Грицюк П.М. Застосування методу найближчих сусідів до короткострокового прогнозування ціни акцій / П.М. Грицюк // Вісник Української академії банківської справи. - 2013.– № 1(34).– С. 54-59.

УДК 004.8:338.27

***П. М. Грицюк**, д-р екон. наук, доц., завідувач кафедри економічної кібернетики Національного університету водного господарства та природокористування*

ЗАСТОСУВАННЯ МЕТОДУ НАЙБЛИЖЧИХ СУСІДІВ ДО КОРОТКОСТРОКОВОГО ПРОГНОЗУВАННЯ ЦІНИ АКЦІЙ

У статті розглянуті особливості застосування методу найближчих сусідів до короткострокового прогнозування ціни акцій. Для підвищення ефективності методу введені і використані поняття "гомогенні найближчі сусіди" та "компетентні паттерни".

Ключові слова: часовий ряд, паттерн, найближчий сусід, справджуваність прогнозу, гомогенні паттерни, компетентні паттерни.

Постановка проблеми. Розробка методів прогнозування економічної динаміки має важливе значення для ефективного планування розвитку економіки та своєчасного передбачення кризових явищ [1–3]. Головним критерієм якості прогнозної моделі вважається її похибка. Однак не завжди прогнозна модель, яка характеризується хорошою точністю, здатна успішно передбачати зміни в динаміці процесу. Тому, крім похибки, важливими характеристиками прогнозної моделі є такі показники, як кореляція між фактичною динамікою процесу та прогнозними значеннями і справджуваність прогнозу – частка правильно передбачених знаків зміни прогнозованого показника. Роль справджуваності вперше була відзначена в роботах І. Б. Загайтова та Л. П. Яновского [4]. Для оцінювання якості прогнозної моделі використовують метод ретроспективного аналізу [1]: на минулому матеріалі виконують прогнозування і, порівнюючи результати з відомими даними, розраховують похибку та справджуваність прогнозу.

Аналіз останніх досліджень і публікацій. Результати діяльності економічних систем відображаються у вигляді часових рядів їх параметрів. Кожен часовий ряд є сумішшю детермінованої та хаотичної компоненти. Після роботи Лоренца [5] стало зрозумілим, що хаотична поведінка системи може бути викликана спеціальним видом нелінійності, присутньої у законах динаміки еволюціонуючої системи. У цьому випадку кажуть про динамічний хаос. Початком математичного дослідження хаосу в курсах валют стали роботи Брока [6], в яких досліджені курси валют на американських та канадських біржах. У роботі [7] було виявлено наявність динамічного хаосу та нелінійності для обмінних курсів валют 12 країн Південної Азії. Прогнозам динаміки часових рядів методами теорії хаосу присвячені роботи російських дослідників [8, 9].

Невирішені раніше частини загальної проблеми. Одним з ефективних методів дослідження нелінійних систем є метод найближчих сусідів [10]. Метод спирається на ідею про повторюваність фрагментів часового ряду, породженого динамічною системою. Обмеженість методу полягає в тому, що його застосовують лише для аналізу стаціонарних часових рядів. Невирішеною проблемою залишається адаптація методу для нестаціонарного випадку. Метою даного дослідження є удосконалення методу найближчих сусідів (НС) шляхом введення поняття "гомогенні найближчі сусіди", врахування зв'язку факторів Open–Close та оптимізації таких параметрів моделі, як кількість найближчих сусідів та розмірність паттерна.

У роботі [11] розроблені нові прогнозні моделі та адаптовані класичні методи для прогнозування динаміки процесів зерновиробництва в Україні. Ці процеси належать до класу процесів з пам'яттю, оскільки в них присутня регулярна повторюваність фрагментів динаміки. Оскільки міжрічні зміни врожайності зернових в Україні можуть досягати 50–100 %, точність прогнозування врожайності при середній похибці 15–20 % вважається хорошою. Щодо справджуваності прогнозів врожайності, то І. Б. Загайтов і Л. П. Яновский вважають необхідним досягнення рівня 80 % для практичної застосовності моделі.

На відміну від динаміки врожайності, динаміка ціни акцій є переважно стохастичним процесом з незначними проявами детермінізму. Такий характер динаміки ускладнює задачу побудови прогнозної моделі. В той же час варіація ціни акцій є набагато меншою від варіації врожайності зернових. Тому побудова прогнозної моделі ціни фінансового активу з середньою похибкою 2–3 % не є занадто складною задачею. В таких умовах вирішальним критерієм якості стає не

похибка прогнозу, а його справджуваність. Адже тільки правильне передбачення знаку зміни ціни акції дозволяє отримати прибуток при біржовій торгівлі. Тому головною метою даної роботи є побудова прогнозної моделі, яка забезпечує максимальну справджуваність прогнозів.

У даній роботі розглянуто застосування різних варіантів прогнозних моделей НС для випадку однокрокового прогнозу тижневих значень ціни акцій. Статистика ціни акцій, зазвичай, представляється у форматах Open, High, Low, Close. Тому при прогнозуванні значення Close з горизонтом 1 тиждень, допустимим буде використання не лише попередніх зачень Close, а й нового значення ціни Open, яке стає відомим на початку прогнозного тижня. В той же час недопустимим є використання факторів High і Low, значення яких визначаються лише в кінці прогнозного тижня.

Виклад основного матеріалу. Алгоритм методу найближчих сусідів ґрунтується на порівнянні відомих статистичних даних з новими елементами. Для нового запису, продовження якого необхідно спрогнозувати, знаходять найбільш подібні записи у минулому та ідентифікують їх як найближчих сусідів. Метод найближчих сусідів спирається на так звану "гіпотезу компактності" [12]. Ця гіпотеза стверджує, що різні відображення одного і того ж образу у просторі ознак породжують геометрично близькі точки, утворюючи компактні "згустки".

Умовою ефективного застосування методу НС є вдалий вибір наступних параметрів: вид метрики, яка служить для оцінки близькості між об'єктами; розмірність навчальних зразків (паттернів), кількість K найближчих сусідів, які беруться до уваги при побудові прогнозу. Вибір малого значення параметра K приведе до сильного розкиду значень прогнозу; велике значення K може спричинити сильну зміщеність моделі. Отже, значення K повинно бути достатньо великим, щоб мінімізувати ймовірність помилкової класифікації, і досить малим, щоб K близьких сусідів розташовувалися досить близько до точки запиту.

В ролі об'єктів, близькість яких аналізується, ми обрали паттерни.

Означення 1. Паттерном u_i розмірності D називається послідовність сусідніх елементів часового ряду $u_i = x_{i-D+1}, x_{i-D+2}, \ldots, x_i$. Тут i – номер паттерна, D – його розмірність.

Вдалий вибір параметра D важливий для ефективного прогнозування. При малих значеннях D знайдеться багато близьких сусідів, які недостатньо повно відображатимуть локальну динаміку процесу зміни ціни. При великих значеннях D відбувається нівелювання поняття близького сусіда через те, що відстані між різними парами досліджуваних зразків будуть мало відрізнятися (прокляття розмірності). Отже, необхідно підібрати оптимальне значення розмірності D, яке є достатнім для ідентифікації характерних паттернів і не є занадто великим для того, щоб їх "знеособити". Для вибору оптимальних значень параметрів K і D використовують алгоритм кросвалідації [13], який полягає у багаторазовому повторенні прогнозування на відомих даних із щоразовим зміщенням базового відрізка моделі на один крок вперед. Результуючі усереднені значення похибки та справджуваності прогнозу дозволяють уникнути випадковості оцінки моделі.

Ще одним оптимізуючим параметром є глибина ретроспекції R. Мале значення параметра R приведе до недостатньо надійних оцінок якості моделі. При великих значеннях R стануть помітними неоднорідності динаміки часового ряду на різних ділянках ретроспекції. Однак глибина ретроспекції впливає лише на точність оцінки середньої похибки та справджуваності, і не впливає на ефективність прогнозування. Логічно вважати мінімальним значенням глибини ретроспекції R період 10 тижнів, який дозволяє оцінити параметри з точністю до 10 %. Максимальне значення глибини ретроспекції становить 20–25 тижнів. При більших значеннях R якість моделі погіршується через змішування різних типів динаміки, для кожного з яких властиві різні оптимальні параметри прогнозування.

Технологія прогнозування. Проаналізуємо можливості різних варіантів методу найближчих сусідів щодо прогнозування часових рядів ціни акцій. Задача полягає у побудові прогнозу z_{n+1} для продовження часового ряду $X = x_1, x_2, \ldots, x_n$. Побудова прогнозу методу НС здійснюється шляхом простого або зваженого усереднення виходів K найближчих сусідів. Зазвичай, вагу W найближчих сусідів оцінюють по їх відстані до замикаючого паттерна, обчисленій з використанням обраної метрики. Будемо оцінювати близькість двох паттернів y_i та y_j через евклідову відстань d_{ij} між ними згідно з означенням

$$d_{ij} = \sqrt{(x_{i-D+1} - x_{j-D+1})^2 + (x_{i-D+2} - x_{j-D+2})^2 + \ldots + (x_i - x_j)^2} \,. \qquad (1)$$

Основна ідея методу найближчих сусідів полягає в тому, що близькі паттерни на короткому відрізку часу еволюціонують однаково. Для того, щоб оцінити очікувану зміну замикаючого паттерна $y_n = (x_{n-D+1}, x_{n-D+2}, \ldots, x_n)$, знайдемо K найближчих до нього паттернів $y_{n1}, y_{n2}, \ldots, y_{nK}$ згідно з критерієм (1). У процесі еволюції системи ці паттерни переходять у паттерни $y_{n1+1}, y_{n2+1}, \ldots, y_{nK+1}$, які утворюються в результаті ковзання вікна шириною D вздовж часового ряду на один елемент вперед. Кожен паттерн $y_m = (x_{m-D+1}, x_{m-D+2}, \ldots, x_m)$, який є близьким сусідом замикаючого паттерна y_n, продукує частковий прогноз x_{m+1}. У найпростішому випадку прогноз z_{n+1} визначається простим усередненням часткових прогнозів

$$z_{n+1} = \frac{1}{K}(x_{n1+1} + x_{n2+1} + \ldots + x_{nK+1}) \qquad (2)$$

Більш точний прогноз можна отримати як зважене усереднення часткових прогнозів згідно із співвідношенням

$$z_{n+1} = \sum_{i=1}^{K} w_i x_{ni+1} / \sum_{i=1}^{K} w_i \qquad (3)$$

Тут w_i – ваги, які приписуються кожному з часткових прогнозів. Для обчислення ваги використовують значення відстані d_{in} від замикаючого паттерна до *i*-го близького сусіда, розрахованої згідно із співвідношенням (1). Обчислення ваги виконують за співвідношенням

$$w_i = \max - d_{in}, i = \overline{1, K} \qquad (4)$$

Тут $\max$ – максимальне значення відстані до замикаючого паттерна на множині прогнозуючих паттернів. Таким чином, максимальній відстані відповідатиме нульова вага, а мінімальній відстані – максимальна вага. Вага визначає прогнозну силу даного паттерна.

Використовуючи описаний вище підхід, нами була побудована модель найближчих сусідів (модель 1), призначена для однокрокового прогнозування ціни акцій. Модель була протестована на відрізку часового ряду тижневих значень ціни акцій (Close) корпорації Exxon Mobil Corporation. Довжина відрізка складала 25 тижнів: від 2176-го рівня ряду (12.09.2011) до 2200-го рівня ряду (27.02.2012). В результаті комп'ютерних експериментів було визначено оптимальне значення розмірності фазового простору $D=7$ та кількості "найближчих сусідів" K=6. Середня похибка прогнозування за описаною моделлю складає 3,6 %, справджуваність прогнозу – 50 %, коефіцієнт кореляції між фактичними та прогнозними значеннями – 79,8 %. Усереднені значення оцінок моделі визначалися методом ретроспективної крос-валідації на контрольному відрізку. Оскільки справджуваність прогнозів є низькою, то якість отриманої моделі слід також вважати низькою. Причиною цього є нестаціонарність ряду ціни, в результаті чого вдається знайти дуже мало близьких сусідів – відрізків часового ряду із значенням ціни, близьким до ціни на завершальному етапі спостережень.

Модель гомогенних паттернів. Для покращення якості моделі найближчих сусідів необхідно розширити прогнозну базу – кількість найближчих сусідів. З цією метою введемо поняття гомогенних паттернів.

Означення 2. Два паттерни $u = (u_1, u_2, \ldots, u_D)$ і $v = (v_1, v_2, \ldots, v_D)$ називаються гомогенними, якщо виконується співвідношення $u_i / v_i = h \quad (i = \overline{1, D})$. Коефіцієнт h називають коефіцієнтом гомогенності паттерна u по відношенню до паттерна *v*.

У реальних часових рядах, які є переважно стохастичними, гомогенні паттерни зустріти надзвичайно складно. Тому доцільно ввести дещо слабше поняття квазігомогенних паттернів.

Означення 3. Два паттерни $u = (u_1, u_2, \ldots, u_D)$ і $v = (v_1, v_2, \ldots, v_D)$ називаються квазігомогенними, якщо виконується співвідношення

$$h \cdot (1-\alpha) < u_i / v_i < h \cdot (1+\alpha) \quad i = \overline{1, D}, \quad \alpha > 0, \alpha << 1 \qquad (5)$$

Поняття квазігомогенних паттернів дозволяє значно розширити базу прогнозування у методі найближчих сусідів. Тепер, порівнюючи два паттерни y_m і y_n, необхідно, перш за все, визначити середнє значення коефіцієнта гомогенності

$$h_{mn} = \frac{1}{D} \sum_{i=1}^{D} \frac{x_{m-D+i}}{x_{n-D+i}} \qquad (6)$$

Якщо при цьому виконується співвідношення (5), паттерни можна вважати близькими сусідами. Модифікований евклідовий критерій близькості двох векторів y_m та y_n тепер матиме наступний вигляд (7):

$$d_{mn} = \sqrt{(x_{n-D+1} - x_{m-D+1} / h_{mn})^2 + (x_{n-D+2} - x_{m-D+2} / h_{mn})^2 + \ldots + (x_n - x_m / h_{mn})^2} \qquad (7)$$

Використовуючи нову метрику, побудуємо удосконалений алгоритм методу найближчих сусідів з урахуванням поняття гомогенності. Частковий прогноз на базі *m*-го паттерна тепер будується на основі співвідношення

$$z_{m+1} = x_{m+1} / h_{mn} . \qquad (8)$$

Тут h_{mn} – коефіцієнт гомогенності паттерна – "найближчого сусіда" y_m по відношенню до кінцевого паттерна часового ряду y_n. Узагальнений прогноз будується за співвідношенням (3).

Результати прогнозування з використанням моделі гомогенних найближчих сусідів (модель 2) є наступними.

Оптимальні параметри моделі залишилися незмінними: розмірність паттерна $D = 7$; кількість "найближчих сусідів" $K = 6$. Середня похибка прогнозування, визначена на відрізку від 2176-го рівня ряду до 2200-го рівня, складає 2,3 %; справджуваність прогнозу – 72,5 %; коефіцієнт кореляції між фактичними та прогнозними значеннями – 90,8 %.

У порівнянні з моделлю 1 помітно збільшилася кореляція і справджуваність прогнозів. Це пояснюється тим, що "гомогенних найближчих сусідів" є набагато більше, ніж звичайних найближчих сусідів. Розширення прогнозної бази привело до покращення якості прогнозної моделі.

Компетентні паттерни. Можливі два варіанти локальної динаміки ряду ціни для поточного тижня. При першому варіанті, якщо значення ціни Open перевищує (є нижчим) попереднє значення ціни Close, то і прогнозоване значення ціни в кінці поточного тижня Close1 також перевищує (є нижчим) попереднє значення ціни Close. Такий хід процесу назвемо трендовим. При другому варіанті, якщо значення ціни Open перевищує (є нижчим) попереднє значення ціни Close, то прогнозоване значення ціни в кінці поточного тижня Close1 буде нижчим (перевищить) від попереднього значення ціни Close. Такий хід процесу назвемо реверсивним.

Для подальшого покращення якості моделі ми використали ідею, запропоновану в роботах Н. Г. Загоруйка [12]. Всі паттерни, які використовуються для отримання локальних прогнозів, поділяються на два класи: "компетентні" та "некомпетентні". Якщо паттерн видає частковий прогноз, який по типу динаміки співпадає з динамікою Close–Open, він вважається компетентним. Локальний тип динаміки (трендова, реверсивна) визначається методом ретроспекції на контрольному відрізку (25 тижнів), який передує прогнозованому тижню.

Роль компетентних і некомпетентних прогнозуючих паттернів у побудові прогнозу визначається введенням додаткового вагового множника wk_i. Якщо для передісторії поточного тижня є характерною трендова поведінка, всі часткові прогнози, які відповідають трендовій динаміці, домножуються на підсилюючий коефіцієнт $wk_i = 4$. Якщо ж превалює реверсивний тип динаміки, часткові прогнози, які відповідають трендовій динаміці, домножуються на зменшуючий коефіцієнт $wk_i = 1/2$.

$$wk_i = \begin{cases} 4, & \text{якщо } (xo_{i+1} - x_i) * (x_{ni+1} - x_i) \ge 0 \text{ і трендова динаміка}; \\ 1/2, & \text{якщо } (xo_{i+1} - x_i) * (x_{ni+1} - x_i) \ge 0 \text{ і реверсивна динаміка}; \\ 1, & \text{якщо } (xo_{i+1} - x_i) * (x_{ni+1} - x_i) < 0; \end{cases} \qquad (9)$$

Тут xo_{i+1} значення ціни Open на початку нового прогнозного тижня, x_{ni+1} – значення часткового прогнозу, x_i – значення ціни Close в кінці попереднього тижня. Значення множників були підібрані нами шляхом комп'ютерних експериментів. Остаточний прогноз отримується шляхом зваженого усереднення часткових прогнозів згідно із співвідношенням

$$z_{n+1} = \sum_{i=1}^{K} w_i wk_i x_{ni+1} / \sum_{i=1}^{K} w_i wk_i \qquad (10)$$

Результати прогнозування за методом компетентних найближчих сусідів (модель 3) представлені на рис. 1. Середня похибка прогнозування на відрізку від 2176-го рівня ряду до 2200-го рівня ряду складає 2,3 %, справджуваність прогнозу – 84 %, коефіцієнт кореляції між фактичними та прогнозними значеннями складає 90,3 %. Якість моделі суттєво залежить від локальної динаміки ціни акцій. Рис. 2 демонструє рівень справджуваності прогнозів, який дозволяє досягти метод компетентних найближчих сусідів шляхом оптимізації параметрів *D* і K на різних інтервалах часового ряду.

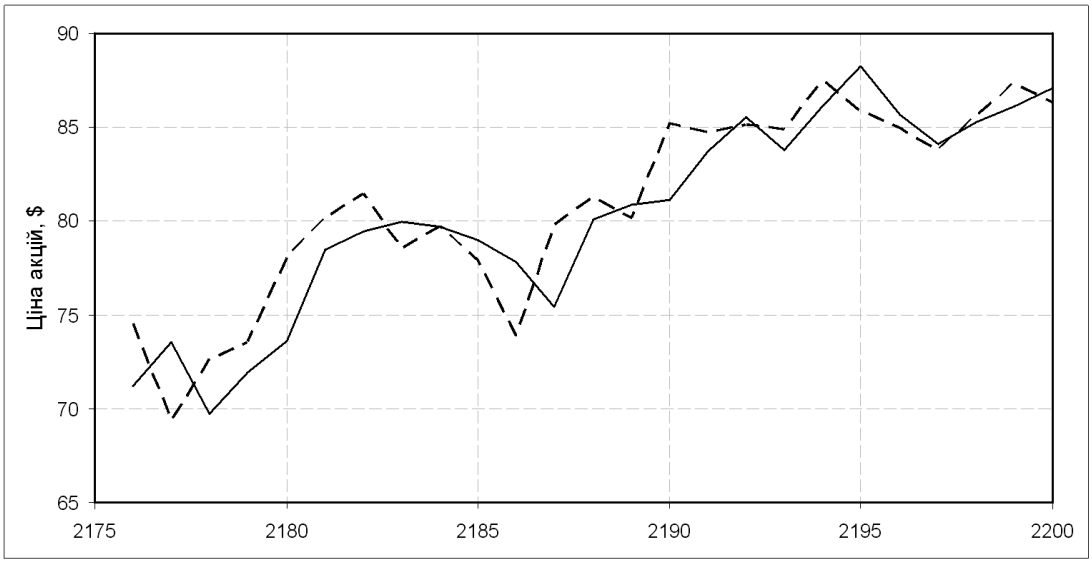

Рисунок 1 – Прогнозування ціни акцій корпорації Exxon Mobil Corporation. Метод компетентних найближчих сусідів

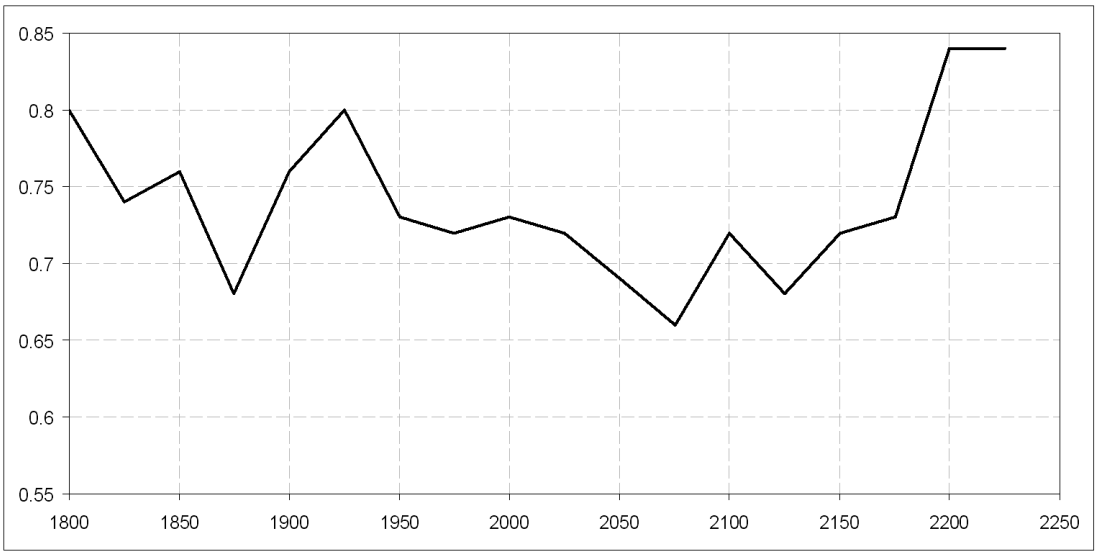

Рисунок 2 – Справджуваність прогнозів, отриманих за моделлю компетентних найближчих сусідів, на різних ділянках часового ряду

Висновки. Завдяки введенню поняття "гомогенні найближчі сусіди" та розділенню прогнозуючих паттернів на компетентні та некомпетентні нам вдалося підняти рівень справджуваності прогнозу від 50 до 70–80 %. Враховуючи випадковий характер поведінки ціни, отримані результати прогнозування можна розцінити як успішні, а розроблений алгоритм прогнозування рекомендувати для практичного використання. Таким чином, використання поняття гомогенних компетентних паттернів дозволяє підвищити ефективність використання методу найближчих сусідів для прогнозування нестаціонарних часових рядів ціни акцій.

Список літератури

1. Ганчук А. А. Методи прогнозування / А. А. Ганчук, В. М. Соловйов, Д. М. Чабаненко. – Черкаси : Брама-Україна, 2012. – 140 с.
2. Клебанова Т. С. Модели прогнозирования и анализа кризисних явлений в экономике / Т. С. Клебанова, Л. С. Гурьянова, Е. А. Сергиенко // Сучасні проблеми прогнозування соціально-економічних процесів: концепції, моделі, прикладні аспекти. – Бердянськ, 2012. – С. 58–73.
3. Моделі і методи соціально-економічного прогнозування : підручник / В. М. Геєць, Т. С. Клебанова, О. І. Черняк та ін. – Х. : ВД "ІНЖЕК", 2008. – 396 с.

4. Яновский Л. П. Принципы, методология и научное обоснование прогнозов урожая по технологии “ЗОНТ”: монография / Л. П. Яновский. – Воронеж : ВГАУ, 2000. – 376 с.
5. Lorenz E. N. Deterministic non-periodic flow / E. N. Lorenz // Journal of Atmospheric Science. – 1963. – V. 20. – P. 120–141.
6. Brock W. A. Distinguishing random and deterministic systems / W. A. Brock // Journal of Economic Theory. – 1986, Vol. 40. P. 168–195.
7. Das A. Chaotic analysis of the foreign exchange rates / A. Das, P. Das // Applied Mathematics and Computations, 2007. 31. – P. 388–396.
8. Безручко В. П. Математическое моделирование хаотических временних рядов / В. П. Безручко, Д. А. Смирнов. – Саратов : Гос. УНЦ “Колледж”, 2005. – 532 с.
9. Григорьев В. П. Исследование математической модели фьючерсных рынков / В. П. Григорьев, А. В. Козловских, Д. А. Марьясов // Рынок ценных бумаг, 2005. – № 9 (288). – С. 38–42.
10. McNames J. Innovations in local modeling for time series prediction / J. McNames // Ph.D. Thesis, Stanford University, 1999.
11. Грицюк П. М. Аналіз, моделювання та прогнозування динаміки врожайності озимої пшениці в розрізі областей України / П. М. Грицюк. – Рівне : НУВГП, 2010. – 350 с.
12. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. – Новосибирск : ИМ СО РАН, 1999. – 270 с.
13. Bishop C. M. Neural Networks for Pattern Recognition / C. M. Bishop. – Oxford University Press, 1995. – P. 504.

Отримано 29.04.2013

Summary

This paper discusses the peculiarities of nearest neighbors method to the short-term forecasting stock prices. To improve the efficiency of the method introduced and used terms “homogeneous nearest-neighbor” and “competent pattern.”