

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ПОТОКІВ МЕДИЧНИХ ДАНИХ ЗА УМОВ НЕВИЗНАЧЕНОСТІ

І. Г. ПЕРОВА, Є. М. БРАЖНІКОВА

У роботі запропоновано інформаційну технологію аналізу потоків медичних даних за умов невизначеності. Невизначеність у даному випадку виражається у тому, що заздалегідь невідомим є загальна кількість пацієнтів, в ході розв'язання деяких задач початкова кількість ознак і діагнозів може змінюватися в процесі діагностування, а також дані можуть бути нерозміченими або розміченими так, що кількість пацієнтів із відомим діагнозом буде достатньою для навчання (репрезентативна вибірка) або недостатньою (нерепрезентативна вибірка). У результаті роботи проводиться нечітке діагностування кожного з пацієнтів у послідовному режимі.

Ключові слова: інформаційна технологія, нео-фаззі нейрон, діагностуюча нейро-фаззі система, контрольоване навчання, самонавчання, активне навчання.

ВСТУП

Основним завданням медичної діагностики є визначення можливих діагнозів хворого на основі знань предметної області та даних його обстеження. До них належать значення ознак (у моменти їхнього спостереження), значення анатомо-фізіологічних особливостей (постійні у часі) і значення подій, що відбулися (в моменти, коли вони відбувалися). На початковому етапі розвитку всі завдання вирішував один чоловік. Його зір і слух були основними вимірювальними інструментами, а досвід визначав якість діагнозу. У міру розвитку засобів вимірювань функції вимірювача параметрів організму людини передавалися приладам. Постановку діагнозу з урахуванням результатів виконаних вимірів продовжував здійснювати фахівець з фундаментальною діагностичною підготовкою.

Виходячи з розглянутої задачі, можна виділити наступні труднощі, з якими стикаються лікарі на етапі проведення медичної діагностики:

- неоднозначність діагнозу, тобто наявність супутніх захворювань, які сильно викривляють картину симптомів;
- часто, невиправдано велика кількість показників, що також призводить до викривлення даних;
- масове обстеження, тобто відсутність індивідуального підходу до кожного пацієнта, а, як відомо, значення чутливості і специфічності для кожної людини необхідно підбирати персонально;
- лікування захворювання вже на етапі його прояви і розвитку, а не на етапі профілактики його появи;
- кількість існуючих захворювань є більш ніж на порядок більшою, ніж кількість методів їх діагностики, що відповідно підвищує ризик неправильної постановки діагнозу;
- відсутність статистики по рідкісних формах захворювання, оскільки ці захворювання зустрічаються нечасто;

– суб'єктивізм лікаря, оскільки саме лікар ставить остаточний діагноз і як показує практика, більше половини всіх лікарських помилок відбувається за рахунок цього фактору (Акопов, 2001).

Medical Data Mining – це галузь обчислювального інтелекту, що займається інформаційною обробкою медичних даних [1–3]. Серед безлічі можливих завдань, що виникають в рамках Medical Data Mining особливе місце займає проблема діагностування стану організму за множиною показників, вимірюваних у тих чи інших шкалах.

З формальної точки зору завдання медико-біологічної діагностики ускладнюються цілим рядом обставин, таких як:

- неопуклість і перекриття класів, що відповідають різним діагнозам;
- «збуреність» (включаючи аномальні викиди) та нестаціонарність масивів вихідних даних;
- можливість наявності як надмалих навчальних вибірок (пропорційних розмірності вектору ознак), так і дуже великих обсягів інформації, що вписуються в концепцію Big Data;
- можливість і необхідність подання вихідної інформації у формі потоків даних (Data Stream), що послідовно надходять на обробку в online режимі.

Зазначені обставини істотно обмежують клас відомих систем, придатних для роботи в описаних умовах.

Таким чином, актуальною є задача розробки та дослідження інформаційної технології для медичного діагностування в послідовному режимі за умов суттєвої апіорної та поточної невизначеності на основі адаптивних гібридних нейро-фаззі систем і методів їхнього навчання.

1. ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ПОТОКІВ МЕДИЧНИХ ДАНИХ

У роботах [4–14] розглянуто системи, що проводять обробку медичних даних у послідовному режимі за різних початкових умов. Всі ці системи можуть бути

об'єднані в інформаційну технологію аналізу потоків медичних даних за умов невизначеності. Саме така структура вирішує науково-прикладну проблему розробки та дослідження інформаційної технології медичного діагностування в послідовному режимі за умов суттєвої апіорної та поточної невизначеності на основі адаптивних гібридних нейро-фаззі систем та їхнього навчання.

Невизначеність у даному випадку виражається у тому, що заздалегідь невідомим є загальна кількість пацієнтів, у ході розв'язання деяких задач початкова

кількість ознак і діагнозів може змінюватися в процесі діагностування, а також дані можуть бути нерозміченими або розміченими так, що кількість пацієнтів із відомим діагнозом буде достатньою для навчання (репрезентативна вибірка) або недостатньою (нерепрезентативна вибірка).

Першим етапом роботи будь-якої системи обробки даних є етап препроцесінгу даних, основні підходи до якого широко розглянуті у роботах [15–17] і зображені на рис. 1.

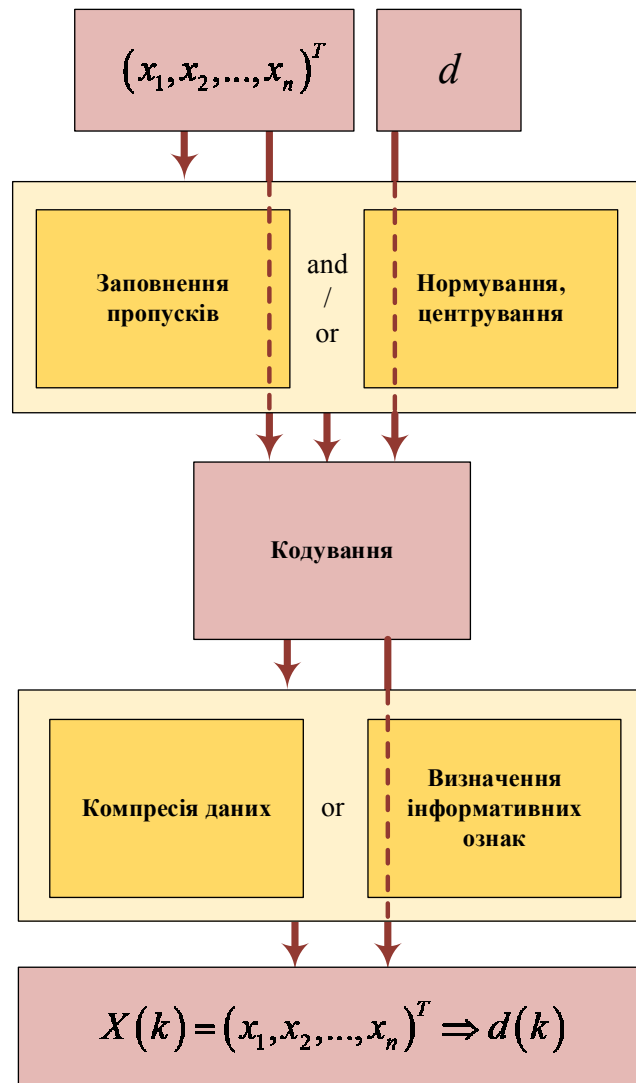


Рис. 1. Етапи препроцесінгу даних

На першому етапі ознаки одного пацієнта $(x_1, x_2, \dots, x_n)^T$ та діагноз d надходять на вхід системи препроцесінгу даних. Ця система складається із обов'язкових та необов'язкових етапів. Перші два етапи препроцесінгу є необов'язковими і залежать від початкового стану даних та вибору подальшої системи діагностування.

Якщо вхідні ознаки мають пропуски, тоді вони мають бути заповнені за допомогою методу адаптивного заповнення пропусків, що розглянуто у [15]. Цей під-

хід дозволяє проводити заповнення в послідовному онлайн-режимі орієнтуючись виключно на міру близькості між пацієнтами. Якщо пропусків немає, цей етап може бути пропущений.

Наступним необов'язковим етапом є нормування і центрування даних, що проводиться в послідовному режимі за допомогою онлайн підходу на основі нейронних мереж.

Обчислення середнього можна здійснювати за допомогою рекурентного співвідношення схемна реалізація якого наведена на рис. 2

$$\bar{x}_i(k) = \bar{x}_i(k-1) + \frac{1}{k}(x_i(k) - \bar{x}_i(k-1)).$$

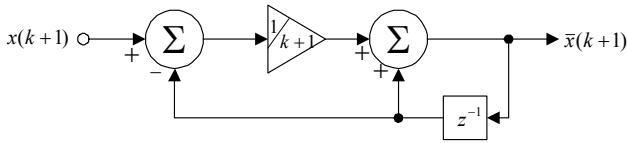


Рис. 2. Нейрон для обчислення середнього
Для розрахунку вибіркової дисперсії

$$\sigma_x^2(k+1) = \frac{1}{k+1} \sum_{p=1}^{k+1} (x(p) - \bar{x}(k+1))^2$$

можна скористатися рекурентним співвідношенням

$$\sigma_x^2(k+1) = \sigma_x^2(k) + \frac{1}{k+1} ((x(k+1) - \bar{x}(k+1))^2 - \sigma_x^2(k)),$$

що реалізується схемою, наведеною на рис. 3.

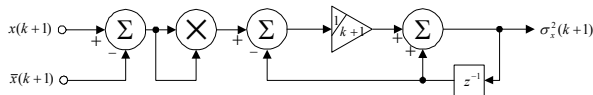


Рис. 3. Нейрон для обчислення дисперсії

Порівняння двох чисел x_1 і x_2 для кодування даних у гіперкуб $[0, 1]^n$ або $[-1, 1]^n$ можна проводити, використовуючи співвідношення:

$$\begin{cases} \max\{x_1, x_2\} = x_1 - 0.5(1 - \text{sign}(x_1 - x_2))(x_1 - x_2), \\ \min\{x_1, x_2\} = x_1 - 0.5(1 + \text{sign}(x_1 - x_2))(x_1 - x_2) \end{cases}$$

і нейроноподібну структуру, показану рис. 4.

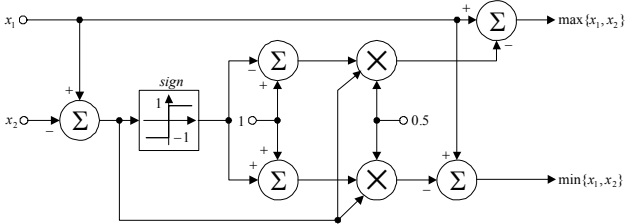


Рис. 4. Нейрон для порівняння двох чисел

Третім етапом є обов'язковий етап кодування даних, що враховує тип шкал, в яких вимірюються медичні дані. Для цього пропонується використовувати адаптивне кодування медичних даних окремо для кількісних, рангових і категоріальних ознак [16].

На наступному етапі (за необхідністю) проводиться оцінка інформативності медичних ознак, що може бути проведена шляхом компресії вхідного простору ознак в онлайн-режимі або відбору інформативних показників за допомогою модифікованої нейронної мережі, побудованої на нейронах Ойя у задачах оцінювання інформативності у послідовному online режимі [17].

Таким чином, останнім етапом є отримання передоброблених даних

$$\tilde{X}(k) = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)^T \Rightarrow d(k),$$

що можуть бути використані як вхідні для інформаційної технології аналізу потоків медичних даних за умов невизначеності.

Структурна схема інформаційної гібридної технології аналізу потоків медичних даних за умов невизначеності наведено на рис. 5. На вхід поступають ознаки пацієнтів у послідовному режимі та їхні діагнози (за наявності). робота якої залежить від того чи відомий діагноз і якщо відомий, то репрезентативна вибірка даних чи ні (тобто від значення, що міститься у $d(k)$).

Тобто маємо три ситуації:

1. Якщо вибірка є репрезентативною, то інформаційна технологія аналізу потоків медичних даних переходить у режим контрольованого навчання, який може бути реалізований шляхом використання трьох адаптивних нейро-фаззі систем. Це може бути адаптивне швидке діагностування за допомогою багатовимірного нео-фаззі нейрону [9, 13], що є однією із найшвидших систем діагностування за рахунок використання оптимального алгоритму навчання. Другим варіантом є адаптивна діагностична нейро-фаззі система реального часу DNFS [4], що має властивості змінювати свою архітектуру за потреби, тобто еволюціонувати. Еволюцією вважається зміна архітектури системи, а саме кількості її входів та виходів (ознак та діагнозів). У ситуаціях, коли в процесі діагностування у пацієнта змінюється кількість симптомів (ознак) або з'являється новий діагноз архітектура DNFS змінюється за допомогою адаптації-еволюції, що в кінцевому результаті перетворює її в систему $DNFS_{n+1,m+1}$ [11, 14]. Запропонована глибинна гібридна система обчислювального інтелекту з архітектурною адаптацією для нечіткого медичного діагностування $DNFS_{n+1,m+1}$ характеризується високою швидкістю за рахунок оптимальних алгоритмів навчання та можливістю продовжувати процес діагностування без перенавчання всієї архітектури системи.

Третім варіантом є адаптивна вейвлет діагностична нейро-фаззі система реального часу [18], що має високу точність за рахунок використання модифікації експоненційно зваженого рекурентного методу найменших квадратів.

2. Якщо вибірка не є репрезентативною, то інформаційна технологія аналізу потоків медичних даних переходить у режим активного навчання та асоціації, тобто переключається між режимами навчання-самонавчання за потреби.

У цьому режимі діагноз лише деяких пацієнтів є відомим, але при цьому кількість цих відомих діагнозів недостатня для навчання систем, що базуються на контрольованому навчанні (нерепрезентативна навчальна вибірка). Першим варіантом є використання методу нечіткої класифікації даних медико-біологічних дослі-

джен, що працює в умовах дефіциту інформації і здатен проводити класифікацію на екстра-малій кількості відомих діагнозів [7]. Другим варіантом є швидке медичне діагностування за допомогою нейро-фаззі автоасоціативної пам'яті [10], що здатна проводити асоціацію за ознаками кількох пацієнтів, щоб виявити ступінь близькості пацієнта із невідомим діагнозом до

нього. Третім варіантом є система онлайн нейро-фаззі медичного діагностування із активним навчанням [12], що автоматично переключається між режимами контрольованого навчання та самонавчання залежно від того розмічені дані потрапляють на її вхід чи нерозмічені.

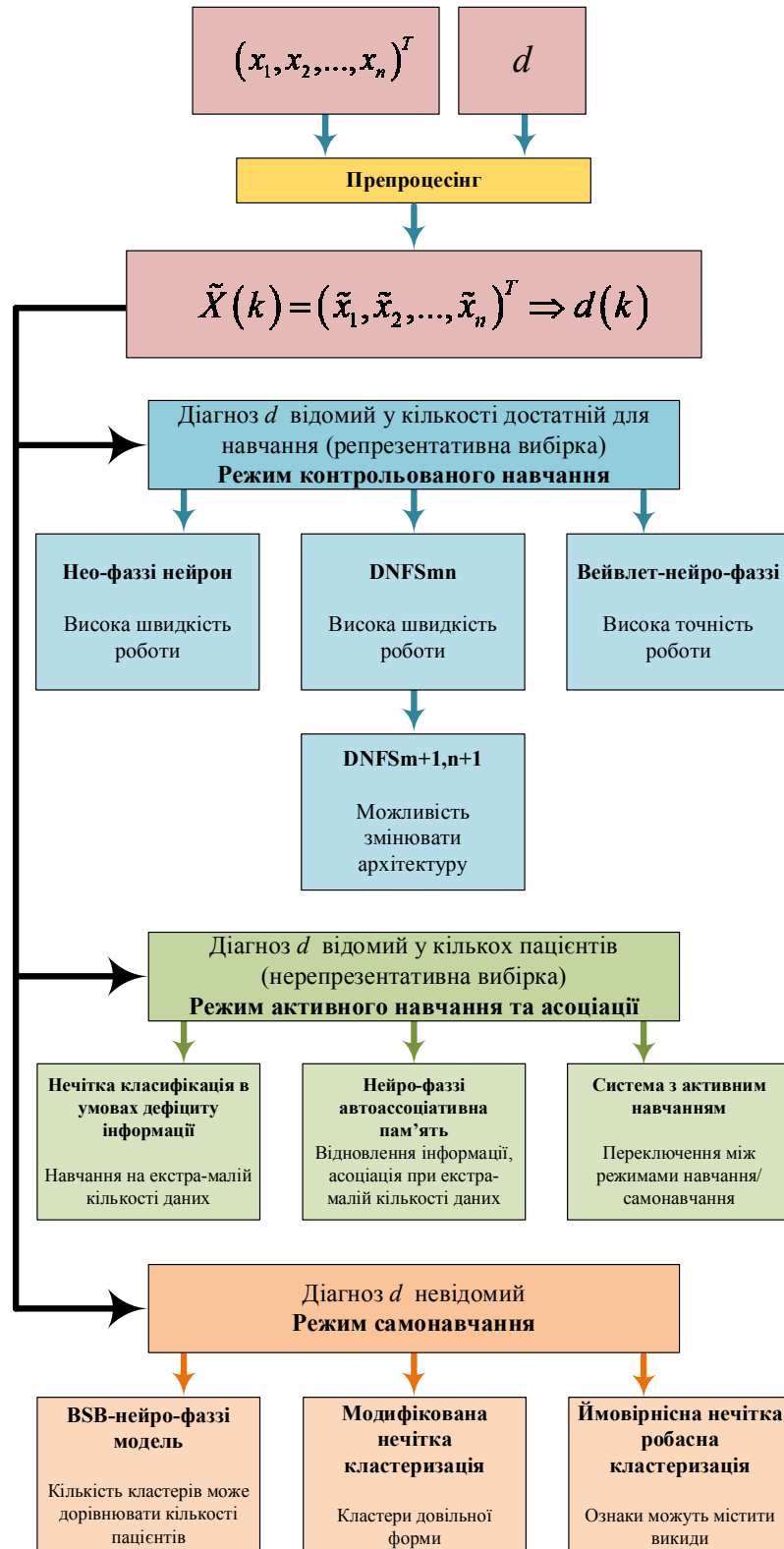


Рис. 5. Структурна схема інформаційної технології аналізу потоків медичних даних за умов невизначеності

3. Якщо вибірка є не нерозміченою, тобто діагнози жодного з пацієнтів є невідомими, то інформаційна технологія аналізу потоків медичних даних переходить у режим самонавчання.

Першою системою в рамках такого режиму є модифікований метод нечіткої кластеризації медичних даних [19], що дозволяє проводити послідовне діагностування в режимі самонавчання із використанням кластерів-діагнозів довільної форми.

Друга система в рамках режиму самонавчання використовується в ситуаціях, коли медичні ознаки можуть містити аномальні викиди, завади та інші артефакти. Для цього використовується метод адаптивної робастної нечіткої ймовірнісної кластеризації на основі манхеттенської метрики [8].

В ситуаціях, коли кількість кластерів-діагнозів заздалегідь невідома і існує вірогідність того, що їхня кількість буде порівнянна із кількістю даних, слід використовувати нечітку кластеризацію на основі BSB-нейро-фаззи моделі [5–6], в якій на першому етапі усі дані розподіляються між вершинами гіперкуба, а потім за допомогою функції належності спеціальної форми можливо об'єднати кілька вершин із розрахуванням функції належності до вершини-діагнозу.

ВИСНОВКИ

Розроблена інформаційна технологія аналізу потоків медичних даних за умов невизначеності дозволяє проводити нечітке діагностування, оцінку інформативності медичних ознак та заповнення пропущених значень в послідовному онлайн режимі із високою швидкістю.

Література

- [1] de Oliveira, J., Pedrycz, W. *Advances in Fuzzy Clustering and its Applications*. John Wiley&Sons Ltd. – 2007. 454p.
- [2] Berka, P., Rauch, J., & Zighed, D. *Data mining and medical knowledge management cases and applications*. New-York: Hershey. – 2009. 440 p.
- [3] Giannopoulou, E. *Data mining in medical and biological research*. New York: ITAe. – 2008. 331 p.
- [4] Бодянский Е. В., Винокурова Е. А., Мулеса П. П., Перова И. Г. Диагностирующая нейро-фаззи-система и ее адаптивное обучения в задачах интеллектуальной обработки данных медико-биологических исследований / Системні технології. Регіональний міжвузівський збірник наукових праць. – Випуск 2(91). – Дніпропетровськ, 2014. – С. 125–135.
- [5] Бодянский Е. В., Перова И. Г. Нечеткая кластеризация биомедицинских данных на основе BSB-нейро-фаззи-модели // Наукові праці. – Миколаїв, Вид-во ЧДУ ім. Петра Могили. – 2014. – Вип. 238. – Т.250. Комп'ютерні технології. – С. 93–100.
- [6] Perova, I., Bodyanskiy, Y., Sachenko, A., Karpinski, M., Rudyk, P. Fuzzy clustering of biomedical datasets using BSB-neuro-fuzzy-model // *CEUR Workshop Proceedings*, Lviv, 2018. pp. 21–28.
- [7] Бодянский Е. В., Перова И. Г. Нечеткая классификация данных медико-биологических исследований в условиях дефицита информации. Системы обработки информации. – Х.: ХУПС, 2015. – Вип. 11(136). – С. 161–163.

- [8] I. Perova, Ye. Bodyanskiy Adaptive fuzzy clustering based on Manhattan metrics in medical and biological applications // *Вісник національного університету “Львівська політехніка”* №826, 2015. p. 8–12.
- [9] И. Г. Перова, Е. Н. Бражникова, И. П. Плюсс Нео-фаззи подход в задачах online медико-биологической диагностики // *Прикладная радиоэлектроника*. Харьков: ХНУРЭ, 2016. – т. 15, №1. – С. 51–57.
- [10] I. Perova, Ye. Bodyanskiy Fast medical diagnostics using autoassociative neuro-fuzzy memory // *International Journal of Computing*, 16 (1), 2017. P. 34–40.
- [11] I. Perova, I. Pliss Deep hybrid System of Computational Intelligence with Architecture Adaptation for Medical Fuzzy Diagnostics // *I.J. Intelligent System and Applications*, 2017, 7, 12–21 DOI: 10.5815/ijisa.2017.07.02
- [12] I. Perova, Ye. Bodyanskiy “Medical online neuro-fuzzy diagnostics system with active learning” *International Journal of Advances in Computer and Electronics Engineering* Volume 2, Issue 7, July 2017. pp. 1–10
- [13] S. M. K. Mahmoud, I. Perova, I. Pliss, Multidimensional neo-fuzzy-neuron for solving medical diagnostics tasks in online-mode // *Journal of Applied Computer Science*, 2017, Vol. 25, № 1. pp. 39–48
- [14] I. Pliss, I. Perova “Diagnostic Neuro-Fuzzy System and Its Learning in Medical Data Mining Tasks in Conditions of Uncertainty about Numbers of Attributes and Diagnoses” *Automatic Control and Computer Sciences*, 2017, 51(6), pp. 391–398. DOI: 10.3103/ S0146411617060062
- [15] Mulesa P., Perova I. Fuzzy Spacial Extrapolation Method Using Manhattan Metrics for Tasks of Medical Data Mining *Computer Science and Information Technologies CSIT'2015*. – Lviv, Ukraine. – 2015, p. 104–106. DOI: 10.1109/STC-CSIT.2015.7325443
- [16] Перова И. Г., Бодянский Е. В. Об одной метрике в задачах интеллектуального анализа медицинских данных // 19-th International conference on System Analysis and Information Technology SAIT 2017, May 22–25, 2017 Institute for Applied System Analysis of National Technical University of Ukraine “KPI”, Kyiv, Ukraine. – С. 173.
- [17] I. Perova, Ye. Bodyanskiy Adaptive Human Machine Interaction Approach for Feature Selection-Extraction Task in Medical Data Mining / *International Journal of Computing*, 17(2) 2018, 113–119 p.
- [18] Ye. Bodyanskiy, I. Perova, O. Vynokurova, I. Izonin “Adaptive Wavelet Diagnostic Neuro-Fuzzy System for Biomedical Tasks” *Proc. of 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, Lviv-Slavske, Ukraine, February 20 – 24, 2018, pp.299-303
- [19] Перова И. Г. Адаптивная обработка данных медико-биологических исследований методами вычислительного интеллекта // *Восточно-европейский журнал передовых технологий*. – Х.: 2014. – №1(67), с. 24–28.

Надійшла до редколегії 26.12.2018



Перова Ірина Геннадіївна, к.т.н., с.н.с., доцент, доцент кафедри біомедичної інженерії Харківського національного університету радіоелектроніки. Галузь наукових інтересів – гібридні нейро-фаззи системи обчислювального інтелекту під час інтелектуальної обробки потоків медико-біологічних даних.



Бражнікова Єлізавета Миколаївна, аспірант кафедри біомедичної інженерії Харківського національного університету радіоелектроніки. Галузь наукових інтересів – інтелектуальна обробка медико-біологічної інформації.

УДК 519.7:004.8

Перова І. Г. **Информационная технология анализа потоков медицинских данных в условиях неопределенности** / И. Г. Перова, Е. Н. Бражникова // Прикладная радиоэлектроника: науч.-техн. журнал. – 2019. – Том 18, № 1, 2. – С. 46–51.

В статті розглянута інтелектуальна технологія аналізу потоків медичних даних в умовах неопределенності, яка дозволяє проводити нечітке діагностування, оцінку інформативності і заповнення пропущен-

них значень медичних ознак в послідовному режимі онлайн з високим швидкістю.

Ключевые слова: інформаційна технологія, нео-фаззи нейрон, діагностуюча нейро-фаззи система, контролюване навчання, самонавчання, активне навчання.

Ил.: 05. Библиогр.: 19 назв.

UDC 519.7:004.8

Perova I. G. **Information technology of medical data stream mining in conditions of uncertainty** / I. G. Perova, E. M. Brazhnikova // Applied Radio Electronics: Sci. Journ. – 2019. – Vol. 18. № 1, 2. – P. 46–51.

The article considers the information technology of medical data stream analysis under the conditions of uncertainty which allows carrying out fuzzy diagnostics, evaluation of information capability and filling in the missed values of medical features in an online mode with high operation speed.

Keywords: information technology, neo-fuzzy neuron, diagnostic neuro-fuzzy system, control learning, self-learning, active learning.

Fig. 05. Ref.: 19 items.