чином, отримуємо множину лінгвістичних образів, що і є проміжним результатом задачі прогнозування з використанням лінгвістичного моделювання.

Надалі буде розглядатися підхід лінгвістичного моделювання для побудови лінгвістичного образу вхідного часового ряду.

Одним із підходів є використання методів розпізнавання образів для реалізації процедур прогнозування. Тому зупинимося як раз саме на розпізнаванні образів. Більшість різноманітних математичних методів розв'язання задач розпізнавання образів можна поділити на два основних класи.

Перший можна позиціонувати з теорією рішень, його ще називають дискримінантним підходом. У цьому випадку об'єкти характеризуються наборами чисел — результатами деякої множини вимірів, що називають ознаками. Розпізнавання

образів при застосуванні цього підходу зазвичай роблять за допомогою розбиття простору ознак на області [14].

Другий клас розвивається у межах синтаксичного (або структурного) підходу. Особливостями цього підходу є розпізнавання образів, в яких міститься важлива інформація про структуру образу, а від самої процедури розпізнавання вимагається, щоб вона давала можливість не тільки віднести об'єкт до деякого класу (тобто визначити його класифікацію), але й дати опис тих сторін об'єкта, які виключають можливість його віднесення до іншого класу.

Коли об'єкти складні та кількість можливих описів велика, незручно рахувати, що кожний опис визначає клас. Таку ситуацію маємо у задачах ідентифікації зображень, відбитків пальців. Таку саму картину спостерігаємо при розпізнаванні динамічних образів. У цих випадках розпізнавання може бути проведеним із використанням опису кожного об'єкта, а не просто за допомогою методів класифікації.

Для того, щоб представити ієрархічну структурну інформацію, яка є у кожному образі, використовуємо прості часткові образи. Цей підхід базується на аналогії між структурою образів і синтаксисом мов. У межах синтаксичного підходу вважається, що образи будуються з поєднаних різноманітним чином часткових образів, так само як фрази та речення формуються шляхом приєднання слів, а слова складаються із літер. Їх ще називають непохідними елементами.

Легко бачити, що такий підхід є корисним у таких випадках, коли розпізнавати обрані прості часткові образи легше, ніж самі образи. «Мову», яка забезпечує структурний опис образів у термінах непохідних елементів та операцій композиції цих елементів, називають мовою розпізнавання образів. Правила композиції непохідних елементів традиційно задають за допомогою так званої граматики мови опису образів. Процес розпізнавання образів здійснюється після ідентифікації в об'єкті непохідних елементів і складення опису об'єкта. Безпосередньо розпізнавання полягає у синтаксичному аналізі (або граматичному розбиранні) «речення», яке описує об'єкт. Ця процедура виявляє, чи є це речення синтаксично вірним по відношенню до заданої граматики. Паралельно синтаксичний аналіз дає деякий структурний опис речення.

Синтаксичний підхід до розпізнавання образів дає можливість описувати достатньо велику множину складних об'єктів шляхом використання невеликої множини непохідних елементів і граматичних правил. І в цьому нам буде допомагати рекурсивна природа апарату граматик.

Граматичне правило (або правило підстановки) може бути застосованим будь-яку кількість разів, так що виявляється можливим достатньо компактним чином подати деякі структурні характеристики нескінченної множини речень. Практичний зиск цього підходу залежить від здатності розпізнавати непохідні елементи образів і їхні взаємовідносини, що подані у вигляді операцій компонування.

Різноманітні відношення, що визначені між частковими образами, традиційно можуть бути подані логічними та математичними операціями.

Обробка символьних послідовностей ставить певні проблеми. Символи групуються у слова, слова складають речення, але не вільним чином, а за певними правилами. Щоб виявити закономірності у розташуванні в реченні, необхідно ви-

значатися з поданням, у межах якого можна було б ці закономірності не тільки описувати, але й віднаходити у символьних послідовностях.

Ці питання є основними для вивчення лінгвістичних особливостей символьних послідовностей. Саме щодо лінгвістичних вимог, свого часу Наомом Хомським у середині минулого століття було запропоновано теорію формальних граматик, яка й стала одним із основних розділів математичної лінгвістики.

Згідно з етапами побудови лінгвістичної моделі вихідна задача буде розбита на такі підзадачі:

— підзадачу отримання різницевих рядів;
— підзадачу інтервалізації;
— підзадачу лінгвістизації;
— підзадачу побудови матриці переходів.

Підзадача отримання різницевих рядів. Призначенням даної підзадачі є отримання рядів, котрі характеризують динаміку зміни руху курсору «мишки»: швидкість (різницевий ряд 1-го порядку), прискорення (різницевий ряд 2-го порядку) тощо. Таким чином, різницеві ряди являються похідними від вихідного ряду.

Дано: вектор цілих чисел $\overline{X}$ з потужністю $n = \left|\overline{X}\right|$.

Результати: вектор цілих чисел $\overline{D}$ з потужністю $k = \left|\overline{D}\right|$.

Обмеження:

$$\forall d_i \in \overline{D}: \quad d_i = x_{i+1} - x_i ,$$

де $i \in [0; n-1); x_{i+1}, x_i \in \overline{X}; k = n-1$.

Підзадача інтервалізації. Призначенням даної підзадачі є побудова алфавіту користувача шляхом розбиття відсортованого різницевого ряду на множину інтервалів, кожний елемент якої характеризує певну літеру алфавіту.

Дано:

— гіпотетичну потужність алфавіту *a*;

— вектор цілих чисел $\overline{D}$ з потужністю $k = \left|\overline{D}\right|$.

Результати: вектор пар цілих значень $\overline{I}$ з потужністю $k = \left|\overline{I}\right|$.

Обмеження:

$$\forall x \in \overline{I} : x^1 \le x^2 , \tag{1}$$

$$\forall x_i, x_{i+a} \in \overline{I} : x_i{}^2 \le x^1{}_{i+1} , \tag{2}$$

де $i \in [0; n-1)$, $n \le a$, $a << k$;

$$\exists x \in \overline{I} : \forall d \in \overline{D}, \quad d \in \left[x^1; x^2 \right], \tag{3}$$

$$\forall d_i; d_{i+1} \in \overline{D} : d_i \le d_{i+1} , \tag{4}$$

де $$i \in [0; k-1),$$

$$x_0 \in \overline{I} : x_0 \in \left(-\infty; x_1^1\right),$$

$$x_n \in \overline{I} : x_n \in \left(x_{n-1}^2; +\infty\right). \quad (5)$$

Підзадача лінгвістизації. Призначенням даної підзадачі є отримання лінгвістичного ланцюжка шляхом знаходження відповідної літери алфавіту для кожного значення різницевого ряду. Літера алфавіту однозначно відповідає певному інтервалу із множини інтервалів, що отримані в результаті розв'язання попередньої задачі (рис. 4).

Дано:

— вектор цілих чисел $\overline{D}$ з потужністю $k = \left|\overline{D}\right|$, що відповідає обмеженню, представленому у формулі (3);

— вектор пар цілих значень $\overline{I}$ з потужністю $n = \left|\overline{I}\right|$ з обмеженнями, що наведені у формулах (1) та (3), а також у (4) та (5).

Результати: вектор цілих чисел $\overline{A}$ з потужністю $k = \left|\overline{A}\right|$.

Обмеження:

$$\forall x_i \in \overline{A} : \exists d_i \in \overline{D}, \exists y_i \in \overline{I}, d_i \in \left[y_j^1; y_h^2\right], x_i = j, \quad (6)$$

де $i \in [0; k); j \in [0; n)$.

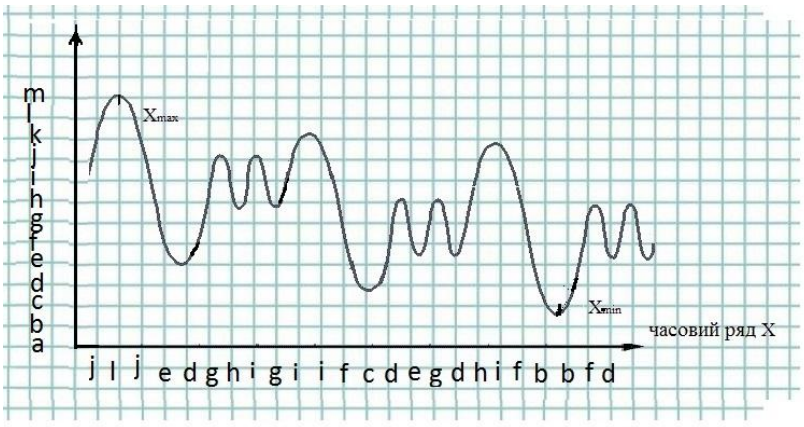

Рис. 4. Загальна схема переходу до лінгвістичного ланцюжка

Підзадача побудови матриці переходів. Призначенням даної підзадачі є побудова матриці переходів між двома літерами алфавіту в реченні. Алфавіт і його літери визначені у підзадачі інтервалізації, а речення — у задачі лінгвістизації.

Дано:

— вектор цілих чисел $\overline{A}$, що відповідає обмеженню 6 з потужністю $k = \left|\overline{A}\right|$;

— потужність множини інтервалів n, що отримана в результаті розв'язання підзадачі інтервалізації.

Результати: квадратна матриця раціональних чисел $\overline{P}$ розмірністю n.

Обмеження:

$$\forall x_{ij} \in \overline{P} : x_{ij} \in [0,0;1,0].$$

У разі наявності інформації, яку отримано у вигляді графічного зображення, маємо наступне. За допомогою ланцюжкового коду Фримена (рис. 5) переходимо до символьного запису послідовності даних.

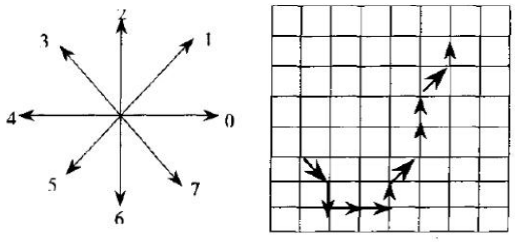

Рис. 5. Перетворення графічних даних до символьного вигляду

Згідно з рис. 5, ми отримуємо послідовність 7, 0, 0, 2, 1, 2, 2, 1, 2, яку так само можна замінити на символьний ряд з довжиною абетки $N = 8$.

Отриману послідовність аналізують на наявність граматичних конструкцій. На виході отримуємо список граматичних конструкцій з імовірностями їхньої наявності в процесі, а також матрицю ймовірностей переходу із символу до символу. Цей етап тісно перекликається з моделюванням (прихованих) марківських процесів, а також з методом подібних траєкторій.

Висновки

Наведено математичний апарат, що об'єднує в собі три види подачі інформації задля її аналізу та прогнозування. Перший вид — звичний числовий, що зустрічається чи не в 99 % відомих авторам джерел, другий — графічний, що ще й досі отримується з аналогових приладів реєстрації і третій — символьний (метод лінгвістичного моделювання), що ще не є розповсюдженим. Для методу лінгвістичного моделювання сформульовано змістовну та математичну постановки задачі знаходження лінгвістичних образів часових рядів. Уведено поняття лінгвістичного моделювання та поетапно описано спосіб застосування даного підходу для розв'язання поставленої задач. Також, для більш повного розгляду задачі, введено додаткові підзадачі, описано їхні математичні постановки та наведено приклади їхнього виконання.

Кожен метод має свої переваги, недоліки та сфери застосування. Викладення докладного їхнього дослідження та порівняння є предметом подальших публікацій авторів.

Наведено правила переходу всіх трьох видів отримання інформації одне в одне.

Наведені методи є універсальними як з боку виду отриманої інформації так і з боку наявності у цій інформації нелінійностей та нестаціонарностей. Але мають загальний недолік усіх статистичних методів — брак історичної інформації.

1. Documents du site de la «Commission sur la Mesure de la Performance Économique et du Progrès Social». URL: http://www.insee.fr/fr/publications-et-services

2. Райзберг Б.А., Лозовский Л.Ш., Стародубцева Е.Б. Современный экономический словарь. 2-е изд., испр. Москва: ИНФРА-М, 1999. 479 с.

3. Barcaccia B. Quality Of Life: Everyone Wants It, But What Is It? Forbes. Education, 10 May 2016.

4. World Happiness Report. Overview. Helliwell, J., Layard, R., & Sachs, J. World Happiness Report 2016, Update (Vol. I). New York: Sustainable Development Solutions Network, 2016.

5. Health and Happiness. The Lancet. 387: (1251). 26 March 2016.

6. Van der Krieke et al. Temporal Dynamics of Health and Well-Being: A Crowdsourcing Approach to Momentary Assessments and Automated Generation of Personalized Feedback. *Psychosomatic Medicine*. 2016. **1**.

7. Fayyad U.M. Piatetsky-Shapiro G., Smyth P. From Data Mining to KnowledgeDiscovery in Databases. *AI Magazine*. 1996. N 17(3). P. 37–54

8. Плескач В.Л., Затонацька Т.Г. Інформаційні системи і технології на підприємствах: підручник. — Київ: Знання, 2011. 718 с.

9. Rabiner L.R. A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE. Feb. 1989. Vol. 77. N 2. P. 257–284.

10. Kollios G., Gunopulos D., Tsotras V.J. Nearest neighbor queries in a mobile environment. *In Spatio-Temporal Database Management*. 1999. P. 119–134.

11. Баклан И.В., Селин Ю.Н. Анализ поведения экономических часовых рядов с использованием структурных подходов. *Вісник Херсонського національного технічного університету*. Херсон: ХНТУ, 2006. № 2. С. 29–34.

12. Баклан І.В., Селін Ю.М., Петренко О.О. Структурний підхід до розпізнавання образів у системах безпеки. Національна безпека України: стан, кризові явища та шляхи їх подолання: зб. наук. праць за матеріалами міжнар. наук.-прак. конф. (7–8 груд. 2005, м. Київ). Київ: НАУ-ЦНПСД, 2005. С. 375–380.

13. Баклан І.В. Лінгвістичне моделювання: основи, методи, деякі прикладні аспекти. *Систем. технології*. 2011. № 3. С. 10–19.

14. Fu K.S. Sequential Methods in Pattern Recognition and Machine Learning. Academic Press, 2012.

Надійшла до редакції 24.01.2017