Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

ШАФРОНЕНКО АЛІНА ЮРІЇВНА

УДК 004.032.26

МЕТОДИ ДИНАМІЧНОГО ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ З ПРОПУСКАМИ

05.13.23 – системи та засоби штучного інтелекту

Автореферат

дисертації на здобуття наукового ступеня

кандидата технічних наук

Харків - 2014

Дисертацією є рукопис.

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

Науковий керівник — доктор технічних наук, професор
Бодянський Євгеній Володимирович,
Харківський національний університет радіоелектроніки,
професор кафедри штучного інтелекту.

Офіційні опоненти: доктор технічних наук, професор
Пелешко Дмитро Дмитрович,
Національний університет «Львівська політехніка»,
МОН України, професор кафедри інформаційних технологій видавничої справи;

доктор технічних наук, доцент
Субботін Сергій Олександрович,
Запорізький національний технічний університет,
МОН України, професор кафедри програмних засобів.

Захист відбудеться «___» _____ 2014 р. о _____ годині на засіданні спеціалізованої вченої ради Д 64.052.01 у Харківському національному університеті радіоелектроніки за адресою: Україна, 61166, м. Харків, пр. Леніна, 14.

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

Автореферат розісланий «___» _____2014р.

Учений секретар
спеціалізованої вченої ради О.А. Винокурова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Останнім часом для розв’язання задач інтелектуального аналізу даних все частіше використовуються методи обчислювального інтелекту. В багатьох задачах, пов’язаних з обробкою емпіричних кількісних даних, досить часто зустрічається ситуація, коли вихідні дані містять пропуски. Задачам відновлення пропущених спостережень приділялось достатньо уваги, при цьому найефективнішими в даній ситуації виявилися підходи, що засновані на математичному апараті гібридних систем обчислювального інтелекту. Разом з тим, описані підходи до відновлення пропусків працездатні лише у випадках, коли вихідна інформація у вигляді таблиці «об’єкт - властивість» задана апріорно та кількість її рядків або стовпців не може змінюватися в процесі обробки. Існує велика кількість задач, коли дані поступають на обробку послідовно у реальному часі, при цьому завчасно невідомо, котрий з векторів-образів, що оброблюються, містять пропуски. Такі задачі розглядаються в динамічному інтелектуальному аналізі даних.

Не дивлячись на велику кількість наукових праць, все ще існує проблема розробки методів для роботи із викривленими даними в умовах нечітких класів, що перетинаються або перекриваються, а також методів, що можуть працювати в режимі послідовної обробки. Тому розробка методів та моделей динамічного аналізу даних для адаптивного відновлення викривлених даних, що представлені у вигляді таблиць «об’єкт - властивість» або часових рядів, та адаптивної нечіткої кластеризації є актуальною.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана в рамках держбюджетних НДР №245 «Еволюційні гібридні системи обчислювального інтелекту зі змінною структурою для інтелектуального аналізу даних» (№ДР 0110U000458); НДР №273 «Нейро-фаззі системи для поточної кластеризації та класифікації послідовностей даних за умов їх викривленості відсутніми та аномальними спостереженнями» (№ДР 0113U000361). В рамках зазначених НДР здобувачкою в якості виконавця розроблено методи: адаптивної кластеризації даних, кластеризації даних з пропусками на основі нейро-фаззі мережі Кохонена, запропоновано нейронну мережу, що дозволяє розв’язувати задачу відновлення пропусків в таблицях «об’єкт-властивість» в on-line режимі з постійною корекцією елементів таблиці, що відновлюються.

Мета і задачі дослідження. Метою дисертаційної роботи є розробка методів динамічного інтелектуального аналізу викривлених даних, що послідовно надходять на обробку в on-line режимі. Відповідно до поставленої мети необхідно розв’язати такі наукові задачі:

- аналіз існуючих методів та підходів до відновлення та кластеризації викривлених даних;
- розробка методів адаптивного відновлення пропущених даних в таблицях «об'єкт-властивість» і часових рядах;
- розробка нейро-фаззі методів для відновлення і кластеризації викривлених даних;

– розробка методів адаптивної нечіткої кластеризації даних з пропусками з використанням оптимізаційних процедур;

– імітаційне моделювання та розв’язання практичних завдань.

Об’єкт дослідження – процес динамічного інтелектуального аналізу і обробки даних, представлених у вигляді часових рядів і таблиць «об'єкт - властивість», викривлених пропусками та викидами.

Предмет дослідження – методи динамічного інтелектуального аналізу даних з пропусками.

Методи дослідження. Основними методами дослідження є методи обчислювального інтелекту: динамічний інтелектуальний аналіз даних - для знаходження прихованих залежностей в інформації; методи машинного навчання, за допомогою яких були синтезовані нові моделі і методи навчання нейронних мереж, що дозволяють виконувати відновлення даних; теорія нечіткої кластеризації – для розробки методів кластеризації викривлених і відновлених даних в умовах класів, що перетинаються; імітаційне моделювання - для визначення ефективності застосування розроблених систем.

Наукова новизна отриманих результатів. До нових, одержаних особисто автором, належать такі результати:

1. Вперше запропоновано адаптивну нейро-фаззі систему, що дозволяє розв’язувати задачу відновлення пропусків в таблицях «об’єкт-властивість», що містять апріорі невідому кількість пропусків в on-line режимі з постійною корекцією елементів таблиці, що відновлюються, та центроїдів кластерів та забезпечує високу швидкодію та простоту чисельної реалізації.

2. Вперше запропоновано адаптивну нейронну мережу на основі активаційних функцій спеціального типу, що дозволяє налаштовувати в процесі навчання не тільки синаптичні ваги, а й структуру, забезпечує високу швидкодію та призначена для обробки викривлених нестаціонарних нелінійних стохастичних та хаотичних сигналів, що поступають на обробку в реальному часі.

3. Вперше запропоновано нейро-фаззі методи для відновлення та кластеризації викривлених викидами та пропусками даних в таблицях «об’єкт - властивість» на основі самоорганізовної нейро-фаззі мапи Кохонена, що дозволяє обробляти дані в on-line режимі та забезпечує роботу з класами, що перетинаються.

4. Дістали подальший розвиток методи кластеризації даних з пропусками, що засновані на рекурентній оптимізації спеціального виду цільових функцій, в яких на відміну від існуючих, спостереження замінюються оцінками, що отримані в процесі розв’язання оптимізаційної задачі; методи імовірнісної та можливісної адаптивної нечіткої кластеризації даних з пропусками, що на відміну від існуючих дозволяють опрацьовувати інформацію на основі стратегії найближчого прототипу-центроїда та забезпечують роботу в on-line режимі, а сам процес обробки інформації може бути організовано на основі самоорганізовної мапи Кохонена.

Практичне значення отриманих результатів. Запропоновані методи, орієнтовані на послідовну адаптивну обробку даних представлених в таблицях «об’єкт - властивість» та часових рядах спотворених пропусками та викидами, забезпечують суттєве підвищення якості обробки інформації за умов її дефіциту та

викривленості, оскільки всі відомі аналоги орієнтовні на обробку даних в пакетному режимі.

Методи нечіткої кластеризації даних з пропущеними значеннями були використанні для швидкого діагностування медичної рентгенівської техніки, що вийшла з ладу, в ТОВ Харківського технічного центру рентгенівського сервісу «Спектр» (акт від 19.03.2014), адаптивні методи відновлення даних з пропусками для контролю якості рентген-апаратів, виготовлених на Заводі рентгенівського обладнання «Квант» м. Харків. Тестування рентгенівської техніки, дозволяє виявляти несправності в реальному часі та істотно скоротити терміни діагностування рентгенівської техніки, що ремонтується (акт від 16.04.2014). Отримані теоретичні результати можуть бути використані для інтелектуального аналізу даних та обробки медичної, технічної економічної інформації та ін.

Основні результати дисертаційної роботи використовуються у навчальному процесі Харківського національного університеті радіоелектроніки на кафедрі штучного інтелекту в курсах «Штучні нейронні мережі: архітектури, навчання та застосування» та «Інтелектуальний аналіз даних» (акт від 19.05.2014), та в держбюджетних науково-дослідних роботах згідно з тематичними планами науково-дослідних робіт №273 та № 245 (акт від 12.05.2014).

Особистий внесок здобувача. Внесок авторки в публікаціях, написаних у співавторстві такий: [1] - розроблено адаптивний метод кластеризації даних з пропусками на основі нейро-нечіткої мапи Кохонена; [2] - запропонована адаптивна нейронна мережа для відновлення пропусків в режимі on-line з постійним корегуванням відновлених елементів; [3] - запропонована адаптивна нейронна мережа для відновлення пропусків та кластеризації в режимі on-line с постійним корегуванням відновлених елементів та центроїдів кластерів; [4] - запропоновано адаптивний ймовірнісний метод нечіткої кластеризації даних з пропусками; [5] - запропонована нейронна мережа, що еволюціонує, для відновлення даних з пропусками; [6] - запропоновано методи імовірнісної та можливісної кластеризації даних з пропусками на основі стратегії найближчого прототипу – центроїду; [7] - запропонована адаптивна нейронна мережа для відновлення пропущених спостережень в даних; [8] - запропонована адаптивна нейронна мережа для відновлення пропусків в режимі on-line на основі ансамблів адалін; [9] - запропонована адаптивна нейронна мережа для відновлення даних з пропусками в режимі on-line; [10] – модифіковано адаптивний алгоритм відновлення таблиць даних; [11] - запропоновано метод нечіткої кластеризації даних з пропусками; [12] - розвинений метод адаптивного нечіткого відновлення даних з пропусками в режимі on-line; [13] – розвинений метод кластеризації даних з пропусками на основі нечіткого можливісного підходу; [14] - розвинений метод адаптивної кластеризації даних з пропусками на основі нейро-фаззі підходу; [15] - розглянуто метод навчання нейронної мережі Т. Кохонена на основі ансамблів адалін; [16] - розвинені методи імовірнісної та можливісної нечіткої кластеризації даних; [17] - розвинений метод нечіткої кластеризації даних на основі стратегії найближчого прототипу.

Апробація результатів дисертації. Основні положення та результати дисертаційної роботи були представлені, доповідалися й обговорювалися на

міжнародних наукових конференціях таких як: 15-й, 16-й, 17-й Міжнародних молодіжних форумах «Радіоелектроніка та молодь в XXI столітті» (Харків, Україна, 2011-2013 рр.); Міжнародних наукових конференціях «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (Євпаторія, Україна, 2011-2013 рр.); Міжнародних науково-технічних конференціях «Автоматизація: проблеми, ідеї, рішення» (Севастопіль, Україна, 2011-2012 рр.); International Conference "Computer Science and Information Technologies" (Львів, Україна, 2011р.); International Conference «Artificial Intelligence Methods and Techniques for Business and Engineering Applications» (Rzeszow, Poland, 2012 p.); International Conference "Information, models and analyses" (Varna, Bulgaria, 2013 p.); 20th East West Fuzzy Colloquium (Zittau, Germany, 2013 p).

Публікації. Основні положення дисертаційної роботи опубліковано в 17 наукових роботах: тому числі 1 монографії (розділ), що видано за кордоном, 6 статтях у фахових періодичних виданнях з технічних наук (з них 5 в Україні, серед яких 3 у виданнях, що входять до міжнародних наукометричних баз, та 1 за кордоном) та 10 тез доповідей на міжнародних наукових конференціях.

Структура дисертації. Дисертація складається зі вступу, п'яти розділів, висновків, списку використаних джерел зі 140 найменувань (15 стор.), та 1 додатку (на 5 сторінках). Повний обсяг дисертації складає 166 сторінок, з них 131 сторінка основного тексту, містить 35 рисунків та 12 таблиць (рисунки та таблиці, що займають окрему площу на 11 стор.).

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтована актуальність теми дисертаційної роботи, сформульовано мету і завдання досліджень, наведено відомості щодо публікацій, апробації роботи та особистого внеску здобувача.

У **першому розділі** проведено огляд стану проблеми динамічного аналізу даних, а саме: методи послідовної обробки нестаціонарних сигналів за умов дефіциту поточної інформації та малої вибірки спостережень. Розглянуто та проаналізовано переваги та недоліки відомих нейро-фаззі систем, еволюційних мереж, самоорганізовних мап Кохонена, а також моделі відновлення даних з пропусками. Розглянуто проблеми кластеризації викривлених та відновлених даних, проаналізовано переваги і недоліки розглянутих методів динамічного інтелектуального аналізу даних, який дозволив зробити висновок, що для задач відновлення та кластеризації викривлених даних за умов апріорної та поточної невизначеності, найефективнішими є методи обчислювального інтелекту, насамперед штучні нейронні мережі та нейро-фаззі системи, які навчаються за допомогою оптимізаційних процедур або методів еволюційних обчислень. Проведено аналіз відомих архітектур нейро-фаззі систем, нейронних мереж, що дістали найбільшого поширення в задачах динамічного інтелектуального аналізу даних з пропусками: адаптивна нейро-фаззі система, нейро-фаззі система з налаштованими функціями належності та еволюційні фаззі та нейро-фаззі системи. Показано, що ці архітектури мають свої недоліки та переваги, та містять обмеження при розв'язанні задач, де спостереження надходять на обробку в послідовному

режимі за умов дефіциту апріорної та поточної інформації та за умов короткої вибірки даних.

На основі проведеного аналізу визначено задачі дослідження, що полягають у розробці адаптивних нейро-фаззі методів кластеризації та відновлення викривлених даних, адаптивних нейронних мереж для розв'язання задач динамічного інтелектуального аналізу викривлених даних в таблицях «об'єкт-властивість» та часових рядів, а також методів їх навчання, що враховують особливості задач обробки даних з пропусками та викидами.

Другий розділ присвячено розробці та дослідженню методів відновлення даних з пропусками в таблицях «об'єкт – властивість» та часових рядах, які мають покращені апроксимуючі та екстраполюючі властивості й швидкісні методи навчання, що мають як фільтруючі, так і слідкуючі властивості та дозволяють оброблювати сигнали з пропусками в послідовному (on-line) режимі.

Запропоновані адаптивна нейронна мережа, що заснована на ансамблях Adalines, та адаптивна нейро-фаззі система для відновлення пропусків в таблицях «об'єкт-властивість». При подачі на вхід вузла системи-нео-фаззі нейрона векторного сигналу $\tilde{x}_k = (\tilde{x}_{k1}, \tilde{x}_{k2}, ..., \tilde{x}_{kn})^T$, на його виході з'являється скалярне значення

$$y_k = \sum_{i=1}^{n} f_i(\tilde{x}_{ki}) = \sum_{i=1}^{n} \sum_{l=1}^{h} w_{li}(k-1)\mu_{li}(\tilde{x}_{ki}),$$

яке визначається як синаптичними вагами $w_{li}(k-1)$, так і функціями належності $\mu_{li}(\tilde{x}_{ki})$.

Оцінювання синаптичних ваг може бути проведене як за допомогою стандартного методу найменших квадратів, так і в послідовному режимі, який в даному випадку набуває вигляду

$$\begin{cases} w_j(N_F+1) = w_j(N_F) + r^{-1}(N_F+1)(\tilde{x}_{N+1,j} - \mu(\tilde{\underline{x}}_{N+1,j})w_j(N_F))\mu^T(\tilde{x}_{N+1,j}), \\ r_j(N_F+1) = \alpha r(N_F) + \left\|\mu(\tilde{\underline{x}}_{N+1,j})\right\|^2, \end{cases}$$

де N_F - кількість повністю заповнених рядків; r_j - коефіцієнт кроку навчання, $0 \le \alpha \le 1$ - параметр згладжування.

Запропоновано метод багатовимірної нечіткої екстраполяції, який дозволяє опрацьовувати нестаціонарні викривлені пропусками сигнали, що надходять на обробку послідовно в реальному часі. Для кожного рядка $\tilde{\underline{x}}_i$, що містить пропуски, для оцінки відмінності між ним і всіма іншими рядками використовується поняття «часткової відстані» (PD), прийнятої в нечіткій кластеризації та модифікованої у вигляді

$$D_P^2(\tilde{\underline{x}}_i, \tilde{\underline{x}}_k) = (n/(n_i + n_k - n_{ik})) \sum_{j=1}^{n} (\tilde{x}_{ij} - \tilde{x}_{kj})^2 \delta_j,$$

де $\delta_j = \begin{cases} 0, \text{ якщо у } \tilde{\underline{x}}_i, \text{ або у } \tilde{\underline{x}}_k \text{ в } j \text{ позиції є пропуск,} \\ 1 \text{ в іншому випадку;} \end{cases}$

n_{ik} - кількість загальних пропусків в одній і тій же позиції у $\tilde{\underline{x}}_i$ та $\tilde{\underline{x}}_k$.

Використовуючи поняття належності, прийняте в стандартному методі с-середніх, розраховується рівень близькості $\tilde{\underline{x}}_i$ до $\hat{N}$, у вигляді

$$U_l(i) = \frac{D_P^{-2[l]}}{\sum_{q=1}^{\hat{N}} D_P^{-2[q]}}, \quad l = 1,2,...,\hat{N}.$$

Таким чином, кожний вектор $\tilde{\underline{x}}_i$ апроксимується виразом

$$\hat{\underline{x}}_i = \sum_{l=1}^{\hat{N}} U_l(i)\tilde{\underline{x}}_l.$$

Введений метод відновлення даних з пропусками у часових рядах. В основі запропонованого підходу до відновлення спостережень лежить використання класичних ортогональних поліномів.

Для часового ряду $\tilde{x}_k$, $k = 1,2,...,N$ введено апроксимуючу Т-систему

$$T_h^{\Sigma}(k) = \sum_{l=0}^{h} w_l T_l(k),$$

де $w_l = \frac{\sum_{k=1}^{N} \tilde{y}_k T_l(\tilde{x}_k)}{\sum_{k=1}^{N} T_l^2(\tilde{x}_k)}$, при цьому припускається, що в деякі моменти дискретного часу k вимірювання загублені.

Далі введено до розгляду дві підмножини $X_P = \{\tilde{x}_k, k\}$, які містять $N_P \le N$ спостережень, та $X_G = \{k\}$, які містять моменти часу з відсутніми спостереженнями. Тоді коефіцієнти апроксимуючого полінома можна розрахувати за допомогою співвідношень:

$$w_l = \frac{\sum_{\substack{k=1 \\ k\in X_P}}^{N} \tilde{x}_k T_{lN}(k)}{\sum_{\substack{k=1 \\ k\in X_P}}^{N} T_{lN}^2(k)}, \; w_0 = \frac{1}{N_P} \sum_{\substack{k=1 \\ k\in X_P}}^{N} \tilde{x}_k,$$

де індекс N в T_{lN} означає, що відповідні поліноми ортогональні на інтервалі $k \in [1, N]$. Далі нескладно відновити відсутні спостереження у вигляді

$$\hat{x}_k = \sum_{l=0}^{h} w_l T_{lN}(k) \quad \forall k \in X_G.$$

Якщо ж дані надходять на обробку послідовно, слід організувати on-line обробку інформації, для цих цілей пропонується використовувати рекурентний метод найменших квадратів у формі

$$\begin{cases} w(N) = w(N-1) + \dfrac{P(N-1)(\tilde{x}_N - w^T(N-1)T_N(N))}{1+T_N^T(N)P(N-1)T_N(N)}T_N(N), \\ P(N) = P(N-1) - \dfrac{P(N-1)T_N(N)T_N^T(N)P(N-1)}{1+T_N^T(N)P(N-1)T_N(N)}, \end{cases}$$

де $w=(w_0,w_1,...,w_h)^T$ - вектор параметрів, $T_N(k)=(T_{0N}(k),T_{1N}(k),...,T_{hN}(k))^T$ Т-система ортогональних поліномів.

Однак при цьому необхідно зауважити, що з приходом нового спостереження $\tilde{x}_{N+1}$, істотно змінюється структура апроксимуючих поліномів так, що $T_N(k) \neq T_{N+1}(k)$. Щоб зберегти структуру апроксимуючих поліномів, можна організувати обробку даних на ковзному вікні s. Тоді оцінка з фіксованою структурою поліномів може бути записана у вигляді

$$w_s(N) = \left(\sum_{\substack{\tilde{k}=1, \\ k=N-s+1 \\ k\in X_P}}^{s,N} T_s(\tilde{k})T_s^T(\tilde{k}) \right)^{-1} \sum_{\substack{\tilde{k}=1, \\ k=N-s+1 \\ k\in X_P}}^{s,N} T_s(\tilde{k})\tilde{x}_k = P_s(N) \sum_{\substack{\tilde{k}=1, \\ k=N-s+1 \\ k\in X_P}}^{s,N} T_s(\tilde{k})T_s^T(\tilde{k}).$$

На базі ортогональних поліномів синтезована архітектура ортогональної нейронної мережі, яка має добрі апроксимуючі властивості і високу швидкість налаштування синаптичних ваг та характеризується простотою навчання як синаптичних ваг, так і архітектури. Архітектура ортогональної нейронної мережі для обробки даних з втраченими спостереженнями реалізує нелінійне відображення

$$\hat{y}_k = \hat{f}(\hat{x}_k) = \sum_{i=1}^{n}\sum_{l=0}^{h} w_{li}T_{li}(\hat{x}_{ki})$$

де $k=1,2,...$- поточний дискретний час, w_{li} - синаптичні ваги, що налаштовуються, $T_{li}(\bullet)$ - l-та ортогональна активаційна функція для вхідного сигналу $\hat{x}_{ki}$, $i=1,2,...,n$.

Наведена архітектура містить $2n+1$ орто-синапсів і $(h+1)(2n+1)$ синаптичних ваг, що підлягають визначенню при цьому, що дуже важливо, вихідний сигнал $\hat{y}_k$ лінійно залежить від ваг w_{li}. Кількість активаційних функцій $h+1$ в кожному орто-синапсі $O-S_i$ вихідного шару вибирається досить довільно, тому, якщо буде встановлено, що синтезована нейронна мережа не забезпечує необхідну якість обробки інформації, кількість цих функцій може бути збільшено (або зменшено, якщо необхідно) в on-line режимі, безпосередньо в процесі навчання.

Таким чином мережа набуває еволюційні властивості, адаптуючи свою структуру. Оцінка $w(N)$ повинна бути скорегована таким чином:

$$w^*(N) = \left(\begin{array}{c|c} \displaystyle\sum_{k=N-s+1}^{N} T(\hat{x}_k)T^T(\hat{x}_k) & R_{l,h+1}(\hat{x}_k) \\ \hline R_{l,h+1}^T(\hat{x}_k) & R_{h+1,h+1}(\hat{x}_k) \end{array} \right)^{-1} \left(\begin{array}{c} \displaystyle\sum_{k=N-s+1}^{N} T(\hat{x}_k)y_k^* \\ \hline r_{h+1}(\hat{x}_k) \end{array} \right),$$

де $((h+1)n\times n)$- матриця $R_{l,h+1}(\hat{x}_k)$ утворена елементами $\sum_{k=N-s+1}^{N} T_{li}(\hat{x}_{ki})T_{h+1,j}(\hat{x}_{kj})$, $l=0,1,...,h;\ i=1,2,...,n;\ j=1,2,...,n;$ $(n\times n)$- матриця $R_{h+1,h+1}(\hat{x}_k)$, утворена елементами $\sum_{k=N-s+1}^{N} T_{h+1,i}(\hat{x}_{ki})T_{h+1,j}(\hat{x}_{kj});$ $(n\times 1)$ - стовпець $r_{h+1,i}(\hat{x}_{ki})$, утворений елементами $\sum_{k=N-s+1}^{N} T_{h+1,i}(\hat{x}_{ki})y_k^*$.

У **третьому розділі** розглянуто задачі кластеризації даних з пропусками, що містяться в таблиці «об'єкт - властивість» в режимі on-line.

Сама кластеризація, реалізується шляхом мінімізації цільової функції

$$E(U_q(k),w_q)=\sum_{k=1}^{N}\sum_{q=1}^{m}U_q^{\beta}(k)D^2(\tilde{x}_k,w_q)$$

при обмеженнях

$$\sum_{q=1}^{m}U_q(k)=1,\ k=1,...,N,\ 0\le\sum_{k=1}^{N}U_q(k)\le N, q=1,...,m,$$

де $U_q(k)\in[0,1]$ - рівень належності вектора $\tilde{x}_k$ до q-го кластера; w_q - прототип (центроїд) q-го кластера; β - фаззіфікатор (зазвичай $\beta=2$); $D^2(\tilde{x}_k,w_q)$ - відстань між $\tilde{x}_k$ і w_q у прийнятій метриці.

Для роботи в on-line режимі, коли дані надходять на обробку послідовно, пропонується використовувати адаптивну модифікацію FCM алгоритму

$$\begin{cases} w_q(k)=w_q(k+1)+\eta(k)U_q^2(k+1)(\tilde{x}_k+w_q(k+1)), \\ U_q(k)=\left\|\tilde{x}_k-w_q(k)\right\|^{-2}\Big/\sum_{l=1}^{m}\left\|\tilde{x}_k-w_l(k)\right\|^{-2}, \end{cases}$$

де $\eta(k)$- параметр шагу навчання.

В тому випадку, коли дані надходять на обробку послідовно і мають пропуски, запропонована модифікація FCM процедури, яка заснована на стратегії часткових відстаней (PDS FCM). Беручи замість традиційної евклідової метрики часткову відстань (PD)

$$D_P^2(\tilde{x}_k,w_q)=\frac{n}{\delta_{k\Sigma}}\sum_{i=1}^{n}(\tilde{x}_{ki}-w_{qi})^2\delta_{ki},$$

цільову функцію кластеризації

$$E(U_q(k),w_q)=\sum_{k=1}^{N}\sum_{q=1}^{m}U_q^{\beta}(k)\frac{n}{\delta_{k\Sigma}}\sum_{i=1}^{n}(\tilde{x}_{ki}-w_{qi})^2\delta_{ki}$$

і вирішуючи задачу нелінійного програмування, приходимо до методу, що може бути записаний у вигляді

$$\begin{cases} U_q(k+1) = (D_P^2(\tilde{x}_{k+1}, w_q(k)))^{\frac{1}{1-\beta}} \Big/ \sum_{l=1}^{m} (D_P^2(\tilde{x}_{k+1}, w_q(k)))^{\frac{1}{1-\beta}}, \\ w_{qi}(k+1) = w_{qi}(k) + \eta(k+1) U_q^{\beta}(k+1)(\tilde{x}_{k+1,i} - w_{qi}(k))\delta_{ki}, \end{cases}$$

де $\delta_k = (\delta_{k1}, ..., \delta_{kn})^T$.

Основний недолік ймовірнісних алгоритмів нечіткої кластеризації пов'язаний з обмеженням на суму належностей, яка повинна дорівнювати одиниці. У можливісних алгоритмах цільова функція кластеризації має вигляд

$$E(U_q(k), w_q, \mu_q) = \sum_{k=1}^{N}\sum_{q=1}^{m} U_q^{\beta}(k) D^2(\tilde{x}_k, w_q) + \sum_{q=1}^{m} \mu_q \sum_{k=1}^{N} (1 - U_q(k))^{\beta},$$

де скалярний параметр $\mu \geq 0$ визначає відстань, на якому рівень належності приймає значення 0.5, таким чином якщо $D^2(\tilde{x}_k, w_q) = \mu_q$, то $w_q(k) = 0{,}5$.

У режимі on-line обробки інформації співвідношення для можливісної кластеризації, можуть бути представлені у вигляді

$$\begin{cases} U_q(k+1) = 1 \Big/ 1 + \dfrac{\left\| \tilde{x}_k - w_q(k) \right\|^2}{\mu_q(k)}, \\ w_q(k+1) = w_q(k) + \eta(k+1) U_q^2(k+1)(\tilde{x}_{k+1} - w_q(k)), \\ \mu_q(k+1) = \sum_{p=1}^{k+1} U_q^2(p) \left\| \tilde{x}_p - w_q(k+1) \right\|^2 \Big/ \sum_{p=1}^{k} U_q^2(p). \end{cases}$$

Беручи замість евклідової метрики часткову відстань (PD), введено цільову функцію кластеризації у вигляді

$$E(U_q(k), w_q, \mu_q) = \sum_{k=1}^{N}\sum_{q=1}^{m} U_q^{\beta}(k) \frac{n}{\delta_{k\sum}} \sum_{i=1}^{n} (\tilde{x}_{ki} - w_{qi})^2 \delta_{ki} + \sum_{q=1}^{m} \mu_q \sum_{k=1}^{N} (1 - U_q(k))^{\beta},$$

та процедуру, яка має рекурентну форму

$$\begin{cases} U_q(k) = 1 \Big/ 1 + (D_P^2(\tilde{x}_{k+1}, w_q(k)) \big/ \mu_q(k))^{\frac{1}{\beta-1}}, \\ w_{qi}(k+1) = w_{qi}(k) + \eta(k+1) U_q^{\beta}(k+1)(\tilde{x}_{k+1,i} - w_{qi}(k))\delta_{ki}, \\ \mu_q(k+1) = \sum_{p=1}^{k+1} U_q^{\beta}(p) D_P^2(\tilde{x}_p, w_q(k+1)) \Big/ \sum_{p=1}^{k+1} U_q^{\beta}(p). \end{cases}$$

Таким чином, процес нечіткої можливісної кластеризації даних з пропусками може бути реалізований за допомогою нейро-фаззі мережі Кохонена.

Далі, вводячи в розгляд робастну цільову функцію, яка заснована на мірі подібності

$$E_S(U_q(k), w_q) = \sum_{k=1}^{N}\sum_{q=1}^{m} U_q^{\beta}(k) S(\tilde{x}_k, w_q) = \sum_{k=1}^{N}\sum_{q=1}^{m} \frac{U_q^{\beta}(k)\sigma^2}{\sigma^2 + \left\|\tilde{x}_k - w_q\right\|^2},$$

$$S(\tilde{x}_k, w_q) = \frac{1}{1 + \dfrac{\left\|\tilde{x}_k - w_q\right\|^2}{\sigma^2}} = \frac{\sigma^2}{\sigma^2 + \left\|\tilde{x}_k - w_q\right\|^2} = \frac{\sigma^2}{\sigma^2 + D^2(\tilde{x}_k, w_q)},$$

де σ^2 параметр ширини функції, імовірнісні обмеження

$$\sum_{q=1}^{m} U_q(k) = 1,$$

функцію Лагранжа

$$L_S(U_q(k), w_q, \lambda(k)) = \sum_{k=1}^{N}\sum_{q=1}^{m} \frac{U_q^{\beta}(k)\sigma^2}{\sigma^2 + \left\|\tilde{x}_k - w_q\right\|^2} + \sum_{k=1}^{N} \lambda(k)\left(\sum_{q=1}^{m} U_q(k) - 1\right)$$

(тут $\lambda(k)$ - невизначені множники Лагранжа),

вирішуючи систему рівнянь Каруша-Куна-Таккера, приходимо до рішення

$$\begin{cases} U_q(k+1) = \dfrac{(S(\tilde{x}_{k+1}, w_q(k)))^{\frac{1}{\beta-1}}}{\sum\limits_{l=1}^{m}(S(\tilde{x}_{k+1}, w_l(k)))^{\frac{1}{\beta-1}}}, \\ w_q(k+1) = w_q(k) + \eta(k+1) U_q^{\beta}(k+1)\, \tilde{x}_{k+1} - w_q(k) \Big/ (\sigma^2 + \left\|\tilde{x}_{k+1} - w_q(k)\right\|^2)^2 = \\ = w_q(k) + \eta(k+1)\varphi_q(k+1)(\tilde{x}_{k+1} - w_q(k)), \end{cases}$$

де $\varphi_q(k+1) = \tilde{x}_{k+1} - w_q(k) \Big/ (\sigma^2 + \left\|\tilde{x}_{k+1} - w_q(k)\right\|^2)^2$ - функція сусідства робастного WTM-правила самонавчання.

Для вирішення задачі робастної кластеризації даних з пропусками введемо в розгляд часткову міру подібності (PCM), що є гібридом часткового відстані (PD) і міри подібності (SM) та має вигляд

$$S_P(\tilde{x}_k, w_q) = \frac{\sigma^2}{\sigma^2 + D_P^2(\tilde{x}_k, w_q)}.$$

Таким чином, використання часткової міри подібності, заснованої на частковій відстані, дозволяє вирішувати завдання нечіткої кластеризації даних, що містять як пропуски, так і аномальні спостереження.

$$
\begin{cases}
U_q^{(\tau+1)}(k) = \dfrac{1}{1+\left(\dfrac{S^{-1}(\hat{x}_k, w_q^{(\tau)}(k))}{\mu_q^{(\tau)}(k)}\right)^{\frac{1}{\beta-1}}}, \\
\hat{x}_{ki}^{(\tau)} = w_{qi}^{(\tau)},\ w_q^{(\tau)}(k) = \arg\max\limits_q \{S_P(\tilde{x}_k^{(\tau)}, w_1^{(\tau)}(k)), ..., S_P(\tilde{x}_k^{(\tau)}, w_m^{(\tau)}(k))\}, \\
w_q^{(Q)}(k) = w_q^{(0)}(k+1), \\
w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1)\dfrac{(U_q^{(Q)}(k))^\beta}{(\sigma^2 + \left\|\hat{x}_{k+1} - w_q^{(\tau)}(k+1)\right\|^2)^2}(\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)), \\
\mu_q^{(\tau+1)}(k) = \dfrac{\sum\limits_{p=1}^{k}(U_q^{(\tau+1)}(p))^\beta S_P^{-1}(\hat{x}_p, w_q^{(\tau+1)}(k))}{\sum\limits_{p=1}^{k}(U_q^{(\tau)}(p))^\beta},
\end{cases}
$$

де τ - прискорений машинний час.

Четвертий розділ присвячено розробці методів адаптивної нечіткої кластеризації викривлених даних з використанням оптимізаційних процедур. Запропоновано можливісні та ймовірнісні методи нечіткої кластеризації даних, які засновані на стратегії оптимального розширення та стратегії найближчого прототипу – центроїда.

В ситуаціях, коли кількість пропусків занадто велика, стратегія часткових відстаней може виявитися неефективною, у зв'язку з чим може виникнути необхідність поряд з вирішенням власне завдання нечіткої кластеризації одночасно відновлювати відсутні спостереження. У цій ситуації ефективним виявляється підхід, в основі якого лежить, так звана, стратегія оптимального розширення (OCS FCM). У зв'язку з цим цікава сама задача on-line кластеризації даних з використанням стратегії оптимального розширення, адаптованої на випадок, коли інформація оброблюється в послідовному режимі, а її обсяг заздалегідь не визначений.

Стратегія оптимального розширення полягає в тому, що елементи підмассива X_G розглядаються як додаткові змінні, за якими також проводиться оптимізація прийнятої цільової функції E. У цьому випадку пакетний метод кластеризації даних з пропусками на основі стратегії оптимального розширення може бути записаний у вигляді такої послідовності кроків:

1. Завдання початкових умов для роботи методу: $\beta > 0$; $1 < m < N$; $\varepsilon > 0$; $w_q^{(0)}$; $1 \le q \le m$; $\tau = 0,1,2,...,Q$; $X_G^{(0)} = \{-1 \le \hat{x}_{ki}^{(0)} \le 1\}$,

де $X_G^{(0)} - N_G (1 \le N_G \le (n-1)N)$ початкової оцінки $\hat{x}_{ki}^{(0)}$ відсутніх спостережень $\tilde{x}_{ki} \in X_G$.

2. Розрахунок рівней належності шляхом вирішення задачі оптимізації:

$$U_q^{(\tau+1)}(k) = \arg\min_{U_q(k)} E(U_q(k), w_q^{(\tau)}, X_G^{(\tau)}) = \frac{(D^2(\hat{x}_k^{(\tau)}, w_q^{(\tau)}))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}(D^2(\hat{x}_k^{(\tau)}, w_l^{(\tau)}))^{\frac{1}{1-\beta}}} = \frac{(\left\|\hat{x}_k^{(\tau)} - w_q^{(\tau)}\right\|^2)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}(\left\|\hat{x}_k^{(\tau)} - w_l^{(\tau)}\right\|^2)^{\frac{1}{1-\beta}}},$$

(тут вектор $\hat{x}_k^{(\tau)}$ відрізняється від $\tilde{x}_k$ тим, що відсутні спостереження $\tilde{x}_{ki} \in X_G$ замінені оцінками $\hat{x}_{ki}^{(\tau)}$, що розраховані на τ-ой епосі обробки даних).

3. Розрахунок центроїдів кластерів шляхом вирішення задачі оптимізації:

$$w_q^{(\tau+1)} = \arg\min_{w_q} E(U_q^{(\tau+1)}(k), w_q, X_G^{(\tau)}) = \frac{\sum_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}\hat{x}_k^{(\tau)}}{\sum_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}}$$

4. Перевірка умов останова: якщо $\left\|w_q^{(\tau+1)} - w_q^{(\tau)}\right\| < \varepsilon \ \forall\ 1 \le q \le m$ або $\tau = Q$, метод закінчує роботу, інакше йти до кроку 5.

5. Відновлення пропущених спостережень шляхом вирішення задачі оптимізації

$$X_G^{(\tau+1)} = \arg\min_{X_G} E(U_q^{(\tau+1)}(k), w_q^{(\tau+1)}, X_G),$$

звідки витікає

$$\hat{x}_{ki}^{(\tau+1)} = \sum_{q=1}^{m}(U_q^{(\tau+1)}(k))^{\beta} w_{qi}^{(\tau+1)} \Big/ \sum_{q=1}^{m}(U_q^{(\tau+1)}(k))^{\beta}.$$

Використовуючи стратегію оптимального розширення, в якості цільової функції можливісної нечіткої кластеризації використовується вираз

$$E(U_q(k), w_q, \mu_q, X_G) = \sum_{k=1}^{N}\sum_{q=1}^{m} U_q^{\beta}(k) D^2(\hat{x}_k, w_q) + \sum_{q=1}^{m}\mu_q \sum_{k=1}^{N}(1 - U_q(k))^{\beta},$$

а його мінімізація веде до системи співвідношень

$$\begin{cases} U_q^{(\tau+1)}(k) = 1 \Big/ 1 + (D^2(\hat{x}_k^{(\tau)}, w_q^{(\tau)}) \Big/ \mu_q^{(\tau)})^{\frac{1}{\beta-1}}, \\ w_q^{(\tau+1)} = \sum_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}\hat{x}_k^{(\tau)} \Big/ \sum_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}, \\ \hat{x}_{ki}^{(\tau+1)} = \sum_{q=1}^{m}(U_q^{(\tau+1)}(k))^{\beta} w_{qi}^{(\tau+1)} \Big/ \sum_{q=1}^{m}(U_q^{(\tau+1)}(k))^{\beta}, \\ \mu_q^{(\tau+1)} = \sum_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta} D^2(\hat{x}_k^{(\tau+1)}, w_q^{(\tau+1)}) \Big/ \sum_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}. \end{cases}$$

Аналогічно рекурентній нечіткій ймовірнісній кластеризації на основі стратегії оптимального розширення можна організувати процес можливісної

кластеризації за допомогою процедури

$$\begin{cases} U_q^{(\tau+1)}(k+1) = \dfrac{1}{1+\left(\dfrac{\left\|\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)\right\|^2}{\mu_q^{(\tau)}}\right)^{\frac{1}{\beta-1}}}, \\ w_q^{(0)}(k+1) = w_q^{(Q)}(k), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1)(U_q^{(\tau+1)}(k+1))^{\beta}(\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)), \\ \hat{x}_{k+1,i}^{(\tau+1)} = \dfrac{\sum\limits_{q=1}^{m}(U_q^{(\tau+1)}(k+1))^{\beta} w_{qi}^{(\tau+1)}(k+1)}{\sum\limits_{q=1}^{m}(U_q^{(\tau+1)}(k+1))^{\beta}}, \\ \mu_q^{(\tau+1)}(k+1) = \dfrac{\sum\limits_{p=1}^{k+1}(U_q^{(\tau+1)}(p))^{\beta}\left\|\hat{x}_p^{(\tau+1)} - w_q^{(\tau+1)}(k+1)\right\|^2}{\sum\limits_{p=1}^{k+1}(U_q^{(\tau+1)}(p))^{\beta}}. \end{cases}$$

До переваг цього методу можна віднести те, що з його допомогою можна в on-line режимі виявляти появи нових кластерів.

Стратегія найближчого прототипу-центроїда може бути розглянута в якості гібрида стратегії оптимального розширення та часткових відстаней і складається з послідовності кроків:

1. Завдання початкових умов для роботи методу: $\beta > 0$; m, необхідної точності $\varepsilon > 0$, прототипів (центроїдів) кластерів $w_q^{(0)}$, кількості епох обробки $\tau \le Q$, $X_G^{(0)} = \{-1 \le \hat{x}_{ki}^{(0)} \le 1\} - N_G$ довільних оцінок відсутніх спостережень $\tilde{x}_{ki} \in X_G$.

2. Розрахунок рівнів належності:

$$U_q^{(\tau+1)}(k) = \left(\sum_{l=1}^{m}\left(\left\|\hat{x}_k^{(\tau)} - w_l^{(\tau)}\right\|^2\right)^{\frac{1}{1-\beta}}\right)^{-1}\left(\left\|\hat{x}_k^{(\tau)} - w_q^{(\tau)}\right\|^2\right)^{\frac{1}{1-\beta}}.$$

3. Розрахунок центроїдів кластерів:

$$w_q^{(\tau+1)} = \left(\sum_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}\right)^{-1}\sum_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}\hat{x}_k^{(\tau)}.$$

4. Перевірка умов останова: якщо $\left\|w_q^{(\tau+1)} - w_q^{(\tau)}\right\| < \varepsilon \ \forall\ q$ або $\tau = Q$, останов; інакше йти до кроку 5.

5. Оцінка відсутніх спостережень шляхом знаходження прототипу $w_q^{(\tau+1)}$ найближчого до $\tilde{x}_k$ в сенсі часткової відстані

$$D_P^2(\tilde{x}_k, w_q) = \frac{n}{\delta_{k\sum}}\sum_{i=1}^{n}(\tilde{x}_{ki} - w_{qi})^2\delta_{ki},$$

тобто знаходження $w_q^{(\tau+1)} = \arg\min\limits_q \{D_P^2(\tilde{x}_k, w_1^{(\tau+1)}), ..., D_P^2(\tilde{x}_k, w_m^{(\tau+1)})\}$ і заміна відсутніх спостережень $\tilde{x}_{ki}$ координатами $\hat{x}_{ki}^{(\tau+1)} = w_{qi}^{(\tau+1)}$. Далі йти до кроку 2.

Далі можна записати стратегію найближчого прототипу у рекурентній формі

$$\begin{cases} U_q^{(\tau+1)}(k) = (\sum\limits_{l=1}^{m} (\left\| \hat{x}_k^{(\tau)} - w_l(k) \right\|^2)^{\frac{1}{1-\beta}})^{-1} (\left\| \hat{x}_k^{(\tau)} - w_q(k) \right\|^2)^{\frac{1}{1-\beta}}, \\ \text{де } \hat{x}_{ki}^{(\tau)} = w_{qi}(k), \;\; w_q(k) = \arg\min\limits_q \{D_P^2(\tilde{x}_k, w_1(k)), ..., D_P^2(\tilde{x}_k, w_m(k))\}, \\ w_q(k+1) = w_q(k) + \eta(k+1)(U_q^{(Q)}(k))^{\beta} (\hat{x}_k^{(Q)} - w_q(k)) \;\; \forall q = 1, 2, ..., m. \end{cases}$$

Можливісна стратегія найближчого прототипу-центроїда у загублених спостереженнях може бути записана у вигляді послідовності кроків:

1. Завдання початкових умов для роботи методу: $\beta > 0$; m, необхідної точності $\varepsilon > 0$, прототипів (центроїдів) кластерів $w_q^{(0)}$, кількості епох обробки $\tau \le Q$, $X_G^{(0)} = \{-1 \le \hat{x}_{ki}^{(0)} \le 1\} - N_G$ довільних оцінок відсутніх спостережень $\tilde{x}_{ki} \in X_G$.

2. Розрахунок рівнів належності:

$$U_q^{(\tau+1)}(k) = 1 \Big/ 1 + (\left\| \hat{x}_k^{(\tau)} - w_q^{(\tau)} \right\|^2 \Big/ \mu_q^{(\tau)})^{\frac{1}{\beta-1}}.$$

3. Розрахунок центроїдів кластерів:

$$w_q^{(\tau+1)}(k) = \sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta} \hat{x}_k^{(\tau)} \Big/ \sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta}.$$

4. Перевірка умов останова: якщо $\left\| w_q^{(\tau+1)} - w_q^{(\tau)} \right\| < \varepsilon \;\; \forall \; q$ або $\tau = Q$, останов; інакше йти до шага 5.

5. Оцінка відсутніх спостережень шляхом знаходження прототипу $w_q^{(\tau+1)}$ найближчого до $\tilde{x}_k$ в сенсі часткової відстані

$$D_P^2(\tilde{x}_k, w_q) = \frac{n}{\delta_{k\Sigma}} \sum_{i=1}^{n} (\tilde{x}_{ki} - w_{qi})^2 \delta_{ki},$$

тобто знаходження $w_q^{(\tau+1)} = \arg\min\limits_q \{D_P^2(\tilde{x}_k, w_1^{(\tau+1)}), ..., D_P^2(\tilde{x}_k, w_m^{(\tau+1)})\}$ і заміна відсутніх спостережень $\tilde{x}_{ki}$ координатами $\hat{x}_{ki}^{(\tau+1)} = w_{qi}^{(\tau+1)}$.

6. Розрахунок скалярного параметра відстані

$$\mu_q^{(\tau+1)} = \sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta} \left\| \hat{x}_k^{(\tau+1)} - w_q^{(\tau+1)} \right\|^2 \Big/ \sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta}.$$

7. Далі йти до кроку 2.

Аналогічно ймовірнісній адаптивній кластеризації на основі стратегії найближчого центроїда можна організувати процес можливісної кластеризації у вигляді

$$\begin{cases} U_q^{(\tau+1)}(k) = 1 \Big/ 1 + (\left\| \hat{x}_k^{(\tau)} - w_q(k) \right\|^2 \Big/ \mu_q^{(\tau)})^{\frac{1}{\beta-1}}, \\ \text{де } \hat{x}_{ki}^{(\tau)} = w_{qi}(k), \;\; w_q(k) = \arg\min\limits_q \{D_P^2(\tilde{x}_k, w_1(k)), ..., D_P^2(\tilde{x}_k, w_m(k))\}, \\ w_q(k+1) = w_q(k) + \eta(k+1)(U_q^{(Q)}(k))^\beta (\hat{x}_k^{(Q)} - w_q(k)) \;\; \forall q = 1, 2, ..., m, \\ \mu_q^{(\tau+1)} = \sum\limits_{p=1}^{k} (U_q^{(\tau+1)}(p))^\beta \left\| \hat{x}_k^{(\tau)} - w_q(k) \right\|^2 \Big/ \sum\limits_{p=1}^{k} (U_q^{(\tau+1)}(p))^\beta . \end{cases}$$

У п’ятому розділі викладено результати проведених експериментальних досліджень та їх використання для розв’язання практичних задач динамічного інтелектуального аналізу даних з пропусками. Проведені експериментальні дослідження щодо моделювання методів відновлення та кластеризації викривлених пропусками і викидами даних, виконана експериментальна оцінка похибок відновлених даних та її змінення при роботі в режимі on-line, коли дані надходять на обробку послідовно в реальному часі, наведено таблиці зміни похибок та продемонстровано у вигляді графіків, наведено таблиці результатів якості кластеризації відновлених даних за основними критеріями та проведено порівняльний аналіз роботи запропонованих методів з відомими. Продемонстровані також результати впровадження методів, запропонованих в дисертаційній роботі з метою їх практичного використання. Розв’язано задачу виявлення прихованих залежностей й інтелектуальної обробки даних. Результатом динамічного інтелектуального аналізу та обробки викривлених даних, наданих заводом рентгенівського обладнання та сервісним центром, були відновлені пошкоджені дані. Застосування запропонованих методів динамічного інтелектуального аналізу даних з пропусками показали свою ефективність для визначення якості виготовленого заводом рентгенівського обладнання та надали можливість швидко проводити тестування. Розв’язано задачу відновлення викривлених пропусками та викидами даних, якими оперує сервісний центр рентгенівського обладнання, що дало можливість пришвидшити роботу відновлення обладнання, що вийшло з ладу, а також надало можливість завчасно ідентифікувати можливу несправність та надати інформацію. Результати впроваджені та підтверджено відповідними актами.

У додатку наведені документи щодо впровадження результатів досліджень та їх практичного використання.

ВИСНОВКИ

У дисертаційній роботі представлені результати, які відповідно до поставленої мети є вирішенням актуальної науково-практичної задачі розробки методів динамічного інтелектуального аналізу викривлених даних, що містять як пропуски, так і аномальні спостереження, які базуються на самонавчанних нейро-фаззі моделях і системах. Отримані результати мають важливе практичне значення для створення систем відновлення і кластеризації даних, які надходять на обробку послідовно в реальному часі. При проведенні наукових досліджень отримані такі основні результати:

1. Розроблено адаптивну нейро-фаззі систему, що дозволяє розв'язувати задачу відновлення пропусків в таблицях «об'єкт-властивість», що містять апріорі невідому кількість пропусків в on-line режимі з постійною корекцією елементів таблиці, що відновлюються, та центроїдів кластерів та забезпечує високу швидкодію та простоту чисельної реалізації.

2. Розроблено адаптивну нейронну мережу на основі ортогональних активаційних функцій, що дозволяє налаштовувати в процесі навчання не тільки синаптичні ваги, а й структуру, забезпечує високу швидкодію та призначена для обробки викривлених нестаціонарних нелінійних стохастичних та хаотичних сигналів, що поступають на обробку в реальному часі.

3. Розроблено нейро-фаззі методи для відновлення та кластеризації викривлених викидами та пропусками даних в таблицях «об'єкт - властивість» на основі самоорганізовної нейро - фаззі мапи Кохонена, що дозволяє обробляти дані в on-line режимі, забезпечує роботу з класами, що перетинаються.

4. Дістали подальший розвиток методи кластеризації даних з пропусками, що засновані на рекурентній оптимізації спеціального виду цільових функцій, в яких на відміну від існуючих, спостереження замінюються оцінками, що отримані в процесі розв'язання оптимізаційної задачі; методи імовірнісної та можливісної адаптивної нечіткої кластеризації даних з пропусками, що на відміну від існуючих дозволяють опрацьовувати інформацію на основі стратегії найближчого прототипу-центроїда та забезпечують роботу в on-line режимі, а сам процес обробки інформації може бути організовано на основі самоорганізовної мапи Кохонена.

5. Проведено імітаційне моделювання методів відновлення та кластеризації викривлених даних, виконана експериментальна оцінка похибок відновлених даних та її змінення при роботі в режимі on-line. Розв'язана прикладна задача відновлення викривлених даних, наданих Заводом рентгенівського обладнання та сервісним центром. Підвищена швидкість тестування якості зробленого на заводі рентгенівського обладнання за допомогою запропонованих методів. Розв'язано задачу відновлення викривлених пропусками та викидами даних, якими оперує сервісний центр рентгенівського обладнання, що дало можливість пришвидшити роботу відновлення обладнання, яке вийшло з ладу, а також надало можливість завчасно ідентифікувати можливу несправність та надати інформацію. Основні результати дисертаційної роботи використовуються у навчальному процесі та в держбюджетних науково-дослідних роботах.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Bodyanskiy, Ye. Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network / Ye. Bodyanskiy, A. Shafronenko, V. Volkova // Artificial Intelligence Methods and Techniques for Business and Engineering Applications. – ITHEA 2012, Rzeszow, Poland; Sofia, Bulgaria. – P. 287-296 (розділ монографії). (Входить до міжнародних наукометричних баз Data Base Intellectualization (IWGDBI), ITHEA Bibliographic Database, Google Scholar, Cite Seer, DBLP.)

2. Плісс, І.П. Нейромережеве відновлення пропусків у таблицях даних / І.П. Плісс, А.Ю. Шевякова (А.Ю. Шафроненко), Ю.Ю. Шевякова // Наукові праці:

науково-методичний журнал. – Миколаїв: Вид-во ЧДУ ім. Петра Могили, 2011. – Вип.148. Т.160. Комп'ютерні технології. - С.59-61.

3. Шафроненко, А.Ю. Адаптивная кластеризация данных с пропущенными значениями / А.Ю. Шафроненко, В.В. Волкова, Е.В. Бодянский // Радіоелектроніка. Інформатика. Управління .-2011. - №2(25). - С.115-119. (Входить до міжнародних наукометричних баз Index Copernicus, INSPEC, INIS, EBSCO, DOI, Ulrich's.)

4. Bodyanskiy, Ye. Adaptive fuzzy probabilistic clustering of incomplete data/ Ye. Bodyanskiy, A. Shafronenko, V. Volkova // Int. J. "Information, models and analyses". – 2013.-Vol.2, № 2. - P. 112-117. (Входить до міжнародних наукометричних баз Data Base Intellectualization (IWGDBI), ITHEA Bibliographic Database, Google Scholar, Cite Seer, DBLP.)

5. Shafronenko, A. The evolving adaptive neural network for data processing with missing observations / A. Shafronenko, I.Pliss, Ye. Bodyanskiy // Радіоелектроніка. Інформатика. Управління. - 2013. - №2(29). - С.119-125. (Входить до міжнародних наукометричних баз Index Copernicus, INSPEC, INIS, EBSCO, DOI, Ulrich's.)

6. Bodyanskiy Ye. Adaptive fuzzy clustering for data with missing values based on the nearest prototype - centroid strategy / Ye. Bodyanskiy, A. Shafronenko // Вісник національного університету «Львівська політехніка». – 2013. - №771. – С.309-315. (Входить до міжнародної наукометричної баз INSPEC.)

7. Волкова, В.В. Нечітка кластеризації масивів даних з пропущеними значеннями / В.В. Волкова, А.Ю. Шафроненко // Збірник наукових праць «Індуктивне моделювання складних систем». – 2011. - № 3. - С. 27-32.

8. Шевякова, А.Ю. (Шафроненко, А.Ю.) Нейросетевая обработка таблиц с пропусками / А.Ю Шевякова. (А.Ю. Шафроненко), М.З. Стольникова // 15 Международный молодежнй форум «Радиоэлектроника и молодежь в 21 веке», 18-20 апреля 2011 г.: матер. конф. – Харьков, 2011. - Том 9. – С.51-52.

9. Шевякова А.Ю. (Шафроненко, А.Ю.) Адаптивная обработка данных с пропусками / А.Ю. Шевякова (А.Ю. Шафроненко), Ю.Ю. Шевякова, М.З. Стольникова // Международная научная конференция «Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта», 17-21 мая 2011 г.: матер. конф. – Евпатория, 2011. - Том 2. - С. 250-253.

10. Шевякова, А.Ю. (Шафроненко, А.Ю.) Модифицированный адаптивный алгоритм восстановления таблиц данных / А.Ю. Шевякова (А.Ю. Шафроненко), Ю.Ю. Шевякова, М.З. Стольникова // Международная научно-техническая конференция «Автоматизация: проблемы, идеи, решения», 5-9 сентября 2011 г.: матер. конф. – Севастополь, 2011. - С. 278-279.

11. Shafronenko, A. Adaptive fuzzy clustering of data with gaps / A. Shafronenko, Y. Shevyakova, V. Volkova,I. Pliss // Proceedings of the International Conference "Computer Science and Information Technologies", 16-19 November 2011. – Lviv, 2011. – P. 158-159.

12. Шафроненко, А.Ю. Адаптивное нечеткое восстановление данных с пропусками / А.Ю. Шафроненко // 16 Международный молодежнй форум «Радиоэлектроника и молодежь в 21 веке», 17-19 апреля 2012 г.: матер. конф. – Харьков, 2012. - С.43-44.

13. Шафроненко, А.Ю. Кластеризация данных с пропусками на основе нечеткого возможностного похода / А.Ю. Шафроненко, И.П. Плисс, Е.В. Бодянский // Международная научно-техническая конференция «Автоматизация: проблемы, идеи, решения», 3-7 сентября 2012 г.: матер. конф. – Севастополь, 2012. - С. 246-247.

14. Шафроненко, А.Ю. Адаптивная кластеризация данных с пропусками на основе нейро-фаззи похода / А.Ю. Шафроненко, И.П. Плисс, Е.В Бодянский // Международная научная конференция «Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта», 27-31 мая 2012г.: матер. конф. – Евпатория, 2012. – С.431-433.

15. Шафроненко, А.Ю. Обучение нейронной сети Т.Кохонена в задаче обработки данных с пропусками / А.Ю. Шафроненко // 17 Международный молодежнй форум «Радиоэлектроника и молодежь в 21 веке», 22-21 апреля 2013г.: матер. конф. – Харьков, 2013. - Том 6. - С.67-68.

16. Bodyanskiy, Ye. Neuro fuzzy Kohonen network for incomplete data clustering using optimal completion strategy / Ye. Bodyanskiy, A. Shafronenko, V. Volkova // Proceedings 20th East West Fuzzy Colloquium 2013, 25-27 September 2013: – Zittau, 2013. – P.214-223.

17. Бодянский, Е.В. Нечеткая кластеризация данных с пропусками с помощью нейро-фаззи сети Кохонена на основе стратегии ближайшего прототипа / Е.В. Бодянский, И.П. Плисс, А.Ю. Шафроненко // Международная научная конференция «Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта», 20-24 мая 2013 г.: матер. конф. – Евпатория, 2013. – С.417-418.

АНОТАЦІЯ

Шафроненко А. Ю. Методи динамічного інтелектуального аналізу даних з пропусками. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Харків, 2014.

Дисертацію присвячено розробці динамічних методів інтелектуального аналізу викривлених даних в таблицях «об’єкт – властивість» та часових рядах для відновлення даних в on-line режимі, коли дані надходять на обробку послідовно.

Розроблено адаптивну нейро-фаззі систему, що дозволяє розв’язувати задачу відновлення пропусків в on-line режимі з постійною корекцією відновлених елементів та центроїдів кластерів. Розроблено нейро-фаззі методи для відновлення та кластеризації спотворених даних на основі самоорганізовної нейро-фаззі мапи Кохонена, що дозволяє обробляти дані в on-line режимі та забезпечує роботу з класами, що перетинаються. Дістали подальший розвиток методи кластеризації даних з пропусками, що засновані на рекурентній оптимізації спеціального виду цільових функцій, в яких спостереження замінюються оцінками, що отримані в процесі розв’язання задачі; методи адаптивної нечіткої кластеризації даних з пропусками, що дозволяють опрацьовувати інформацію на основі стратегії

найближчого прототипу-центроїда та забезпечують роботу в on-line режимі. Розв’язано задачі відновлення викривлених даних, наданих заводом рентгенівської техніки та сервісним центром, за допомогою запропонованих методів, що дало можливість пришвидшити роботу відновлення обладнання, яке вийшло з ладу, а також завчасно ідентифікувати можливу несправність.

Ключові слова: гібридні системи обчислювального інтелекту, штучні нейронні мережі, нейро-фаззі системи, нечітка кластеризація, самонавчання, викривлені спостереження.

АННОТАЦИЯ

Шафроненко А.Ю. Методы динамического интеллектуального анализа данных с пропусками. – Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Харьковский национальный університет радиоэлектроники, Министерство образования и науки Украины, Харьков, 2014.

Диссертация посвящена разработке методов динамического интеллектуального анализа данных с пропусками, в таблицах «объект - свойство» и временных рядах для восстановления данных в on-line режиме, когда данные поступают на обработку последовательно.

Проведен анализ методов обработки нестационарных сигналов в условиях дефицита текущей информации. Рассмотрены преимущества и недостатки известных нейро-фаззи систем, самоорганизующися карт Кохонена, а также моделей восстановления данных с пропусками. Рассмотрены проблемы кластеризации искаженных и восстановленных данных, проведен анализ известных архитектур нейро-фаззи систем, нейронных сетей, получивших наибольшее распространение. На основе проведенного анализа определены задачи исследования, заключающиеся в разработке адаптивных нейро-фаззи методов кластеризации и восстановления искаженных данных, адаптивных нейронных сетей для решения задач динамического интеллектуального анализа искаженных данных, а также методов их обучения, учитывающие особенности задач обработки данных с пропусками.

Разработана адаптивная нейро-фаззи система, которая позволяет решать задачу восстановления пропусков в таблицах «объект – свойство», содержащих априори неизвестное количество пропусков, в on-line режиме с постоянной коррекцией элементов восстанавливаемой таблицы, а также обеспечивает высокое быстродействие и простоту численной реализации. Разработана адаптивная нейронная сеть, которая позволяет настраивать в процессе обучения не только синаптические весы, но и структуру, обеспечивая высокое быстродействие и предназначена для обработки искаженных нестационарных нелинейных стохастических и хаотических сигналов, которые поступают на обработку в реальном времени. Разработаны нейро-фаззи методы самообучения карты Кохонена, которые позволяют обрабатывать данные в on-line режиме, обеспечивая работу с пересекающимися классами. Получили дальнейшее развитие методы кластеризации данных с пропусками, основанные на рекуррентной оптимизации специального вида

целевых функций, которые отличаются тем, что наблюдения заменяются оценками, полученными в процессе решения оптимизационной задачи; методы вероятностной и возможностной адаптивной нечеткой кластеризации данных с пропусками, которые позволяют обрабатывать информацию на основе стратегии ближайшего прототипа-центроида, а также обеспечивают работу в on-line режиме, а сам процесс обработки информации может быть организован на основе самоорганизующейся карты Кохонена. Проведено имитационное моделирование методов восстановления и кластеризации искаженных данных, проведена экспериментальная оценка ошибок восстановленных данных и ее изменение при работе в on-line режиме. Решена прикладная задача восстановления искаженных данных, предоставленных Заводом рентгеновского оборудования и сервисным центром. Повышена скорость тестирования медицинского рентген оборудования с помощью предложенных методов. Решена практическая задача восстановления искаженных данных, которыми оперирует сервисный центр рентгеновского оборудования, что ускорило ремонтные работы с оборудованием, которые вышло из строя, а также появилась возможность заблаговременно идентифицировать возможную поломку и предоставить необходимую информацию мастеру.

Ключевые слова: гибридные системы вычислительного интеллекта, искусственные нейронные сети, нейро-фаззи системы, нечеткая кластеризация, самообучение, искаженные наблюдения.

ABSTRACT

Shafronenko A. Yu. Methods of dynamic intellectual analysis for data with missing values. – Manuscript.

Dissertation for a candidate of technical science (Ph.D.) degree in speciality 05.13.23 – systems and means of artificial intelligence. – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2014.

The dissertation is devoted to the development of methods of dynamic intellectual analysis for distorting data in the tables “object - property” and the time series for data recovery in on-line mode.

An adaptive neuro-fuzzy system that allows to solve the problem of restoring missing values in on-line mode with correction of recovered elements and centroids of clusters was proposed. A neuro-fuzzy method for recovery of distorted data and clustering based on neuro-fuzzy Kohonen maps, that allows to process the data in on-line mode and provide operation with overlapping class was proposed. Have got further development of methods for data clustering data with missing values, based on recurrent optimization of objective functions in special type whose observations are replaced by estimates obtained in the process of solving the problem; methods of adaptive fuzzy clustering of data with missing values that allow to process information using on strategy of nearest prototype-centroid in on-line mode. The problem of restoration of distorted data provided by the x-ray plant and service center using the proposed methods, making it possible speed up recovery hardware that is out of order, and early identification of potential problem.

Keywords: Hybrid systems of computational intelligence, artificial neural networks, neuro-fuzzy systems, fuzzy clustering, self-learning, distorted observation.