УДК 004.032.26

Є. Бодянський, О. Винокурова, І. Кобилін, П. Мулеса
Харківський національний університет радіоелектроніки,
Проблемна науково-дослідна лабораторія автоматизованих систем управління

АДАПТИВНА МАТРИЧНА НЕЙРО-ФАЗЗІ САМООРГАНІЗОВНА МЕРЕЖА ДЛЯ КЛАСТЕРИЗАЦІЇ БАГАТОВИМІРНИХ ПОТОКІВ ДАНИХ

Запропоновано адаптивну матричну нейро-фаззі самооргазізовну мережу для кластеризації багатовимірних потоків даних (біомедичні масиви спостережень, сигнали цифрового відео, що формують дискретні двовимірні поля тощо).

Ця мережа характеризується простотою обчислювальної реалізації, високими апроксимувальними властивостями, швидкодією процесу навчання і призначена для розв'язання широкого класу задач інтелектуального аналізу потоків даних. Результати низки експериментів як на тестових, так і на реальних даних підтверджують ефективність запропонованого підходу.

Ключові слова: кластеризація багатовимірних потоків даних, багатовимірні часові ряди, матрична нейро-фаззі самоорганізовна мережа, адаптивні процедури навчання.

Time series clustering is wide spread problem in Data Stream Mining tasks and nowadays there are a lot of various approaches for solving such tasks that are based on different a priori assumptions. However, there are cases when well-known methods and algorithms for solving this task are inoperative in real applications. One of such tasks is short time series fuzzy clustering with unevenly distributed in time observations. The time series clustering of data set with missed observations is sufficiently close to this problem. The object of clustering is the sample in total and the observations are recorded by unevenly instants of time. Generated clusters are overlapped in such way that each processed sample can belong to several classes. At that it is assumed also, that all processed data are defined in the form of a fixed data set with unchanged size.

In the connections with that, it seems appropriate the spreading of the fuzzy clustering of short time series with unevenly distributed observations approach to the situation when the data are fed to the processing in online mode in the form of multivariate data stream in the context of Data Stream Mining.

In the paper the fuzzy clustering approach of multivariate short time series with unevenly distributed observations is considered. Such time series are fed to the processing in batch mode or sequentially on-line mode. In the first case we can use the matrix modification of fuzzy C-means method, and in second case we can use the matrix modification of neuro-fuzzy network by T. Kohonen, which is learned using the rule "Winner takes more". Proposed fuzzy clustering algorithms are enough simple in computational implementation and can be used for solving of wide class of Data Stream Mining problems.

Key words: multivariate data stream clustering, multivariate time series, matrix neuro-fuzzy self-organizing network, adaptive learning algorithms.

Вступ

Задачу кластерування та сегментування часових рядів достатньо досліджено в межах інтелектуального аналізу даних (Data Mining) [1–4], і сьогодні існує багато різноманітних алгоритмів її розв'язання, що основані на тих чи інших апріорних припущеннях.

У багатьох практичних застосуваннях виникають ситуації, коли відомі і популярні підходи щодо розв'язання задачі виявляються неефективними та непрацездатними.

Однією з таких задач є нечітка кластеризація коротких часових рядів з нерівномірним розподілом спостережень у часі [5]. Доволі близька до цієї задачі є кластеризація неповних рядів, в яких частина спостережень або втрачена, або взагалі відсутня [6].

Особливістю цих задач є те, що об'єктом кластеризації є не окремі спостереження, а вибірки загалом, самі спостереження фіксуються через нерівновіддалені моменти часу, а кластери, що формуються, перетинаються так, що кожна вибірка, що опрацьовується, може належати відразу декільком класам [7, 8]. Передбачено також, що всю вихідну інформацію задано у формі фіксованого масиву даних, обсяг якого не змінюється.

Ситуація ще більше ускладнюється, якщо вихідну інформацію задано у формі багатовимірних часових рядів, тобто двовимірних полів спостережень. Прикладом таких двовимірних полів можуть бути електромагнітні, термічні і оптичні поля, області забруднення повітря та води, біомедичні масиви спостережень та насамперед сигнали цифрового відео, що формують дискретні двовимірні поля.

Тому доцільно поширити підхід нечіткої кластеризації коротких часових рядів з нерівновіддаленими спостереженнями [5] на ситуацію, коли дані надходять на обробку в on-line режимі у формі багатовимірного потоку інформації в межах концепції Data Stream Mining [9].

Нечітка ймовірнісна кластеризація багатовимірних коротких рядів

Нехай вихідну інформацію задано у формі набору $(q\times n)$ -вимірних матриць $X(k)=\{x_{ip}(k)\}$ (тут $i=1,2,...n$, n – номер окремого спостереження q -вимірної послідовності в k -й реалізації (вибірці), $k=1,2...N$, $p=1,2,...,q$ – p-а координата багатовимірного процесу), що містить N ($N>n$) q -вимірних реалізацій з нерівномірним тактом квантування, при цьому p-а компонента $X(k)$ може бути представлена у вигляді $(1\times n)$-вектора $x_p(\mathrm{k})=(x_{1p}(k),x_{2p}(k),\mathbf{K},x_{np}(k))$. Нерівномірність квантування означає, що $\Delta t_i=t_i-t_{i-1}\neq\Delta t_{i+1}=t_{i+1}-t_i$, тобто $\Delta t_i\neq const$. Приклад однієї такої реалізації наведено на рис.1.

Очевидно, що для оцінювання відстані між двома реалізаціями $X(k)$ та $X(l)$ не можна використати ані традиційну евклідову метрику, ані класичні ймовірнісні критерії. Для оцінювання відстані між одновимірними часовими рядами в [5] було введено так звану PS-відстань, що основана на представленні цих рядів у формі кусково-лінійних функцій та оцінює по суті відмінність форм (нахилів) аналізованих вибірок. При цьому відстань між двома послідовностями, наприклад, $x_p(k)$ та $x_p(l)$, можна записати у вигляді

$$d^2_{PS}(x_p(k),x_p(l))=\sum_{i=1}^{n-1}\left(\frac{x_{i+1,p}(k)-x_{ip}(k)}{t_{i+1}-t_i}-\frac{x_{i+1,p}(l)-x_{ip}(l)}{t_{i+1}-t_i}\right)^2=$$
$$=\sum_{i=1}^{n-1}\left(\frac{x_{i+1,p}(k)-x_{ip}(k)}{\Delta t_{i+1}}-\frac{x_{i+1,p}(l)-x_{ip}(l)}{\Delta t_{i+1}}\right)^2, \quad (1)$$

що задовольняє усі умови, які визначають метрику.

Використовуючи відстань (1), автори [5] ввели процедуру нечіткої кластеризації, що є модифікацією популярного алгоритму нечітких с-середніх (FCM) [7] на випадок одновимірних часових рядів з нерівновіддаленими спостереженнями.

Нескладно помітити, що компоненти відстані (1) є за суттю першими різницями дискретних сигналів $x_p(k)$ і $x_p(l)$, тобто тангенсами кутів нахилу кусково-лінійних функцій:

$$\Delta x_{i+1,p}(k)=\frac{x_{i+1,p}(k)-x_{ip}(k)}{\Delta t_{i+1}}=tg\alpha_{i+1,p}(k),$$
$$\Delta x_{i+1,p}(l)=\frac{x_{i+1,p}(l)-x_{ip}(l)}{\Delta t_{i+1}}=tg\alpha_{i+1,p}(l).$$

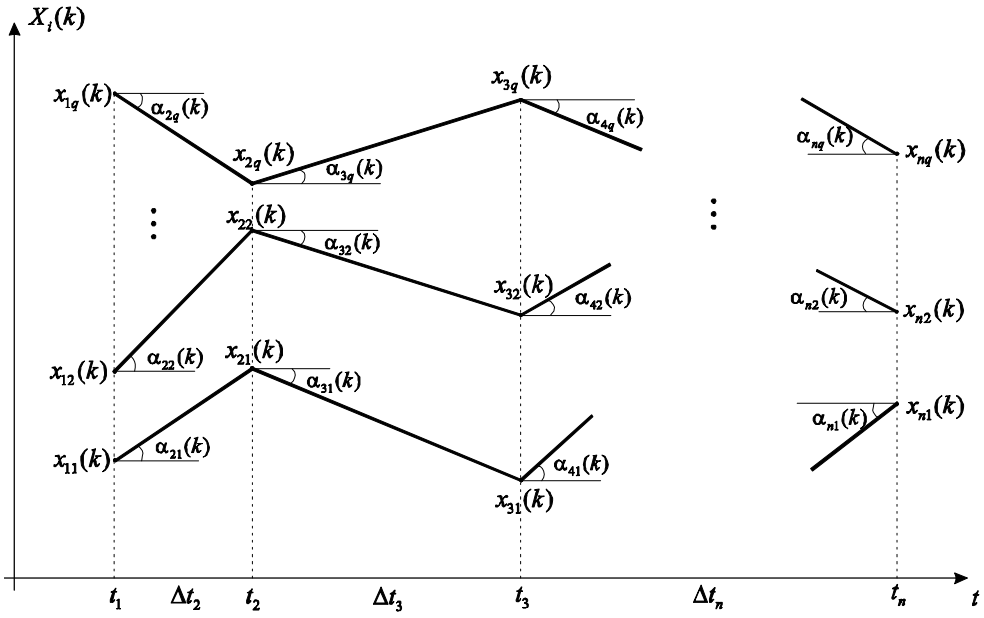

Рис. 1. Багатовимірний часовий ряд з нерівномірним тактом квантування

При цьому ряд, що сформований першими різницями, містить на одну точку менше, ніж вихідна вибірка, тобто не n, а $(n-1)$ спостережень $\Delta x_{2p}(k) = tg\alpha_{2p}(k)$, $\Delta x_{3p}(k) = tg\alpha_{3p}(k)$, …, $\Delta x_{np}(k) = tg\alpha_{np}(k)$.

Оскільки в результаті взяття першої різниці, що є аналогом першої похідної у неперервному випадку, у ряді $x_p(k)$ видаляється його середнє значення, для відновлення вихідної послідовності за її першими різницями необхідно доповнити послідовність цих різниць довільним з вихідних спостережень, наприклад, $x_{np}(k)$.

Тоді відновлюється вихідний ряд елементарно за виразом

$$\begin{cases} x_{n-1,p}(k) = x_{np}(k) - \Delta x_{np}(k)\Delta t_n, \\ x_{n-2,p}(k) = x_{n-1,p}(k) - \Delta x_{n-1,p}(k)\Delta t_{n-1}, \\ \text{M} \\ x_{1p}(k) = x_{2p}(k) - \Delta x_{2p}(k)\Delta t_2. \end{cases} \tag{2}$$

Вводячи далі $(1\times n)$ – вектор $\%_p(k) = (\Delta x_{2p}(k), \Delta x_{3p}(k), ..., \Delta x_{np}(k), x_{np}(k))$, можна переписати (1) у традиційній формі

$$d^2_{PS}(x_p(k), x_p(l)) = \left\| \%(k) - \%(l) \right\|^2, \tag{3}$$

тобто повернутися до стандартної евклідової відстані між першими різницями вихідних вибірок.

Потім на основі метрики (3) нескладно реалізувати будь-який з методів нечіткого кластерного аналізу [8].

Для того, щоб скористатися ідеєю оцінювання відстаней між рядами за їх першими різницями, розглянемо $(q\times n)$ – матрицю

$$\%(k) = \begin{pmatrix} \Delta x_{21}(k) & \Delta x_{31}(k) & \mathbf{L} & \Delta x_{nl}(k) & x_{nl}(k) \\ \text{M} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \text{M} \\ \text{M} & \mathbf{L} & \Delta x_{ip}(k) & \mathbf{L} & \text{M} \\ \text{M} & \mathbf{L} & \mathbf{L} & \mathbf{L} & \text{M} \\ \Delta x_{2q}(k) & \Delta x_{3q}(k) & \mathbf{L} & \Delta x_{nq}(k) & x_{nq}(k) \end{pmatrix}$$

і замість евклідової відстані – сферичну норму:

$$D^2_{PS}(X(k),X(l)) = Tr(\overset{\%}{X}(k)-\overset{\%}{X}(l))(\overset{\%}{X}(k)-\overset{\%}{X}(l))^T, \quad (4)$$

що є узагальненням (3) на матричний випадок.

На основі відстані (4) можна здійснити нечітку кластеризацію масиву реалізації $\overset{\%}{X}(1),\overset{\%}{X}(2),\ldots,\overset{\%}{X}(N)$.

Використовуючи методику нечіткого ймовірнісного кластерного аналізу, розглянемо цільову функцію

$$E(u_j(k),\overset{\%}{C}_j) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k) D^2_{PS}(\overset{\%}{X}(k),\overset{\%}{C}_j) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k) Tr(\overset{\%}{X}(k)-\overset{\%}{C}_j)(\overset{\%}{X}(k)-\overset{\%}{C}_j)^T$$

за наявності стандартних обмежень

$$\sum_{j=1}^{m} u_j(k) = 1, \text{ чи } \sum_{j=1}^{m} u_j(k) - 1 = 0,\ k = 1,2,\ldots N,$$

$$0 < \sum_{j=1}^{m} u_j(k) < N,\ j = 1,2,\ldots,m,$$

де $u_j(k)$ – рівень належності матриці $\overset{\%}{X}(k)$ j-му кластеру з матричним центроїдом $\overset{\%}{C}_j$, m – кількість кластерів, що задається апріорно, $\beta > 1$ – параметр фаззіфікації (fuzzyfier), що визначає «розмитість» границь між кластерами.

Результатом кластеризації є $(N \times m)$-матриця $U = \{u_j(k)\}$, що має назву матриці нечіткого розбиття, та m матриць-центроїдів $\overset{\%}{C}_j$, $j = 1,2,\ldots,m$.

Записавши функцію Лагранжа

$$L(u_j(k),\overset{\%}{C}_j,\lambda(k)) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k) Tr(\overset{\%}{X}(k)-\overset{\%}{C}_j)(\overset{\%}{X}(k)-\overset{\%}{C}_j)^T + \sum_{k=1}^{N}\lambda(k)\left(\sum_{j=1}^{m} u_j(k) - 1\right) \quad (5)$$

(тут $\lambda(k)$ – невизначені множники Лагранжа) і розв'язавши систему рівнянь Каруша–Куна–Таккера

$$\begin{cases} \partial L(u_j(k),\overset{\%}{C}_j,\lambda(k))/\partial u_j(k)) = \beta u_j^{\beta-1}(k) Tr(\overset{\%}{X}(k)-\overset{\%}{C}_j)(\overset{\%}{X}(k)-\overset{\%}{C}_j)^T + \lambda(k) = 0, \\ \partial L(u_j(k),\overset{\%}{C}_j,\lambda(k))/\partial\lambda(k) = \sum_{j=1}^{m} u_j(k) - 1 = 0, \\ \{\partial L(u_j(k),\overset{\%}{C}_j,\lambda(k))/\partial \overset{\%}{C}_{jip}\} = -2\sum_{k=1}^{N} u_j^{\beta}(k)(\overset{\%}{X}(k)-\overset{\%}{C}_j) = \mathbf{0} \end{cases}$$

(тут $\{\partial L(\mathrm{u}_j(k),\overset{\%}{C}_j,\lambda(k))/\partial\overset{\%}{C}_{jip}\}$ -$(q \times n)$-матриця, що сформована частковими похідними, 0 – матриця тієї самої розмірності, що утворена нулями), отримуємо результат [10]:

$$\begin{cases} u_j(k) = \dfrac{(Tr(\overset{\%}{X}(k)-\overset{\%}{C}_j)(\overset{\%}{X}(k)-\overset{\%}{C}_j)^T)^{\frac{1}{1-\beta}}}{\sum_{g=1}^{m}(Tr(\overset{\%}{X}(k)-\overset{\%}{C}_g)(\overset{\%}{X}(k)-\overset{\%}{C}_g)^T)^{\frac{1}{1-\beta}}}, \\ \lambda(k) = -\left(\sum_{g=1}^{m}\left(\beta Tr(\overset{\%}{X}(k)-\overset{\%}{C}_g)(\overset{\%}{X}(k)-\overset{\%}{C}_g)\right)^{\frac{1}{1-\beta}}\right)^{1-\beta}, \\ \overset{\%}{C}_j = \dfrac{\sum_{k=1}^{N} u_j^{\beta}(k)\overset{\%}{X}(k)}{\sum_{k=1}^{N} u_j^{\beta}(k)}, \end{cases} \quad (6)$$

близький за $\beta = 2$ до алгоритму Дж. Бездека [7]. Він є його узагальненням для матричного випадку:

$$\begin{cases} u_j(k) = \dfrac{(Tr(\overset{\%}{X}(k) - \overset{\%}{C}_j)(\overset{\%}{X}(k) - \overset{\%}{C}_j)^T)^{-1}}{\sum\limits_{g=1}^{N}(Tr(\overset{\%}{X}(k) - \overset{\%}{C}_g)(\overset{\%}{X}(k) - \overset{\%}{C}_g)^T)^{-1}}, \\ \overset{\%}{C}_j = \dfrac{\sum\limits_{k=1}^{N} u_j^2(k)\overset{\%}{X}(k)}{\sum\limits_{k=1}^{N} u_j^2(k)}. \end{cases} \quad (7)$$

Оскільки матриці $\overset{\%}{C}_j$, $j = 1,2,\ldots m$ є центроїдами кластерів, що утворені рядами різниць, для відновлення центроїдів вихідних даних $\overset{\%}{C}_j$ необхідно скористатися співвідношеннями (2).

Послідовна on-line нечітка кластеризація багатовимірних рядів на основі модифікованої нейро-фаззі мережі Т. Кохонена

Процедури кластеризації (6), (7) введено з припущенням, що всю інформацію задано у вигляді фіксованого масиву даних $X(1), X(2), \mathbf{K}, X(N)$ і вона не змінюється з часом. Якщо ж дискретні поля $X(k)$ надходять на обробку послідовно в формі потоку даних, можна скористатися підходами, що використовуються в Data Stream Mining і насамперед адаптивними методами [9].

Для послідовної обробки даних якнайкраще пристосовані кластерувальні нейронні мережі – самоорганізовні карти Т. Кохонена [11, 12], що дають змогу в on-line режимі самонавчання провести чітке розбиття потоку векторних спостережень. За умов, коли вихідна інформація надходить у формі ($q \times n$)- матричних спостережень класів, що перетинаються, можна скористатися матричною нейро-фаззі кластерувальною мережею [13].

Скориставшись для пошуку сідлової точки лагранжіана (5) рекурентним алгоритмом нелінійного програмування Ерроу–Гурвица–Удзави, можна записати адаптивні процедури кластеризації багатовимірних коротких часових рядів з нерівномірним тактом квантування у вигляді

$$\begin{cases} u_j(k) = \dfrac{(Tr(\overset{\%}{X}(k) - \overset{\%}{C}_j(k-1))(\overset{\%}{X}(k) - \overset{\%}{C}_j(k-1))^T)^{\frac{1}{1-\beta}}}{\sum\limits_{g=1}^{N}(Tr(\overset{\%}{X}(k) - \overset{\%}{C}_g(k-1))(\overset{\%}{X}(k) - \overset{\%}{C}_g(k-1))^T)^{\frac{1}{1-\beta}}}, \\ \overset{\%}{C}_j(k) = \overset{\%}{C}_j(k-1) - \eta(k)\{\partial L(\mathrm{u}_j(k), \overset{\%}{C}_j, \lambda(k)) / \partial \overset{\%}{C}_{jip}\} = \\ = \overset{\%}{C}_j(k-1) + \eta(k)\mathrm{u}_j^{\beta}(k)(\overset{\%}{X}(k) - \overset{\%}{C}_j(k-1)) \end{cases} \quad (8)$$

для довільного значення фаззіфікатора β (тут $\eta(k)$ – параметр кроку навчання) і

$$\begin{cases} u_j(k) = \dfrac{(Tr(\overset{\%}{X}(k) - \overset{\%}{C}_j(k-1))(\overset{\%}{X}(k) - \overset{\%}{C}_j(k-1))^T)^{-2}}{\sum\limits_{g=1}^{m}(Tr(\overset{\%}{X}(k) - \overset{\%}{C}_g(k-1))(\overset{\%}{X}(k) - \overset{\%}{C}_g(k-1))^T)^{-2}}, \\ \overset{\%}{C}_j(k) = \overset{\%}{C}_j(k-1) + \eta(k)u_j^2(k)(\overset{\%}{X}(k) - \overset{\%}{C}_j(k-1)) \end{cases} \quad (9)$$

для β=2.

Нескладно помітити, що з позиції самонавчання кластерувальних мереж Т. Кохонена другі рекурентні співвідношення (8) та (9) є модифікаціями для матричного випадку правила налаштування на основі принципу «Переможець отримує більше» (WTM) [11], де множник $u_j^{\beta}(k)$ виконує роль функції сусідства.

На рис. 2 наведено архітектуру запропонованої адаптивної матричної нейро-фаззі самоорганізовної мережі.

Отже, для розв'язання задачі нечіткої кластеризації багатовимірних часових рядів можна використати архітектури, що є за суттю самоорганізовними мапами з ($q \times n$)-матричним входом і m матричними вузлами.

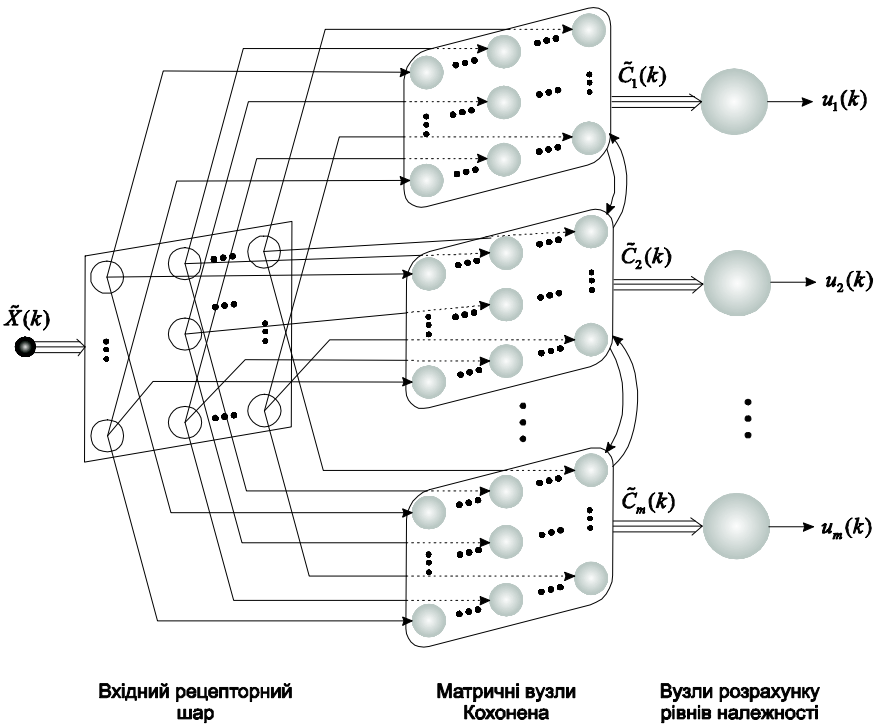

Рис. 2. Архітектура адаптивної матричної нейро-фаззі самоорганізовної мережі

Висновок

Розглянуто задачу нечіткої кластеризації багатовимірних коротких часових рядів з нерівномірним тактом квантування, які можуть бути представлені у формі пакета спостережень чи послідовно надходити на обробку в on-line режимі. В першому випадку можна використати матричну модифікацію методу нечітких С-середніх, а у другому – матричну модифікацію нейро-фаззі мережі Т. Кохонена, що навчається на основі правила «Переможець отримує більше». Запропонована процедура нечіткої кластеризації є достатньо простою в обчислювальній реалізації і може бути використана для широкого класу задач, що пов'язані з Big Data та Data Stream Mining.

1. Liao, T.W. Clustering of time series data-A survey / T. W. Liao // Pattern Recognition. – 2005. – 38. – № 11. – P. 1857–1874. 2. Mitsa, T. Temporal Data Mining / T. Mitsa. -Boca Raton: CRC Press, 2010. – 395 p. 3. Aggarwal, C. C. Data Clustering. Algorithms, and Applications / C. C. Aggarwal, C. K. Reddy. – Boca Raton: CRC Press, 2014. – 621 p. 4. Aggarwal, C. C. Data Mining / C. C. Aggarwal. – N.Y.: Springer, 2015. – 734 p. 5. Möller-Levet, C.S. Fuzzy clustering of short time series with unevenly distributed sampling points / C.S. Möller-Levet, F. Klawonn, K.-H. Cho, O. Wolkenhauer // Lecture Notes in Computers Science. – Heidelberg: Splinger, 2003. – Vol. 2810. – P.330–340. 6. Cruz, L.P. Fuzzy clustering for incomplete short time series data / L. P. Cruz, S. M. Vieira, S. Vinga // Lecture Notes in Artificial Intelligence. – 9273. – Springer Int. Publishing Switzerland, 2015. – P. 353–359. 7. Bezdek, J. C. Pattern Recognition with Fuzzy Objective Function Algorithms / J. C. Bezdek. – N.Y.: Plenum Press. 1981. – 272 p. 8. Höppner, F. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis, and Image Recognition / F. Höppner, F. Klawonn, R. Kruse, T. Runkler. – Chichester: John Wiley and Sons Ltd., 1999. – 289 p. 9. Bifet, A. Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams / A. Bifet. – IOS Press, 2010. – 224 p. 10. Bodyanskiy, Ye. Adaptive matrix fuzzy c-means clustering / Ye. Bodyanskiy, M. Skuratov, V. Volkova // Proc. 19th East-West Fuzzy-Colloquium. – Zittau-Görlitz: HS, 2012. – P.96–103. 11. Kohonen, T. Self-Organizing Maps / T. Kohonen // Berlin: Springer-Verlag. – 1995. – 362 p. 12. Haykin S. Neural Netwroks. A Comprehensive Foundation / S. Haykin. – Upper Saddle River: Prentice Hall, 1999. – 842 p. 13. Bodyanskiy, Ye. Matrix neuro-fuzzy self-organizing clustering network / Ye. Bodyanskiy, M. Skuratov, V. Volkova // Computer Science, Information Technology and Management Science. – 2011. – № 49. – P. 54–58.