Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

БАБІЙ АНДРІЙ СТЕПАНОВИЧ

УДК 004.942

МОДЕЛІ, МЕТОДИ ТА ІНТЕЛЕКТУАЛЬНА ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ НЕОДНОРІДНИХ ПОСЛІДОВНОСТЕЙ

05.13.06 – інформаційні технології

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків-2017

Дисертацією є рукопис.

Робота виконана в Харківському національному університеті радіоелектроніки Міністерства освіти і науки України.

Науковий керівник - доктор технічних наук, професор
Єрохін Андрій Леонідович,
Харківський національний університет радіоелектроніки,
декан факультету комп`ютерних наук

Офіційні опоненти: доктор технічних наук, професор
Шаронова Наталія Валеріївна,
Національний технічний університет
«Харківський політехнічний інститут»,
завідувач кафедри інтелектуальних комп'ютерних систем

доктор технічних наук, професор
Дивак Микола Петрович,
Тернопільський національний економічний університет,
декан факультету комп'ютерних інформаційних технологій

Захист відбудеться "15" грудня 2017 р. о 14^{00} годині на засіданні спеціалізованої вченої ради Д 64.052.08 у Харківському національному університеті радіоелектроніки за адресою: 61166, м. Харків, пр. Науки, 14.

З дисертацією можна ознайомитись у бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Науки, 14.

Автореферат розісланий "10" листопада 2017 р.

Учений секретар
спеціалізованої вченої ради І.П. Плісс

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми.

Зміни, що відбуваються в сучасному суспільстві, спричинені різними факторами, в тому числі і швидким зростанням рівня проникнення інформаційних технологій в різні види діяльності людини.

Існуючі інформаційно-аналітичні системи та інструментальні засоби інтелектуального аналізу даних такі як IBM Cognos Analytics, Microsoft SRSS, MicroStrategy Analytics Platform, Oracle BI, SAP Business Warehouse, дозволяють здійснювати узагальнення великих масивів даних та формування інформації, на основі якої можуть прийматися рішення і оцінюватися поточний стан предметної області із застосуванням широкого діапазону моделей і методів обробки даних.

Разом з тим часто постає проблема, не охоплена існуючими реалізаціями інформаційно-аналітичних систем та інструментальних засобів інтелектуального аналізу даних, а саме обробка різнорідних даних, частина з яких виражена у вигляді чітких кількісних значень, отриманих в результаті вимірювань, а друга частина має суб'єктивну та нечітку природу.

Така задача може виникати, наприклад, в сфері наук про людину, її особисту поведінку, умови та процеси у суспільстві. Ці відомості можуть бути формалізовані у вигляді лінгвістичних змінних або нечітких множин чи інтервальних оцінок.

Дослідженням обробки даних різнорідної природи походження, узгодженню експертних оцінок, основним положенням теорії нечітких множин та її застосуванню для вирішення різних наукових задач були присвячені роботи таких вчених: Л. Заде, Т. Сааті, Е. Руспіні, І. Перфильєвої, Г. Танаки, П. Курран, Л. Г. Раскіна, Є. В. Бодянського та інших.

Сучасні інформаційно-аналітичні системи, як правило, мають інтерфейси для створення додаткових користувацьких модулів, що дає можливість розширювати функціональність існуючих, в тому числі і користувацьких, інформаційних систем, розроблених на основі поширених платформ інтелектуального аналізу даних.

Враховуючи невідповідність між можливостями існуючих інформаційних технологій обробки відомостей, представлених у вигляді неоднорідних послідовностей даних і розповсюдженості потреби в їх аналізі, виникає завдання відбору значущих чинників при проведенні нечіткого багатофакторного регресійного аналізу і завдання фільтрації неоднорідних часових послідовностей даних.

Таким чином, розробка моделей, методів та інтелектуальної інформаційної технології аналізу неоднорідних послідовностей є актуальною науково-практичною задачею.

Зв'язок роботи з науковими програмами, планами, темами.

Дисертаційна робота виконана на кафедрі програмної інженерії Харківського національного університету радіоелектроніки відповідно до завдань НДР «Теорія, методи і моделі управління життєвим циклом інтелектуальних інформаційних

середовищ регіональних соціо-економічних об’єктів» (№ ДР 0115U002430), де здобувач брав участь як виконавець окремих розділів.

Мета та задачі дослідження.

Метою дисертаційної роботи є розробка моделей, методів та інтелектуальної інформаційної технології аналізу неоднорідних послідовностей даних для підвищення ефективності оцінювання поточного стану предметних областей в інформаційно-аналітичних системах.

Для досягнення поставленої мети необхідно здійснити вирішення таких задач:

- провести дослідження існуючих моделей і методів аналізу неоднорідних послідовностей даних;
- розробити моделі неоднорідних послідовностей даних;
- розробити метод відбору значущих чинників при побудові нечіткої багатофакторної регресії для даних, що представлені у вигляді неоднорідних послідовностей;
- вдосконалити метод фільтрації компонент неоднорідних послідовностей даних;
- розробити інтелектуальну інформаційну технологію аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметної області та виконати програмну реалізацію й впровадження результату дослідження при вирішенні практичних задач.

Об’єктом дослідження є процес аналізу неоднорідних послідовностей при оцінюванні поточного стану предметної області.

Предметом дослідження є моделі, методи та інтелектуальна інформаційна технологія аналізу неоднорідних послідовностей для оцінювання поточного стану предметної області на основі коротких вибірок даних.

Методи дослідження: при розробці та дослідженні методу визначення значущих чинників нечіткої регресійної моделі були використані методи теорії нечітких множин та регресійного аналізу. При розробці моделей і методів фільтрації компонент неоднорідних часових послідовностей були використані методи нечіткої апроксимації даних та аналізу динамічних рядів. При проведенні та аналізі результатів експериментальних досліджень були використані елементи математичної статистики.

Наукова новизна одержаних результатів.

В результаті проведення досліджень одержано такі нові результати:

1. Вперше запропоновано метод визначення значущих чинників нечіткої регресійної моделі неоднорідних послідовностей даних, який, на відміну від існуючих, містить етапи підбору коефіцієнтів за критерієм рівнозначності кутів відхилення між вектором похибки і векторами змінних та відбору підмножини значущих чинників з коефіцієнтами, що перевищують порогове значення та дозволяє запобігти перенавчанню нечіткої лінійної регресії та отримати підмножину значущих чинників за скінченну кількість ітерацій.

2. Отримав подальший розвиток метод фільтрації компонент неоднорідних часових послідовностей, в якому, на відміну від існуючих, для виявлення тренду

початкова послідовність ітеративно розбивається на скінчену кількість нечітких розділів, для кожного з яких розраховується усереднене значення із врахуванням функції належності, яка асоційована із нечітким розділенням, що дозволяє підвищити ефективність оцінювання зміни стану предметної області за рахунок фільтрації коливань різних періодів та виділення трендової складової.

3. Отримала подальший розвиток тренд-сезонна модель неоднорідних послідовностей, в якій, на відміну від існуючих моделей, трендова складова подається у вигляді інтерпольованих усереднених значень із врахуванням функції належності, яка асоційована із кожним нечітким розділенням, що дозволяє застосовувати дану модель для коротких вибірок без втрати крайових значень і тим самим підвищити ефективність моделювання стану предметних областей в інформаційно-аналітичних системах.

Практична значимість одержаних результатів. На основі запропонованих моделей та методів удосконалено інформаційну технологію аналізу даних неоднорідних послідовностей, яка, на відміну від існуючих технологій, надає можливість при побудові моделі предметної області додатково враховувати відомості подані у вигляді нечітких даних, та здійснювати відбір значущих чинників, що надає можливість проведення аналізу в умовах коротких вибірок даних.

Інформаційна технологія реалізована у вигляді програмного модуля, робота якого ґрунтується на запропонованій тренд-сезонній моделі даних впорядкованих за часом значень із врахуванням функцій належності, асоційованих із нечіткими розділами, методі фільтрації компонент динамічного ряду із врахуванням крайових значень та методі визначення значущих чинників нечіткої регресійної моделі неоднорідних даних, який містить етап відбору підмножини значущих чинників.

Впроваджено інформаційну технологію у вигляді програмного засобу у діяльність ТОВ «Ендейвер», м.Полтава (акт впровадження від 14.03.2017) та в діяльність Головного управління національної поліції Харківської області (акт впровадження від 20.12.2016).

Результати дисертаційної роботи впроваджені в навчальний процес кафедри програмної інженерії ХНУРЕ (акт впровадження від 15.03.2017).

Особистий вклад здобувача. Усі результати, що виносяться на захист отримані здобувачем особисто. У роботах опублікованих у співавторстві, здобувачу належать: у роботі [1] – метод відшукування значущих чинників на основі методу найменших кутів для побудови нечіткої регресійної моделі; у роботі [2] – метод виявлення сезонних коливань із застосуванням нечіткого згладжування на базі F-перетворення; у роботі [3] – модифікований метод аналізу сезонних коливань злочинності; у роботі [4] – підхід для аналізу тенденцій розвитку злочинності; у роботі [5] – інформаційна система моделювання впливу чинників злочинності; у роботі [6] – метод попереднього аналізу даних в системах обробки інформації про скоєні злочини; у роботі [7] – статистична модель аналізу злочинності; у роботі [10] – інформаційна система для виклику екстрених служб; у роботі [12] – модифікований метод визначення чинників

злочинності на основі методу найменших кутів; у роботі [13] – метод оцінювання злочинності із врахуванням нечіткості; у роботі [14] – застосування F-перетворення для аналізу риноманометричних сигналів; у роботі [15] – застосування F-перетворення для апроксимації риноманометричних сигналів; у роботі [16] – застосування F-перетворення як один з підходів для відшукування значущих ознак риноманометричних сигналів;

Апробація роботи. Основні положення дисертаційної роботи доповідалися на таких міжнародних конференціях і форумах:

- XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), (Львів, 2016 р.);
- First International Conference on Data Stream Mining & Processing (DSMP), (Львів, 2016 р.);
- Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), (Львів, 2015 р.);
- V Міжнародній науково-практичній конференції «Спеціальна техніка у правоохоронній діяльності» (Київ, 2012 р.);
- Міжнародній науково-технічній конференції «Інформаційні системи і технології» (Харків, 2012 р.);
- II міжнародній науково-технічній конференції «Інформаційні технології в навігації і управлінні» (Київ, 2011 р.);
- IV міжнародній науково-практичній конференції «Наука і соціальні проблеми суспільства: інформатизація і інформаційні технології», (Харків, 2011р.);
- VII міжнародному науковому конгресі «Державне управління та місцеве самоврядування», (Харків, 2007 р.);

Публікації. За результатами дисертаційного дослідження опубліковано 16 наукових праць (серед них 3 – одноосібних), у тому числі 7 статей у наукових фахових виданнях України з технічних наук і 1 стаття – за кордоном у виданні, що входить до міжнародної наукометричної бази Scopus, 8 тез у матеріалах міжнародних конференцій (з них 3 включено до наукометричної бази Scopus).

Структура дисертації. Дисертація складається зі вступу, чотирьох розділів, висновків, списку використаних джерел зі 167 найменувань на 16 сторінках та 1 додаток на 4 сторінках, а також містить 25 рисунків і 3 таблиці. Загальний обсяг роботи складає 156 сторінок, включаючи 117 сторінок основного тексту.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність розробки нових технологій обробки даних отриманих від людино-машинних систем, потреба в їх аналізі та побудові прогнозів на основі цих даних. Розглянуто сучасний стан та підходи до розв’язку цієї проблеми, визначені мета, об’єкт, предмет і методи дослідження, наведено задачі, що розв’язуються у дисертаційній роботі, зв’язок з науковими темами, наведено наукову новизну та практичне значення отриманих результатів, перелік публікацій за темою роботи, надано інформацію про особистий внесок автора.

В першому розділі проведено аналіз існуючих публікацій щодо моделей, методів та інформаційних технологій обробки великих масивів даних з різнорідною природою походження.

Розглянуто основні проблеми, які виникають під час аналізу даних представлених у вигляді неоднорідних послідовностей.

Розглянуто методи побудови нечітких регресійних моделей, умови їх використання та обмеження, що існують для застосування цих моделей.

Сформульовані мета та задачі дисертаційних досліджень.

Другий розділ присвячено розробці моделей неоднорідних послідовностей, зокрема вдосконаленої тренд-сезонної моделі.

Описуються показники, що можуть використовуватися для аналізу неоднорідних послідовностей, методи вимірювання взаємозв'язку між неоднорідними послідовностями та моделі аналізу неоднорідних послідовностей.

Обгрунтовано метод побудови нечіткої регресійної моделі для використання в дослідженнях, що пов'язані із обробкою неоднорідних послідовностей, на прикладах аналізу статистичних обліків рівнів злочинності і окремих її видів та аналізу даних риноманометричних вимірювань із врахуванням відомостей з анамнезу.

Запропоновано тренд-сезонну модель із використанням елементів теорії нечітких множин.

Нехай неоднорідна послідовність подається подається у вигляді динамічного ряду: $X = x_1, x_2, x_3 ... x_n$ де $i = \overline{1, N}$ значення якого впорядковані за зростанням часу i. Представимо його відповідно до формули (1):

$$x_i = U_i + V_i + \varepsilon_i \qquad (1)$$

де U_i - тренд, V_i - сезонна компонента, ε_i - випадкова компонента, N - число рівнів спостереження. Щодо U_i вважається, що ця компонента згенерована деякою гладкою функцією, міра гладкості якої заздалегідь невідома. Сезонна компонента V_i має період T_0, таким чином, що $V_{i+T_0} = V_i$.

Для моделювання сезонності на відміну від традиційного підходу, який полягає в тому, щоб за допомогою статистичних методів отримати відповідь про наявність сезонності в певному чітко визначеному переліку місяців, запропоновано застосування нечіткого підходу – де наявність сезонності визначається певним рівнем належності кожного з місяців до певного сезону, а функція належності може приймати не лише два значення 0 та 1, а всі значення з діапазону [0, 1], характеризуючи таким чином належність елемента до множини.

В цьому випадку нечіткі множини дозволяють досить вдало передавати відомості отримані від людини і виражені у вигляді лінгвістичних змінних. Тобто задається нечітке розбиття часового ряду даних, і для центру кожної з отриманих нечітких множин елементів ряду визначається усереднене значення.

Застосовуючи цей підхід, пропонується використовувати для адитивної тренд-сезонної моделі (1) обчислення і подання трендової компоненти із

використанням F-перетворення.

Трендова складова U_i відшукується з використанням F-компоненти, тобто необхідно створити n нечітких розділів, які відповідають умовам розбиття Руспіні, для кожного з них визначити центр розбиття t_k, причому $t = \overline{1..N}$, де $k = \overline{1..n}$, функції належності A_k і значення U_k за формулою (3).

$$U_k = \frac{\sum X_{t_i} A_k(t_i)}{\sum A_k(t_i)}, \qquad (3)$$

де існують такі t, що $A_k(t) > 0$.

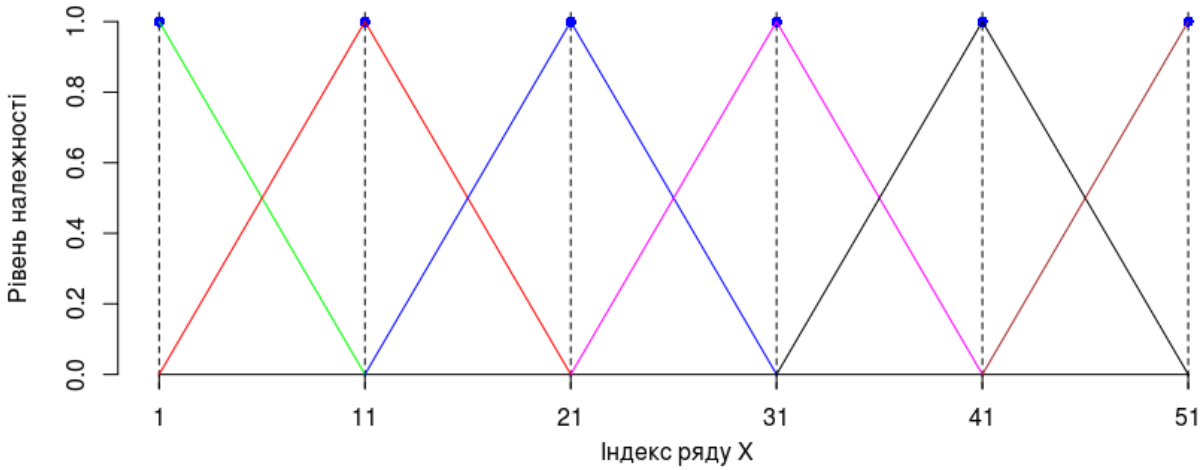

Рисунок 1 – Розбиття на шість нечітких розділів

У випадку, якщо значення тренду U_i необхідно відшукати для значень які відмінні від центрів нечіткого розбиття, тоді використовуються значення F-компоненти, що передує шуканому - U_{d_0} та наступне за ним U_{d_1}, після чого розраховують $U_i^{(1)}$ (4):

$$U_i^{(1)} = U_{d_0} + \frac{U_{d_1} - U_{d_0}}{d_1 - d_0}(i - d_0) \qquad (4)$$

Середню сезонну хвилю можна представити у вигляді $\bar{l}_{ij} = l'_{ij} / \sigma_i$, де l'_{ij} - окремі місячні відхилення.

Для аналізу багатовимірних даних впорядкованих сукупностей чітких та нечітких значень розроблена модель із врахуванням даних попарних порівнянь експертних оцінок для непрямого визначення параметрів функції належності. Запропоновано подати дані попарних порівнянь експертних оцінок у вигляді матриці M, таким чином, що $M = \{m_{ij}\}$, де кожен елемент m_{ij} відображає відношення ступеня належності характеристики двох значень до нечіткої

множини S, $\mu_S(x)$ – відображає рівні належності елементів x до нечіткої множини S і для потреб нечіткої лінійної регресійної моделі апроксимується трикутною функцією належності.

Нехай $\tilde{Y}$ - нечіткі дані неоднорідних послідовностей, $X_j = \{x_{ij}\}$, $j = 1..n, i = 1..m$ - чіткі значення факторів, тоді рівняння регресії прийме такий вигляд (5):

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 X_1 + ... + \tilde{A}_n X_n, \tag{5}$$

де $\tilde{Y}_i = (y_i, e_i)$, $i = 1..m$ - нечітка величина з центром y_i і шириною e_i, $\tilde{A}_j = (a_j, c_j)$, $j = 0..n$ - нечітка величина з центром a_i і шириною c_i. Необхідно мінімізувати функцію (6):

$$S = c_0 + \sum_{j=1}^{n} c_j \sum_{i=1}^{m} x_{ij}, \tag{6}$$

за умови, що $c_0 \geq 0, c_j \geq 0$ та виконуються нерівності (7) і (8)

$$a_0 + \sum_{j=1}^{n} a_j x_{ij} + (1-h)\left[c_0 + \sum_{j=1}^{n} c_j \left| x_{ij} \right| \right] > y_i + (1-h)e_i, \tag{7}$$

$$a_0 + \sum_{j=1}^{n} a_j x_{ij} - (1-h)\left[c_0 + \sum_{j=1}^{n} c_j \left| x_{ij} \right| \right] < y_i - (1-h)e_i, \tag{8}$$

де $h \in \{0;1]$ – коефіцієнт чіткості. Результатом розв'язку цієї задачі є a_j та c_j, тобто нечіткі коефіцієнти. В такій постановці задача розв'язується методом лінійного програмування.У випадку, якщо відомості про чинники неоднорідних послідовносетй даних подані у вигляді нечітких значень, пропонується використати нечітку регресійну модель (9)

$$\tilde{Y} = (\tilde{A}_0 \oplus \tilde{A}_1 \otimes \tilde{X}_1 \oplus ... \oplus \tilde{A}_j \otimes \tilde{X}_j) = \tilde{A} \otimes \tilde{X}, \tag{9}$$

де $\tilde{y}_i = (y_i, e_i)_L$ - компоненти вектора $\tilde{Y}$, нечітка величина з цетром y_i і шириною e_i; $\tilde{x}_{ij} = (x_{ij}, v_{ij})_L$ - компоненти вектора X_j, нечітка величина з центром x_{ij} і шириною v_{ij}; $\tilde{A}_j = (a_j, c_j)_L$ - нечітка величина з центром a_j і шириною c_j. $L(x)$ - функція належності, така що $L(x) = L(-x)$ і $L(0) = 1, L(1) = 0$ а також L зростає на $[0, \infty)$ і може бути обернена на $[0,1]$. Необхідно мінімізувати функцію (10):

$$S = \sum_{i=1}^{n} \max_{1 \leq j \leq p} (\left| a_j \right| v_{ij}, \left| x_{ij} \right| c_j) - \left| L^{-1}(h) \right| e_i, \tag{10}$$

за умови (11) і $c_j \geq 0$:

$$\left| y_i - \sum_{j=1}^{p} a_j x_{ij} \right| \leq \left| L^{-1}(h) \right| \max_{1 \leq j \leq p} (\left| a_j \right| v_{ij}, \left| x_{ij} \right| c_j) - \left| L^{-1}(h) \right| e_i \qquad (11)$$

Розв'язком цієї задачі є значення коефіцієнтів a_j та c_j, тобто $\tilde{A}_j = (a_j, c_j)_L$.

Таким чином, були запропоновані моделі неоднорідних послідовностей, зокрема отримала подальшого розвитку тренд-сезонна модель неоднорідних послідовностей даних, в якій було запропоновано подання трендової складової у вигляді інтерпольованих усереднених значень (F-компонент) із врахуванням функції належності, яка асоційована із кожним нечітким розподілом. Це дозволяє застосовувати дану модель для коротких вибірок без втрати крайових значень.

Третій розділ присвячено розробці методів аналізу неоднорідних послідовностей. Запропонований метод відбору значущих чинників при побудові нечіткої багатофакторної регресії для даних, що представлені у вигляді неоднорідних послідовностей. Вдосконалений метод фільтрації компонент неоднорідних послідовностей даних.

Запропоновано використовувати метод фільтрації на основі F-перетворення на етапі виявлення тренду. Подамо часовий ряд (1) у вигляді $x_{ij} = U_{ij} + V_{ij} + \varepsilon_{ij}$, $i = \overline{1..m}$ де m- число періодів, $j = \overline{1..T_0}$, де T_0 - період спостереження. Алгоритм фільтрації рівнів динамічного ряду з виявленням періодичних коливань у запропонованому методі реалізується такими етапами ітераційного процесу:

Eman 1. Розділимо динамічний ряд на n частин. Визначимо n рівновіддалених точок з індексом t_k, де $k = \overline{1..n}$, які належать до цих нечітких частин, причому $t = \overline{1..N}$, а $t_k = 1 + h(k-1)$, де $N > n, h = (N-1)/(n-1)$.

Eman 2. Визначимо n базисних функцій $A_1 ... A_n$, які покривають всі частини динамічного ряду та відповідають таким умовам: A_k - неперервна; A_k монотонно зростає на $[t_{k-1}, t_k]$ і монотонно спадає на $[t_k, t_{k+1}]$; $A_k : [1..N] \to [0,1], A_k(t_k) = 1$; $A_k(t) = 0$, якщо $t \notin (t_{k-1}, t_{k+1})$, при цьому вважаємо що $t_0 = t_1 = 1$ і $t_{n+1} = t_n = N$; $\sum A_k(t) = 1$ для всіх $t \in [1, N]$. Пропонується використати базисну функцію (12)

$$A_k(t) = \begin{cases} \dfrac{t - t_{k-1}}{t_k - t_{k-1}}, if : t_{k-1} \leq t \leq t_k \\ \dfrac{t_{k+1} - t}{t_{k+1} - t_k}, if : t_k \leq t \leq t_{k+1} \\ 0, в\ інших\ випадках \end{cases} \qquad (12)$$

Етап 3. Використовуючи базисні функції, перетворимо динамічний ряд X в кортеж з n дійсних чисел $[U_1...U_n]$, які визначаються за формулою (3).

F-компоненти, тобто U_k, представляють собою точки, які належать тренду динамічного ряду. Для одержання значень тренду в інших точках скористаємося лінійною інтерполяцією (4) між двома найближчими точками. У результаті отримуємо попередню оцінку тренду $U_i^{(1)}$ і відхилення емпіричного ряду від вирівняного $l'_{ij} = x_{ij} - U_{ij}^{(1)}$

Етап 4. Для кожного року i обчислюється σ_i- середнє квадратичне відхилення (13):

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{T_0} l'^2_{ij} - \frac{\left(\sum_{j=1}^{T_0} l'_{ij}\right)^2}{T_0}}{T_0 - 1}} \quad , \tag{13}$$

на яке діляться окремі відхилення відповідного періоду $\overline{l_{ij}} = l'_{ij} / \sigma_i$

Етап 5. З "нормованих" таким чином відхилень обчислюється в першому наближенні середня сезонна хвиля (14):

$$V_j^{(1)} = \frac{\sum_{i=1}^{m} \overline{l_{ij}}}{m} \quad . \tag{14}$$

Етап 6. Середня хвиля множиться на середнє квадратичне відхилення кожного року і віднімається від рівнів початкового емпіричного ряду (15):

$$\overline{=}{x_{ij}} = x_{ij} - V_j^{(1)} \cdot \sigma_i \quad . \tag{15}$$

Етап 7. Цей ряд знову піддається нечіткому згладжуванню (довжина обирається в залежності від інтенсивності дрібних коливань). В результаті одержується нова оцінка тренда $U_i^{(2)}$.

Етап 8. Відхилення початкового емпіричного ряду $\{x_i\}$ від ряду $\{U_i^{(2)}\}$, одержаного на етапі 5, $l_i^{(2)} = x_i - U_i^{(2)}$ знову піддаються аналогічному опрацюванню відповідно до етапів 2 і 3 для виявлення наступного наближення середньої хвилі.

Цей ітераційний процес повторюється до виділення хвилі із заданою точністю.

Для проведення багатофакторного аналізу неоднорідних послідовностей запропонований метод визначення значущих чинників нечіткої регресійної моделі неоднорідних даних

Нехай $\tilde{Y}$ - нечіткі дані про значення функції, $i = 1..m$ - чіткі фактори.

Етап 1. Для визначення функції належності скористаємось підходом описаним в пункті один третього розділу на основі методу попарних порівнянь.

Створена в такому випадку функція буде мати такий вигляд (16):

$$\mu_{\tilde{Y}_i} = \max\left\{1 - \frac{y - y_i}{e_i}, 0\right\}, \tag{16}$$

де y_i - центр нечіткої величини, e_i - розкид значень нечіткої величини.

Етап 2. Подамо чіткі дані про фактори у вигляді матриця значень m пояснюючих змінних у n спостереженнях (17):

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} \tag{17}$$

а дані про центри нечіткої величини $\tilde{Y}$ у вигляді (18):

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{18}$$

Ці данні пропонується використати для відшукування значущих чинників нечіткої регресійної моделі.

Етап 3. Задамо початкову оцінку $\hat{\mu}_A = 0$ вектора значень залежної змінної y.

Етап 4. Обчислимо вектор кореляцій: $\hat{c} = X^T(y - \hat{\mu}_A)$.

Етап 5. Знайдемо поточний набір індексів A, що відповідає ознакам із найбільшими абсолютними значеннями кореляцій: $A = \{j : |\hat{c}_j| = \hat{C}\}$,

де $\hat{C} = \max\limits_{j=1,\dots,n}\{|\hat{c}_j|\}$.

Етап 6. Знайдемо $s_j = sign(\hat{c}_j)$ для $j \in A$. Розрахуємо матриці X_A, ψ_A за допомогою формули (19):

$$X_A = \left[s_{j_1} x_{j_1}, \dots, s_{j_{|A|}} x_{j_{|A|}} \right], j = (j_1, \dots, j_{|A|}) \in A,\ \psi_A = (1_A^T \varsigma_A^{-1} 1_A)^{-\frac{1}{2}}, \tag{19}$$

де $s_j \in \{+1,-1\}$ і $|A|$- потужність множини A (кількість значень множини A),

$\varsigma = X_A^T X_A$, 1_A – одинична матриця розміру $1 \times |A|$

Етап 7. Розрахуємо вектор $a = X^T u_A$, де $u_A = X_A w_A$, $w_A = \psi_A \varsigma_A^{-1} 1_A$.

Етап 8. Розрахуємо значення $\hat{\gamma} = \min\limits_{j \in A}^{+} \left\{ \dfrac{\hat{C} - \hat{c}_j}{\psi_A - a_j}, \dfrac{\hat{C} + \hat{c}_j}{\psi_A + a_j} \right\}$, (мінімум береться по всім додатнім значенням для кожного j).

Етап 9. Знаходимо значення $\hat{\mu}_A$ для наступної ітерації: $\hat{\mu}_{A+} = \hat{\mu}_A + \hat{\gamma} u_A$.

Етап 10. Процес повторюється n раз (де n – кількість факторів), починаючи з етапу 4. Для кожної ітерації обчислюється коефіцієнт Cp Маллоуза .

Етап 11. Для побудови нечіткої регресійної моделі обираємо набір коефіцієнтів, який буде відповідати мінімальному значенню коефіцієнта *Ср.*

Після цього застосовується метод розв'язку задачі побудови нечіткої лінійної регресійної моделі із використанням лише обраного набору факторних змінних(20):

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 X'_1 + ... + \tilde{A}_k X'_k, \qquad (20)$$

де X' - чіткі значення факторів, обрані за допомогою запропонованого методу.

Четвертий розділ присвячено розробці інтелектуальної інформаційної технологію аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметної області та виконанню програмної реалізації і впровадження результату дослідження для вирішення практичних задач.

Розглянемо етапи інтелектуальної інформаційної технології аналізу неоднорідних послідовностей.

Перший етап функціонування інтелектуальної інформаційної технології полягає в отриманні початкових відомостей та значень, які в подальшому використовуються для формування рядів даних. Дані можуть отримуватися або шляхом запиту у вигляді SQL до одного з сховищ даних на основі реляційної СУБД або запита REST, що відправляється до бази даних чи сервісу, який підтримує WEB API для передачі інформації.

Отримані на цьому етапі дані зберігаються. Також інформаційна технологія підтримує можливість самостійного внесення даних користувачем, зазвичай це відбувається за допомогою завантаження даних із файлового ресурсу.

Другий етап інтелектуальної інформаційної технології полягає в формуванні динамічних рядів даних. В залежності від того, яка інформація є в наявності, користувач може вводити як кількісні значення показників та факторів, що зумовлюють значення величини, так і дані подані у вигляді нечітких значень. Дані про нечіткі значення рівнів чинників та/або відгуку можуть задаватися за допомогою внесення самих значень та функції належності до нечіткої множини, так і за допомогою внесення результатів експертних опитувань.

Схематично інтелектуальна інформаційна технологія відображена на рис. 2.

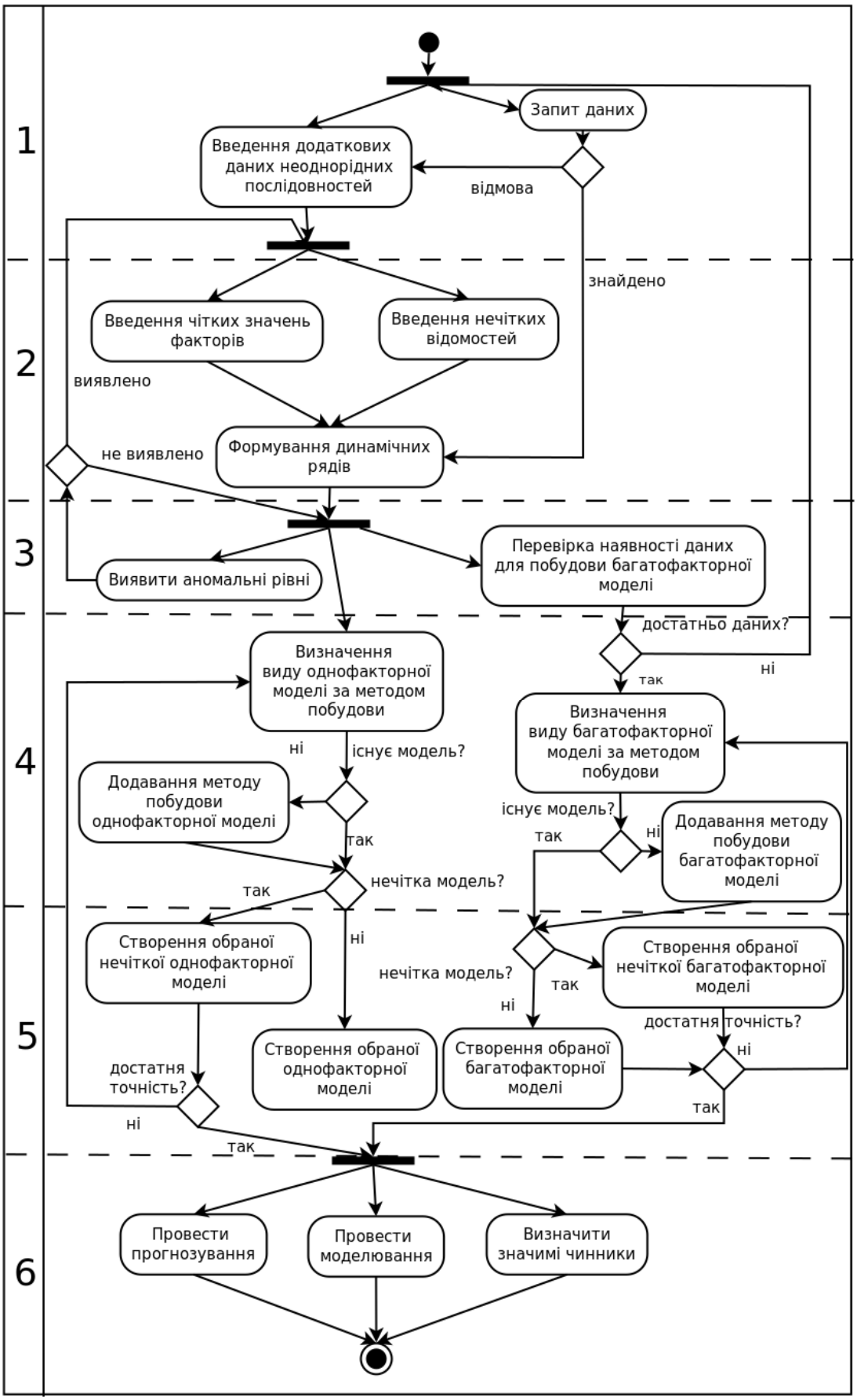

Рисунок 2 – Схема етапів інформаційної технології

В подальшому з цих даних за допомогою непрямого методу побудови функцій належності формуються функції належності нечіткої величини.

Третій етап полягає в перевірці сформованих рядів даних на наявність аномальних рівнів та достатньої кількості значень для проведення багатофакторного аналізу. Якщо кількість даних не достатня, користувач повертається до етапу запиту даних з додаткових джерел даних. Це пов'язано з неможливістю побудови багатофакторної моделі на основі неоднорідних послідовностей за недостатньою кількістю значень рядів даних.

У випадку наявності рівнів даних з підозрою на аномальність, користувач обирає один із варіантів подальшої обробки таких даних, які ґрунтуються на відкиданні цього рівня, або заміні його на інше значення.

Четвертий етап інтелектуальної інформаційної технології полягає у виборі способу побудови моделі.

Користувачу пропонується здійснити вибір способу створення моделі та підключити додаткові модулі, що реалізують моделі. Такі модулі можуть створюватися окремо та розширяти можливості системи. При роботі із додатковим модулем, користувач повинен визначити тип задачі з переліку типових задач, які виникають під час налізу даних.

В процесі роботи, вибір користувача зберігається і задачу можна вибрати із переліку тих проблем, які уже розв'язувалися.

Після формулювання задачі на екрані монітора з'являється схема алгоритму розв'язування поставленої задачі. Тут зображується стандартний алгоритм, тобто алгоритм, який уже використовувався при розгляді подібної задачі. Алгоритм будується за схемою "І-АБО", надаючи користувачу можливість вибору однієї з декількох альтернатив.

Також користувачу надається можливість внести певні зміни в даний алгоритм (застосувати інший метод розв'язування певної підзадачі чи іншу модель). Такі методи та моделі додаються в базу знань і в подальшому фігурують в альтернативах.

Аналогічно формується алгоритм розв'язування задачі в тому випадку, якщо за результатами моделювання виникає потреба у зміні моделі досліджуваного процесу. Після цього автоматично підбираються потрібні модулі і встановлюються потрібні інформаційні зв'язки між ними з використанням таблиць. При формулюванні невідомої задачі програмна система разом із користувачем створює алгоритм її розв'язування у режимі "запитання-відповідь". Такий алгоритм подається у вигляді графа.

Для реалізації такої системи пропонується застосувати підхід, який забезпечує можливість моделей, що застосовуються в системах, адаптуватися до конкретної, специфічної реальності в результаті діалогу з користувачем і можливість системи інтерактивного генерування моделей.

П'ятий етап інтелектуальної інформаційної технології полягає в застосуванні моделей розглянутих в другому розділі за допомогою методів аналізу неоднорідних послідовностей даних. За допомогою розроблених рішень на даному етапі можуть створюватися однофакторні і багатофакторні моделі

неднорідних послідовностей. В залежності від наявності даних експертних оцінювання формуються чіткі або нечіткі моделі. Також на цьому етапі надається можливість створити тренд-сезоні моделі часових рядів, в тому числі для випадку коротких вибірок даних.

Шостий етап інтелектуальної інформаційної технології дозволяє провести моделювання із використанням даних підготованих на попередніх етапах, з використанням параметрів розрахованих для різних моделей.

Відповідно до побудованих моделей можуть бути визначені значимі чинники та проведений аналіз із використанням як багатофакторних, так і однофакторних моделей. Реалізація здійснюється у вигляді модуля, написаного мовою програмування Python.

У ході дослідження було проведено ряд експериментів з аналізу неоднорідних послідовностей із побудовою моделей та за допомогою методів викладених в попередніх розділах із застосуванням запропонованої інтелектуальної інформаційної технології.

Результати, які були отримані в ході експериментів свідчать, що:

- під час побудови тренд-сезонної моделі рівнів злочинності із умовою наявності лише короткого ряду даних, для вилучення трендової складової за допомогою запропонованого методу фільтрації компонент неоднорідних послідовностей враховувалися крайові значення ряду, що дозволило виділити сезонну хвилю і відповідно зменшити час і витрати на формування початкових даних на 33% та підвищити ефективність в 1,5 рази в умовах коротких вибірок даних;

- для нечіткої регресійної моделі, застосованої для моделювання показників носового дихання за допомогою розробленого методу визначення значимих чинників нечіткої регресійної моделі, була визначена множина значущих факторів, що дозволило запобігти перенавчанню нечіткої лінійної регресії та отримати підмножину значущих чинників за скінченну кількість ітерацій і в результаті зменшити кількість помилок на тестовій вибірці на 17% і підвищити ефективність в 1,2 рази у порівнянні із методом на основі крокового нечіткого регресійного аналізу

У додатках наведено акти про впровадження результатів дисертаційної роботи у діяльність ТОВ «Ендейвер», м.Полтава, Головного управління національної поліції Харківської області та навчальний процес кафедри програмної інженерії ХНУРЕ.

ВИСНОВКИ

У дисертаційній роботі розв’язана нова актуальна науково-практична задача розробки моделей, методів та створення на їх основі інтелектуальної інформаційної технології аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметних областей в інформаційно-аналітичних системах. Внаслідок виконання роботи отримані нові наукові та практичні результати.

1. Дослідження сучасних моделей і методів аналізу неоднорідних послідовностей даних для задач оцінювання поточного стану предметної області показало, що існує потреба в створенні методу побудови багатофакторної нечіткої регресійної моделі неоднорідних послідовностей даних із урахуванням значущих чинників. Також на основі проведених досліджень було виявлено, що тренд-сезонні адитивні моделі неоднорідних часових послідовностей, які ґрунтуються на використанні ковзного середнього, потребують вдосконалення пов'язаного із використанням крайових значень часового ряду в умовах коротких вибірок даних.

2. Отримали подальший розвиток моделі неоднорідних послідовностей даних, в тому числі тренд-сезонна модель неоднорідних часових послідовностей шляхом утворення нечітких розділів із асоційованими функціями належності, які враховуються при поданні трендової складової у вигляді інтерпольованих усереднених значень. Це дозволяє застосовувати вдосконалену модель для коротких вибірок без втрати крайових елементів неоднорідної послідовності даних.

3. Вперше запропонований метод визначення значущих чинників нечіткої регресійної моделі даних неоднорідних послідовностей на основі відбору підмножини значущих чинників з коефіцієнтами, які перевищують порогове значення. Коефіцієнти підбирається за критерієм рівнозначності кутів відхилення між вектором похибки і векторами змінних. Запропонований метод дозволяє отримати підмножину значущих чинників за скінченну кількість ітерацій та запобігти перенавчанню нечіткої лінійної регресії.

4. Отримав подальший розвиток метод фільтрації компонент неоднорідних часових послідовностей шляхом застосування ітеративного розбиття початкової послідовності на скінчену кількість нечітких розділів з кожним з яких асоційована функція належності, яка враховується при отриманні усереднених значень для кожного з центрів нечітких розділів, за допомогою яких подається трендова складова. Значення, які знаходяться на ділянках поза центрами нечітких розділів розраховуються за допомогою інтерполяції. Це дозволяє відфільтровувати коливання різних періодів при виділенні трендової складової і тим самим підвищити ефективність оцінювання зміни стану предметної області.

5. На основі запропонованих моделей і методів аналізу неоднорідних послідовностей даних було розроблено інтелектуальну інформаційну технологію аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметних областей в інформаційно-аналітичних системах. Розроблено програмний засіб, який реалізує запропоновану інтелектуальну інформаційну технологію. Застосування запропонованої інтелектуальної інформаційної технології для моделювання показників носового дихання дозволило зменшити кількість помилок на тестовій вибірці на 17% і підвищити ефективність в 1,2 рази. Використання для дослідження сезонної складової дозволило зменшити час і витрати на формування початкових даних на 33% та підвищити ефективність в 1,5 рази.

6. Проведено впровадження моделі, методів та інтелектуальної

інформаційної технології при вирішенні практичних задач в діяльність ТОВ «Ендейвер», м. Полтава, та в діяльність Головного управління національної поліції Харківської області, а також в навчальний процес кафедри програмної інженерії ХНУРЕ, що підтверджено відповідними актами впровадження.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. A. L. Yerokhin, A. S. Babii, A. S. Nechyporenko, O. P. Turuta. A Lars-Based Method of the Construction of a Fuzzy Regression Model for the Selection of Significant Features. Cybernetics and Systems Analysis, Springer US, 2016. V. 52, Issue 4, P. 641–646, DOI:10.1007/s10559-016-9867-5 (Входить до міжнародної наукометричної бази SCOPUS).

2. Зацеркляний М.М., Єрохін А.Л., Бабій А.С.,Турута О.П. Розробка методу виявлення сезоних коливань з застосуванням нечіткого згладжування на базі F-перетвореннння. Біоніка інтелекту, Харків: ХНУРЕ, 2011. №2011’2. С. 89 – 93

3. Бабій А. С., Зацеркляний М. М.. Автоматизація аналізу сезонних коливань рівня злочинності. Право і безпека. Харків: ХНУВС, 2005. Т4. № 3. С.163-166.

4. Бабій А. С., Зацеркляний М. М.. Аналіз тенденцій розвитку злочинності. Системи обробки інформації. Харків: ХУПС, 2007. №4. С. 153-155.

5. Зацеркляний М. М., Бабій А. С. Інформаційна система моделювання впливу чинників злочинності. Право і Безпека. Харків: ХНУВС, 2008. Т.7. № 2. С. 204-209.

6. Зацеркляний М. М., Бабій А. С.. Попередній аналіз даних у системах обробки інформації про скоєні злочини. Право і Безпека. Харків: ХНУВС, 2009. № 1. С. 269-272.

7. Лановий О.Ф., Бабій А.С. Статистичний аналіз злочинності. Вісник НТУ ХПІ, Харків: НТУ «ХПІ», 2006. №19. С. 24 – 30

8. Бабій А.С. Програмна система для аналізу злочинності. Вісник НТУ ХПІ, Харків: НТУ «ХПІ», 2007. №19. С. 12 – 16

9. Бабій А.С. Автоматизація управління діяльністю правоохоронних органів. Державне управління та місцеве самоврядування: тези VII міжнародного наукового конгресу, 29-30 березня 2007 р. Харків:НАДУ, 2007. С. 20-22

10. Єрохін А.Л., Бабій А.С.,Турута О.П. Спеціальна інформаційна система для виклику екстрених служб в Україні. ХНУРЕ, Збірник праць IV міжнародної науково-практичної конференції «Наука і соціальні проблеми суспільства: інформатизація і інформаційні технології», 24-25 травня 2011. Харків: ХНУРЕ, 2011. С. 163

11. Бабій А.С. Побудова СППР для оцінювання злочинності. Збірник праць II міжнародної науково-технічної конференції «Інформаційні технології в навігації і управлінні», 16-17 липня 2011 р., Київ: «ДП ЦНДІ НіУ», 2011. С. 41

12. Зацеркляний М.М., Бабій А.С. Застосування методу найменших кутів для аналізу чинників злочинності. Материалы международной научно-технической конференции “Информационные системы и технологии”, Харьков: НТМТ, 2012, С. 37

13. Петров К.Е., Зацеркляний М.М., Бабій А.С. Оцінювання злочинності із врахуванням нечіткості. Спеціальна техніка у правоохоронній діяльності, Матеріали V Міжнародної науково-практичної конференції, Київ: НАВС, 2012, С.79

14. A. Yerokhin ,A. Nechyporenko, A. Babii, O. Turuta .Usage of F-transform to finding informative parameters of rhinomanometric signals. Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), 2015 Xth International. P. 129-132, DOI:10.1109/STC-CSIT.2015.7325449 (Входить до міжнародної наукометричної бази SCOPUS)

15. A. Yerokhin ,A. Nechyporenko, A. Babii, O. Turuta. Processing and analysis of rhinomanometric signals by F-transform approximation - 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP). P. 314 - 317, DOI: 10.1109/DSMP.2016.7583566 (Входить до міжнародної наукометричної бази SCOPUS)

16. Andriy Yerokhin; Oleksii Turuta; Andrii Babii; Alina Nechyporenko; Ihor Mahdalina. Usage of phase space diagram to finding significant features of rhinomanometric signals. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), P. 70 - 72, DOI: 10.1109/STC-CSIT.2016.7589871 (Входить до міжнародної наукометричної бази SCOPUS)

АНОТАЦІЯ

Бабій А.С. Моделі, методи та інтелектуальна інформаційна технологія аналізу неоднорідних послідовностей. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Харків, 2017.

У дисертаційній роботі запропоновано нове вирішення актуальної науково-практичної задачі розробки моделей, методів та інтелектуальної інформаційної технології аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметних областей в інформаційно-аналітичних системах.

На основі модифікованої тренд-сезонної моделі неоднорідних послідовностей, в якій трендова складова подається у вигляді інтерпольованих усереднених значень із врахуванням функції належності нечітких розбиттів, запропонований метод фільтрації компонент неоднорідних часових послідовностей, що дозволяє застосовувати дану модель для коротких вибірок без втрати крайових значень. Розроблений метод визначення значущих чинників нечіткої регресійної моделі неоднорідних послідовностей даних, який дозволяє запобігти перенавчанню нечіткої лінійної регресії.

На основі запропонованих моделей і методів була розроблена інтелектуальна інформаційна технологія аналізу неоднорідних послідовностей та програмне забезпечення.

Ключові слова:нечіткий регресійний аналіз, тренд-сезонна модель, аналіз даних, тренд, неоднорідні послідовності, часові ряди.

АННОТАЦИЯ

Бабий А.С. Модели, методы и интеллектуальная информационная технология анализа неоднородных последовательностей. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 – информационные технологии. – Харьковский национальный университет радиоэлектроники, Министерство образования и науки Украины, Харьков, 2017.

В диссертационной работе предложено новое решение актуальной научно-практической задачи разработки моделей, методов и интеллектуальной информационной технологии анализа неоднородных последовательностей данных для оценки текущего состояния предметных областей в информационно-аналитических системах.

Основываясь на модифицированной тренд-сезонной модели неоднородных последовательностей, в которой трендовая составляющая представляется в виде интерполированных усредненных значений с учетом функции принадлежности нечетких разбиений, предложен метод фильтрации компонент неоднородных временных последовательностей, что позволяет применять данную модель для коротких выборок без потери краевых значений. Разработан метод определения значимых факторов нечеткой регрессионной модели неоднородных последовательностей данных, который позволяет предотвратить переобучение нечеткой линейной регрессии.

На основе предложенных моделей и методов была разработана интеллектуальная информационная технология анализа неоднородных последовательностей и программное обеспечение.

Ключевые слова: нечеткий регрессионный анализ, тренд-сезонная модель, анализ данных, тренд, неоднородны последовательности, временные ряды.

ABSTRACT

Babii A.S. Models, methods and intelligent information technology for analysis of heterogeneous sequences. – Manuscript.

A thesis for obtaining the candidate degree in technical sciences in the speciality 05.13.06 – information technology. – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2017.

Solution of the actual scientific and practical problem of the development of models, methods and intelligent information technology for the analysis of heterogeneous data sequences to assess the current state of the domain for information-analytical system.

The model of the seasonal component is proposed on the basis of the decomposition approach. In contrast to actual models its trend component is given in the form of interpolated averaged values with regard to membership function, which is

associated with every fuzzy partition that permits to use this model for short series without loss of the boundary data.

The method of filtration of components of heterogeneous time sequences received its further development, in contrast to actual methods the initial sequence is broken into a finite number of fuzzy partitions in order to find out the trend, and for each of these segments the averaged value is calculated with regard to the membership function associated with the fuzzy segment, that permits to take into account the boundary values of the series in order to find the trend component

Trend component were estimated with the usage of the iterative application of F-transformation, that by making fuzzy partitions with defining the centre of partition and membership function for each of them calculate F-component values which represent points belonging to the trend of the heterogeneous sequence.

The method is proposed in the thesis research to determine significant factors of the fuzzy regression model of heterogeneous data; this method, in contrast to the actual methods, includes stages of factors selection in accordance with the criterion of equal significance of angles of deviation between the vector of errors and vectors of variables, as well as selection of subsets of significant factors with coefficients that exceed the threshold value, that allows avoiding any over-fit of the fuzzy linear regression and receiving the subset of significant factors on the basis of the finite number of iterations.

The intellectual information technology of the heterogeneous sequence analysis and software is developed based on proposed models and methods.

This technology gives possibility to flexibly set up for an explicit specific task due to the dialog with the user, because the possibility appears of interactive generation of models that permits to use a number of alternative models that compose the functionality of the system.

The practical implications consist in the possibility to use the proposed intellectual information technology of heterogeneous sequence analysis in order to solve tasks of heterogeneous data processing, whose part is expressed in the form of clearly defined qualitative values received as a result of measurements, and the other part is fuzzy.

Keywords: fuzzy regression analysis, trend-season model, data analysis, trend-seasonal model, trend, heterogeneous sequences, model of social phenomena, time series.