УДК 681.32

Р. Н. КВЄТНИЙ, О. О. ДРУЖИНІНА

Вінницький національний технічний університет, Вінниця

МЕТОДИ ФОРМУВАННЯ НАВЧАЛЬНОЇ ВИБІРКИ В ЗАДАЧАХ МОДЕЛЮВАННЯ ЧАСОВИХ РЯДІВ НЕЙРОННИМИ МЕРЕЖАМИ

Аннотація. Здійснено огляд існуючих підходів до формування навчальної вибірки в задачах моделювання часових рядів радіально-базисними нейронними мережами. Розглядаються методи попередньої обробки кількісних ознак навчальної вибірки. Проведені експериментальні дослідження підтверджують доцільність використання додаткових вхідних компонент та попередньої обробки вхідних даних.

Аннотация. Приведен обзор существующих подходов к формированию обучающей выборки в задачах моделирования временных рядов радиально-базисными нейронными сетями. Рассматриваются методы предварительной обработки признаков обучающей выборки. Проведенные экспериментальные исследования подтверждают целесообразность использования дополнительных входных компонент и предварительной обработки входных данных.

Abstract. Existing approaches to the training set generating issue for time series modeling using radial-basis neural networks tasks were analyzed. Methods of training set features preprocessing are considered. Experimental results confirm the reasonability of additional input components using and input data preprocessing.

Вступ

Задача аналізу часових рядів (ЧР) являє собою широку область досліджень у різних областях науки, яка зазнає найбільш стрімкого розвитку останнім часом. Така задача є актуальною для часових рядів, що описують об’єкти різної природи: біржових котирувань акцій, медицини, сейсмології та багато інших.

Класичні методи ідентифікації часових рядів, стають все менш придатними для моделювання складних нелінійних систем. Адекватним апаратом для побудови моделей практично будь-яких нелінійних структур можуть слугувати методи, побудовані на основі штучного інтелекту, а саме штучні нейронні мережі, які мають здатність до моделювання нелінійних процесів, адаптації та дозволяють працювати з зашумленими даними.

Саме таким інструментом є радіально-базисні нейронні мережі (RBF NN – Radial Basis Function Neural Network), які на фоні інших інтелектуальних засобів відрізняються особливо високою швидкістю навчання, і сьогодні широко використовуються для розв’язання задач моделювання часових рядів [1, 2].

Актуальність

В нейромережевому підході задача прогнозування часових рядів може бути сформульована як задача розпізнавання образів. Радіально-базисні нейронні мережі передбачають використання парадигми навчання «з вчителем». Під навчальною вибіркою будемо розуміти сукупність прецедентів – пар «об’єкт, клас», яка подаються на вхід нейроної мережі.

Аналіз літературних джерел показав, що на сьогодні питанню формування навчальної вибріки приділено недостатньо уваги, хоча ефективність моделювання часових рядів у значній мірі визначаться якістю формування навчальної вибірки [1]. Серед праць науковців які присвячені питанням формування навчальної вибірки слід виділити роботи В.А.Крісілова, Н.В. Пєскова, Р.О. Тарасенко, Д. Н. Олешко, О. А. Блажко, Р. І. Франка, Н. Деві, С. П. Ханта, Х. Мурвейта, М. Вайнтрауба, М. Коена [3-11]. Роботи цих вчених, в переважній мірі, присвячені дослідженню і розробці комплексних критеріїв якості навчальної вибірки. Технології, методики та методи, присвячені безпосередньо вирішенню задачі формування набору прецедентів, а не оцінці його якості в наукових працях освітлені не в достатній мірі. Це і визначило актуальність і задачі даного дослідження.

Мета

Метою даного дослідження є підвищення ефективності моделювання часових рядів засобами радіально-базисних нейронних мереж за рахунок підвищення якості навчальної вибірки.

Постановка задачі

Для ідентифікації часових рядів, оцінки поточного моменту і перспектив розвитку методами розпізнавання образів висувається наступна основна гіпотеза. Нехай в момент часу $t = t_0$ існує деякий набір (вектор) факторів процесу, що досліджується:

$$x(t_0) = \{x_1(t_0), x_2(t_0), x_3(t_0), ..., x_N(t_0)\}. \quad (1)$$

Припускається, що внаслідок існування $x(t_0)$ в момент часу $(t_0 + T)$ реалізується набір векторів, які оцінюють розвиток процесу:

$$Q(t_0+T)=\{q_1(t_0+T),q_2(t_0+T),q_3(t_0+T),...,q_m(t_0+T)\}. \quad (2)$$

Тобто існує деяке відображення простору X в простір Q. Це відображення може бути інтерпретоване як функціональна або як кореляційна залежність.

Компоненти $q_j(j=\overline{1,m})$ можуть задовольняти або не задовольняти нормативні умови. Таким чином, за результатами кожного експерименту, фіксуються значення компонент вектора $Q(t_0+T)$ і визначається до якого з класів Bi (i = 1...k) відноситься отриманий результат (табл. 1).

Таблиця 1 - Таблиця відповідності вхідних компонент заданим класам

N експер.	$q_1(t_0)$	$q_2(t_0)$	...	$q_n(t_0)$	$q_1(t_0+T)$	...	$q_m(t_0+T)$	*Клас*
1	x_{11}	x_{21}	...	x_{n1}	q_{11}	...	q_{m1}	$B_{i1}(i=1...k)$
2	x_{12}	x_{22}	...	x_{n2}	q_{12}	...	q_{m2}	$B_{i2}(i=1...k)$
...	...	...	...	...	...	...	...	...
N	x_{1N}	x_{2N}	...	x_{nN}	q_{1N}	...	q_{mN}	$B_{iN}(i=1...k)$

Дані наведеної таблиці можуть використовуватись як навчальна вибірка без попередньої обробки. Але використання «чистих» вхідних даних для навчальної вибірки веде до збільшення розмірності вхідного вектора та часу для навчання нейромережевого класифікатора. У вирішенні цієї проблеми допомагає використання вхідних компонент, які являють собою виділені ознаки. Наявність надлишкових і непотрібних ознак обтяжуватиме навчальний процес, в той час як замала кількість ознак може бути недостатньою для представлення важливих ознак образу.

Задача полягає у аналізі існуючих методів формування навчальної вибірки та дослідженні залежності ефективності ідентифікації часових рядів засобами радіально-базисних нейолнних мереж від методу формування навчальної вибірки.

Розв'язання задачі

Аналіз літературних джерел показав, що зазвичай для опису ситуації вибирається однакова глибина занурення – використовується метод ковзного вікна (mooving/sliding window) зі стаціонарною шириною. Схематично даний метод зображений на рисунку 1.

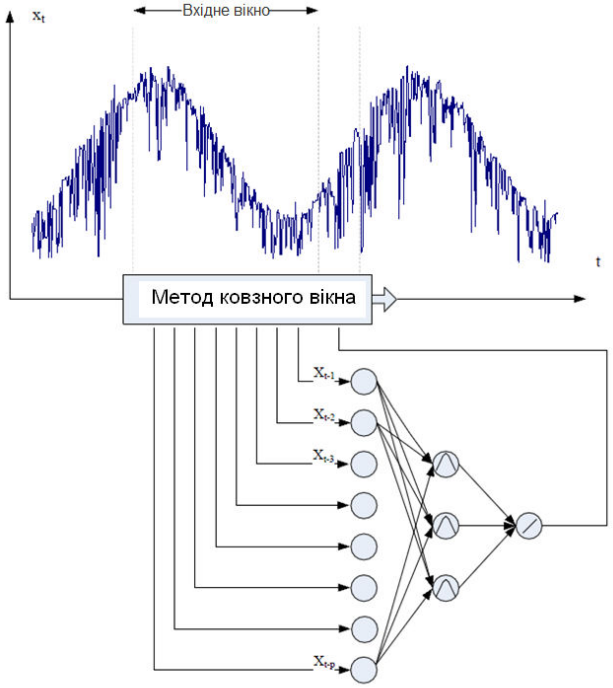

Рисунок 1 – Прогнозування часових рядів методом ковзного вікна

У випадку застосування даного методу необроблені дані спостережень, як правило, перетворюються шляхом центрування і нормалізації або ж подаються на вхід нейронної мережі без попередньої обробки [12-14].

Дані процедури являють собою процедуру лінійного перетворення зразків (Xt) до нормальних стандартних змінних (Zt). Це може бути здійснено шляхом використання наступного рівняння:

$$Z_t = \frac{X_t - \mu}{\sigma}, \tag{3}$$

де μ – середнє відхилення від статистично стабільного процесу, σ – середнє квадратичне відхилення випадкового процесу X_t. За умови стабільного процесу, Z_t задовольняє нормальний розподіл в межах [-3, 3] з нульовим середнім і одиничним середньоквадратичним відхиленням. В роботі [15] зазначається, що навчання НМ буде більш ефективним, якщо дані знаходяться в межах певного діапазону.

Нормалізація зразків (Zt) до компактного інтервалу [0, 1] чи [-1, 1] дає можливість мінімізувати вплив випадкових шумів.

Процедури центрування і нормалізації є корисними у випадках, коли дані змінюються у широкому діапазоні. Процедура нормалізації також часто застосовується для ремасштабування виділених ознак в діапазон між [-1, 1] для представлення в класифікаторах НМ.

Нормалізація до діапазону між [0, 1] може бути виконана шляхом застосування наступного перетворення [16]:

$$Y_t = \frac{Z_t - Z_{\min}}{Z_{\max} - Z_{\min}}, \tag{4}$$

де $Z_{\min}$ – мінімальне значення даних чи виділених ознак; $Z_{\max}$ - максимальне значення даних чи виділених ознак.

Нормалізація до діапазону між [-1, 1] може бути виконана шляхом застосування наступного перетворення [17, 18]:

$$Y_t = \frac{2\left(Z_t - Z_{\min}\right)}{Z_{\max} - Z_{\min}} - 1. \tag{5}$$

В роботах [13, 14] для вхідного представлення даних використовувалось бінарне кодування. Дане кодування являло собою перетворення стандартизованих зразків (Zt) закодованої форми, тоді як зонування являла собою процедуру ре-масштабування і поділу графіку стандартизованих даних на 7 зон (zone+3, zone+2, zone+1, zone0, zone-1, zone-2, zone-3). Наприклад, якщо зразок розташований в зоні zone+2, то бінарне кодування буде представлятись як ‘0100000’. Один зразок потребує сім вхідних нейронів. Тим паче, це потребує відносно великого розміру мережі і збільшення обчислень.

З іншої сторони в роботах [18, 19, 20] було опублікувано іншу технологію обробки даних. Зразки були лінійно перетворені в діапазон між [-7.625, 7.625], який відрізняється від звичайного діапазону стандартизації [-3, 3]. Далі перетворені зразки буди поділені на 61 зону з інтервалом в 0,25 стандартного відхилення. Вони зазначили, що великий діапазон ранжування перетворенних зразків і зонування дозволить ідентифікувати процеси з високими відхиленнями, які досягають 4 стандартних відхилень.

Дослідники в області аналізу часових рядів розглядали множини ознак для стиснення часових рядів. Широко використовувались перетворення Фур’є, що дозволяли підвищити швидкодію даної процедури [21-24], однак цей підхід має декілька недоліків. На практиці, ці перетворення згладжують локальні мінімуми та максимуми, що може призвести до втрати важливої інформації.

В роботі [25] для розв’язання даної задачі пропонується використовувати вейвлети і показані переваги цієї технології над перетвореннями Фур’є.

Однак, аналізуючи нестаціонарні ЧР об’єктів різної природи легко бачити, що для різних ділянок ЧР ця необхідна глибина занурення різна. Таким чином, з однієї сторони на різних ділянках ЧР необхідно формувати образи з різним розміром ситуації, а з іншої сторони розмір вхідних векторів збудження навчальних наборів навчальної вибірки має бути однаковим. Саме це протиріччя створює проблему вибору розміру опису ситуації для нестаціонарних ЧР. Лише в небагатьох джерелах на це звертається увага. Однак, наприклад, в роботі [3] запропоновано модифікований метод формування НВ в

задачах прогнозування ЧР, що відрізняється від методу «вікон» тим, що вхідне «вікно», що задає розпізнаваний образ навчального набору, має перемінний розмір, що залежить від складності ділянки ряду, на якій формується навчальний набір. Для досягнення цього пропонується розділити поняття розміру опису образа і розміру вікна W_i, що переміщається по ЧР. Ділянка ЧР, що розглядається через вікно W_i, розбивається на ділянки, складність яких однакова. Кожна з цих ділянок, узагалі говорячи, має різну довжину. Для кожної з ділянок формується стиснутий опис за рахунок завдання її не переліком значень, а коефіцієнтами апроксимуючої функції. Таким чином розпізнаваний образ має фіксовану довжину і задається вектором стиснутих описів ділянок однакової складності.

Погоджуючись з твердженням, що використання однакового розміру опису ситуації не дозволяє отримати інформативну навчальну вибірку пропонується використовувати сегментований часовий ряд для виділення найбільш важливої інформації з масиву початкових даних, шляхом ідентифікації важливих точок мінімуму та максимуму і вилучення інших незначних коливань [3]. Ступінь сегментації визначається показником *R*, який завжди має перевищувати одиницю, тобто збільшення даного показника буду призводити до зменшення точок (рис 2).

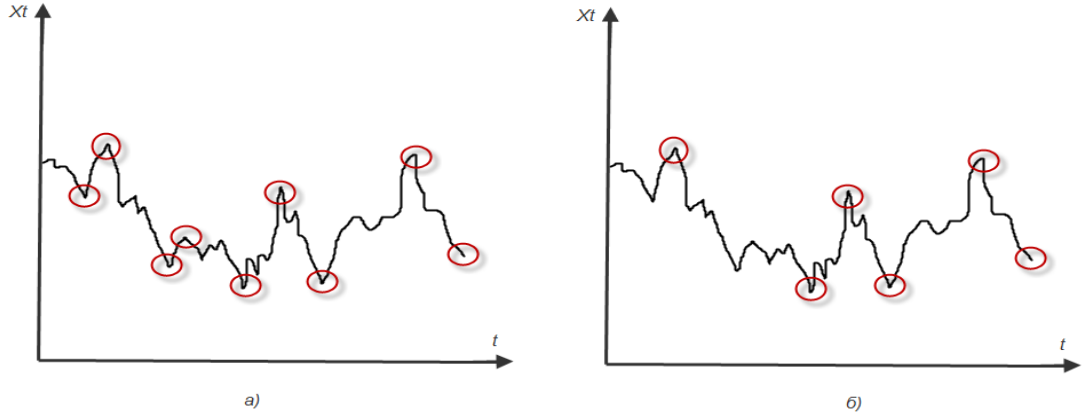

Рисунок 2 – Сегментований часовий ряд: а) R=90%, б) R = 95%

Точка x_{mim} ряду $x_{1,\dots,}x_n$, буде важливими мінімумом (рис. 4 – а), якщо є такі індекси t_1 та t_2, де $t_1<min<t_2$, такі, що:

$$\begin{cases} x_{\min} = \min[x(t_1), x(t_2) \\ \dfrac{x(t_1)}{x_{\min}} \geq R, \\ \dfrac{x(t_2)}{x_{\min}} \geq R. \end{cases} \tag{6}$$

Аналогічно точка a_m ряду $a_{1,\dots,}a_n$, буде важливими максимумом (рис. 4 – б), якщо є такі індекси t_1 та t_2, $t_1<max<t_2$, такі, що:

$$\begin{cases} x_{\max} = \max[x(t_1), x(t_2)], \\ \dfrac{x_{\max}}{x(t_1)} \geq R, \\ \dfrac{x_{\max}}{x(t_2)} \geq R. \end{cases} \tag{7}$$

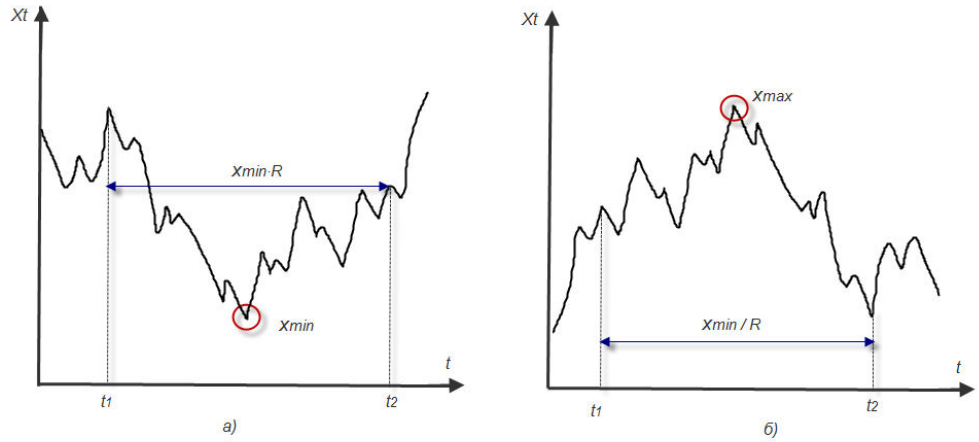

Рисунок 3 – приклади важливого мінімуму (а) і важливого максимуму (б)

Результати експериментальних досліджень

Для аналізу розглянутих підходів та методів щодо підготовки навчальної вибірки було проведено комп’ютерне моделювання. Досліджувалась ефективність ідентифікації часових рядів на основі ймовірнісної нейронної мережі (PNN) з використанням необроблених вхідних даних у навчальній вибірці та навчальних зразків отриманих шляхом попередньої обробки даних та з використанням запропонованого алгоритму сегментації. В результаті застосування запропонованої технології для ідентифікації часових рядів валютних курсів було отримано сигнальний індикатор, значення якого використовувались в режимі реального часу. Сигнальний індикатор у випадку класифікації на 2 класи може приймав 3 можливих значення: 1 – сигнал на купівлю фінансового інструменту, -1 – сигнал на продаж фінансового інструменту, 0 – у випадку відсутності сигналу щодо зімни тенденції.

Достовірність ідентифікації була оцінена за допомогою наступних показників: FRR (помилка першого роду), FAR (помилка другого роду) та Detection Rate (DR). Для задачі ідентифікації ЧР, у випадку класифікації поточної ринкової ситуації на два класи, помилка FRR визначалась як середнє значення пропусків зміни тенденції для класів *Buy* і *Sell* , а помилка FAR як середнє значення хибних виявлень зміни тенденції для цих же класів. Показник Detection Rate (DR) визначався за формулою

$$DR = \frac{кількість\ \ зразків, коректно\ \ класифікованих}{загальна\ \ кількість\ \ зразків\ \ в\ \ множині\ \ даних} \cdot 100\%. \qquad (8)$$

В результаті проведення досліджень була оцінена ефективність технології ідентифікації часових рядів з застосуванням різних моделей НМ на різних часових рядах і інтервалах (табл. 2).

Таблиця 2 – Результати експериментальних досліджень

Часовий ряд	Навчальна вибірка	FRR (%)	FAR (%)	DR (%)
Часовий ряд валютного курсу GBPUSD. Часовий інтервал – 1 год.	побудована на основі тільки вхідних нормалізованих даних	31	18	58
	побудована на основі нормалізованих вхідних даних; додатково враховувались компонента екстремумів та низка технічних індикаторів (Stochastic, MA, RSI).	23	14	72
Часовий ряд споживання мережевого трафіку. Часовий інтервал – 0.5 год	побудована на основі нормалізованих даних	34	15	61
	побудована на основі вхідних нормалізованих даних; видалена сезонна компонента, додатково враховувались компонента екстремумів	25	16	63

Висновки

1. В даній роботі був проведений аналіз існуючих методів формування навчальної вибірки. Аналіз показав, що дослідження методів формування навчальної вибірки та їх впливу на результати нейромережевих моделей є актуальним і необхідним для можливості розширення та уточнення правил організації процесу формування навчальної вибірки.

2. Результати комп'ютерного моделювання підтвердили ефективність використання попередньо оброблених вхідних даних на основі запропонованого алгоритму. Для досліджених в роботі прикладних задач, запропонований підхід дозволив підвищити якість ідентифікації на 11 – 14% для валютних часових рядів і незначне підвищення (2%) було отримано моделюванні часового ряду споживання мережевого трафіку. На основі отриманих даних можна також зробити висновок, що вхідні компоненти, які подаються на вхід нейронної мережі мають бути вибрані в залежності від природи об'єкту, який описує часовий ряд, що досліджується.

Список літератури

1. Квєтний Р. Н. Імовірнісні нейронні мережі в задачах ідентифікації часових рядів [Електронний ресурс] / Р. Н. Квєтний, В. В. Кабачій, О. О. Чумаченко // Наукові праці Вінницького національного технічного університету. – 2010. – № 3. – 6 с. – Режим доступу до журн: http://www.nbuv.gov.ua/e-journals/VNTU/2010-3/2010-3.htm.

2. Чумаченко О. О. Проблема формування навчальної вибірки в нейромережевому моделюванні нелінійних часових рядів : (Тези доповідей XIX міжнародної науково-практичної конференції «Інформаційні технології: наука, техніка, технологія, освіта, здоров'я») / О.О. Чумаченко. – Харків, 2011 – С. 375.

3. Тарасенко Р. А. Выбор размера описания ситуации при формировании обучающей выборки для нейронных сетей в задачах прогнозирования временных рядов / Р. А. Тарасенко, В. А. Крисилов // Труды Одесского политехнического университета. – 2001. – № 2. – С.25 – 28.

4. Тарасенко Р.А. Предварительная оценка качества обучающей выборки для нейронных сетей в задачах прогнозирования временных рядов / Р. А. Тарасенко, В. А. Красилов // Труды Одесского политехнического университета. – 2001. – № 1. – С. 90 – 93.

5. Крисилов В. А. Использование гипотезы λ-компактности при построении обучающей выборки для прогнозирующих нейросетевых моделей [електронний ресурс] / В. А. Крисилов, С. А. Юдин, Д. Н. Олешко. Режим доступу до файлу: http://journal.iasa.kpi.ua/zm456st/2006/No-3/Krisilov_Judin_Ol_N3_06.doc

6. Дюкова Е. В. Поиск информативных фрагментов описаний объектов в дискретных процедурах рапознавания [Електронний ресурс] / Е. В. Дюкова, Н. В. Песков // Журнал вычислительной математики и математической физики. – 2002. – т.42, №5. – С.741-753. Режим доступу до журн.: http://www.ccas.ru/frc/papers/djukova02poisk.pdf

7. Колесникова С. И. Методы анализа информативности разнотипных признаков [Електронний ресурс] / С. И. Колесникова. Режим доступу до файлу: http://www.lib.tsu.ru/mminfo/000063105/inf/06/image/06-069.pdf.

8. Востров Г. Н. Сегментация экономических временных рядов с использованием вейвлет-анализа [Електронний ресурс] / Г. Н. Востров, М. В. Полякова, В. В. Любченко // Труды Одесского политехнического университета. – 2003. – № 1(19). Режим доступу до журн.: http://www.library.ospu.odessa.ua/online/periodic/opu_2003_1(19)/3/3-6.pdf

9. Frank R. J. Input Window Size and Neural Network Predictors [Електронний ресурс] / N. Davey, S. P. Hunt. Режим доступу до файлу: http://homepages.feis.herts.ac.uk/~nngroup/pubs/papers/frank-ijcnn00.pdf

10. Murveit H. Training Set Issues in SRI's DECIPHER Speech Recognition System [Електронний ресурс] / Murveit H., Weintraub M., Cohen M. Режим доступу до файлу: http://acl.ldc.upenn.edu/H/H90/H90-1065.pdf

11. Preprocessing Time Series Data for Classification with Application to CRM / [Yiming Yang, Qiang Yang, Wei Lu, Jialin Pan, Rong Pan, Chenhui Lu, Lei Li, and oth]. Режим доступу до файлу: http://www1.i2r.a-star.edu.sg/~jspan/publications/[AI05]Preprocessing%20Time%20Series%20Data%20for%20Classification%20with%20Application%20to%20CRM.pdf

12. Christophe Paoli Solar radiation forecasting using ad-hoc time series preprocessing and neural networks [Електронний ресурс] / Christophe Paoli, Cyril Voyant, Marc Muselli, Marie-Laure Nivet. Режим доступу до файлу: http://arxiv.org/ftp/arxiv/papers/0906/0906.0311.pdf

13. Hwarng H. B. Detecting process non-randomness through a fast and cumulative learning ART-based pattern recognizer / H. B. Hwarng, C.W. Chong // International Journal of Production Research. – 1995. – № 33. – PP. 1817-1833.

14. Hwarng H. B. X-Bar Chart pattern recognition using neural nets / H. B. Hwarng, N. F. Hubele // ASQC quality congress transactions. – 1991. – PP. 884-888.

15. Barghash M. A. Pattern Recognition of Control Charts Using Artificial Neural Networks - Analysis the Effects of the Training Parameters / M. A. Barghash, N. S. Santarisi // Journal of Intelligent Manufacturing. – 2004. – №15. – PP. 635-644.

16. Gauri S. K. Improved recognition of control chart patterns using artificial neural networks / S. K. Gauri, S. Chakraborty// International Journal of Advanced Manufacturing Technology. –2008. –№ 36. – PP. 1191-1201.

17. Eamonn Keogh Segmenting time series: a survey and novel approach [Електронний ресурс] / Eamonn Keogh, Selina Chu, David Hart, Michael Pazzani. Режим доступу до файлу: http://www-scf.usc.edu/~selinach/segmentation-slides.pdf

18. Guh R. S. On-Line identification of control chart patterns using selforganizing approaches / Guh R.S., and Shiue, Y.R. // International Journal of Production Research. – 2005. №43. – PP. 1225-1254.

19. Guh R.S. On-Line identification and quantification of mean shifts in bivariate processes using a NN-Based approach/ R. S. Guh // Quality and Reliability Engineering International.– 2007. – № 23. – PP. 367-385.

20. Sameer Singh Dynamic time-series forecasting using local approximation / Sameer Singh, Paul McAtackney // In Proceedings of the Tenth IEEE International Conference on Tools with Artificial Intelligence. – 1998. – PP. 392 –399.

21. David S. Stoffer Detecting common signals in multiple time series using the spectral envelope / David S. Stoffer // Journal of the American Statistical Association. – 1999. – PP. 1341–1356.

22. Online data mining for co-evolving time series / [Byoung-Kee Yi, Nikolaos D. Sidiropoulos, Theodore Johnson, H. V. Jagadish, Christos Faloutsos, and Alexadros Biliris] // In Proceedings of the Sixteenth International Conference on Data Engineering. – 2000. – PP. 13-22.

23. Hara K. A training data selection in online-training for multilayer neural networks / K. Hara, K. Nakayama, A. A. Kharaf // Proc. IEEE IJCNN’1998. –1998. – PP.2247-2252.

24. Klevecka I. Leli Pre-Processing of input data of neural networks: the case of forecasting telecommunication network traffic [Електронний ресурс] / I. Klevecka, Leli Janis. Режим доступу до файлу: http://www.telenor.com/en/resources/images/168-178_Pre-processingInput Data-ver1_tcm28-36193.pdf

25. Wang C.H. An Integrated Approach for process monitoring using wavelet analysis and competitive neural network / C.H.Wang, W. Kuo, H.Qi // International journal of production research. –2007. – №45. – PP.227-244.

Відомості про авторів

Квєтний Роман Наумович – д.т.н., проф., завідувач кафедри АІВТ. Вінницький національний технічний університет, м. Вінниця, вул. Хмельницьке шосе 95, (0432) 598243.

Дружиніна Ольга Олегівна – аспірант кафедри АІВТ. Вінницький національний технічний університет, м. Вінниця, вул. Хмельницьке шосе 95, (0432) 598243, oo.druzhinina@gmail.com.