УДК 004.048

Т.В. КОРЧАК[1], О.С. РАДИВОНЕНКО[1], А.М. СКАКОВСЬКА[2]

[1] Національний аерокосмічний університет ім. М.Є. Жуковського «ХАІ», Україна
[2] Сумський державний університет, Україна

ОБРОБЛЕННЯ НЕВИЗНАЧЕНОСТІ В МЕДИЧНИХ ЧАСОВИХ РЯДАХ ДЛЯ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ПРОГНОЗУВАННЯ

Робота присвячена задачам аналізу та прогнозування часових рядів в умовах невизначеності. Модифіковано та розширено метод Скланські-Гонзалеза з метою застосовування лінгвістичного узагальнення трендів для аналізу медичних часових рядів, що містять викиди та пропуски. Було запропоновано зручне подання мети дослідження оброблення часових рядів. Докладно подано характерні ознаки медичних даних. Метод Скланські-Гонзалеза був використаний для опису характеристики часових рядів за динамікою зміни, тривалістю та мінливістю з подальшим обробленням часових рядів. Було подано функціональну систему первинного оброблення медичних даних і моделі Брауна для прогнозування часових рядів.

Ключові слова: *прогнозування, епідемічні дані, адаптивні методи, нейро-нечіткі моделі, динаміка захворюваності.*

Вступ

Медичні часові ряди мають широкий діапазон характеристик і невизначеностей, які відрізняють їх від інших рядів даних. Більш поглиблене вивчення проблеми оброблення часових рядів дозволяє побачити складний процес, спричинений багатьма невизначеностями, які існують в теорії медичних даних [1]. Далі наведено приклад основних видів невизначеностей, що містить в собі медична інформація:

1) неминуча невизначеність – виявляється в мінливості властивостей середовища;

2) статистична помилка - неточність при зборі даних та обробленні;

3) невизначеність при моделюванні - це результат припущень, зроблених під час проведення аналізу або викликаних використанням спрощених моделей.

Характерною особливістю медичних часових рядів є також частота збору даних. Наприклад, в області епідеміології у більшості випадків зустрічаються частоти, такі, як: щоденні, щотижневі, щомісячні, квартальні та річні. Інформація про будь-яку інфекцію надходить кожен день і може бути додатково подана в абсолютних, відносних і середніх показниках. Крім того, часові ряди в медицині розрізняються за відстанню між датами, тобто відрізняють повні (дати реєстрації захворюваності йдуть одна за одною) і часткові (коли принцип рівних інтервалів не дотримується) часові ряди. У реальних умовах динаміки захворюваності отримані часові ряди часто є неповними. Це може відбуватися з кількох причин:

1) несвоєчасне подання інформації про ряд випадків;

2) низька якість ретроспективного аналізу;

3) неможливість установлення причини виникнення захворюваності.

Кожна з цих причин сприяє падінню числа зареєстрованих хворих. Відсутні дані ускладнюють побудову моделі часових рядів та оцінки її параметрів [2].

У разі коли відсутні значення у часових рядах зустрічаються рідко, було запропоновано такі підходи:

1) використання функцій інтерполяції даних, що базуються на загальній тенденції часового ряду;

2) використання простих формул, таких, як загальне, локальне та сукупне середнє;

3) скорочення часових рядів без урахування відсутніх даних.

Ці підходи мають свої недоліки, вони видозмінюють і модифікують первинну структуру часових рядів та ускладняють побудову моделі. Тому актуальною проблемою є розроблення і застосування таких моделей, які могли б відновити відсутні дані без модифікації або втрати інформації про початковий часовий ряд.

Таким чином, першим кроком було запропоновано попередню обробку часового ряду, щоб уникнути невизначеності на етапі прогнозування.

1. Вирішення проблеми

Аналіз часових рядів з використанням лінгвістичної апроксимації

Необхідно проаналізувати наявні статистичні дані - щомісячні абсолютні показники захворюваності. Використовуючи набір властивостей, треба знайти тренди типу - швидке зростання, повільне

зростання, повільно убувають, незмінні, тобто отримати лінгвістичний опис ряду. Необхідно також оцінити такі параметри, як динаміка змін, тривалість і мінливість.

Під динамікою зміни тренда розуміють швидкість зміни. Її може описати нахил прямої, що відображає тренди. Були виділені такі лінгвістичні терми, що відповідають різним нахилам прямої (рис. 1): швидке спадання; повільне спадання; незмінність; повільне зростання; швидке зростання. На рисунку зображено області, що відповідають специфічним лінгвістичним термам. Фактично, кожен терм являє собою гранулу напрямів.

Рис. 1. Візуалізація гранул напрямів, що визначають динаміку мінливості

Використаємо підхід до лінгвістичної апроксимації, запропонований в роботі [3]. Цей підхід базується на рівномірній частково-лінійній апроксимації часових рядів з використанням модифікації ефективного алгоритму Скланські та Гонзалеза.

Функція f є рівномірною ε-апроксимацією часового ряду або безліччю пар точок $\{(x_i, y_i)\}$. Якщо $\varepsilon > 0$ то $\forall i : |f(x_i) - y_i| \le \varepsilon$ якщо функція f – лінійна, то така апроксимація – лінійна рівномірна ε-апроксимація. Конструкція алгоритму – перетин конусів, починаючи з точки p_i часового ряду, що містить круги радіусу ε довкола подальших базових точок p_{i+j}, $j = 1,2...$ доти, поки перетин усіх конусів, що починаються з p_i, не буде пустим. Якщо для p_{i+k} перетин є пустим, то ми будуємо новий конус, що починається в p_{i+k-1} (рис. 2) [4].

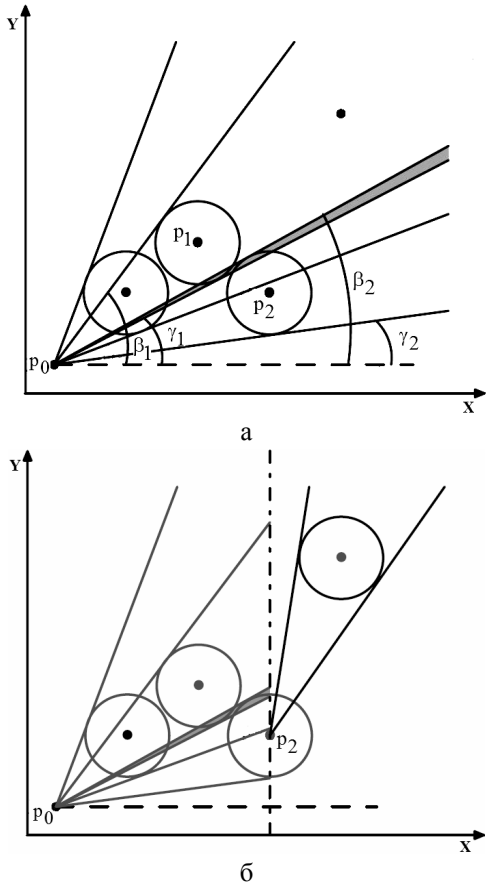


Рис. 2. Робота алгоритму:
а – перетин конусів, що зазначений темно-сірою областю; б - побудова нового конусу, що починається в точці p_2

Метод простий та ефективний, оскільки потребує лише єдиного проходу через дані. У даному підході, підсумовуючи тренди даних часових рядів, визначаються такі аспекти, як динаміка, тривалість і мінливість трендів.

Апробація і тестування запропонованого методу передоброблення часових рядів виконано на основі тестових наборів даних TSDL [5], результати наведено на рис. 3.

Системи нейро-нечіткого виводу Такагі-Сугено

Використання систем нейро-нечіткого виводу зумовлено тим фактом, що вони поєднують роботу нейронної мережі та нечіткої системи і тим самим поєднують переваги обох. За рахунок нейронної мережі можливості нечітких систем можуть бути розширені, що дозволить посилити їхню пристосовуваність і швидкість. Застосування адаптивних нечітких систем і систем на базі нейронних мереж є широко відомими, що здатні на самонавчання.

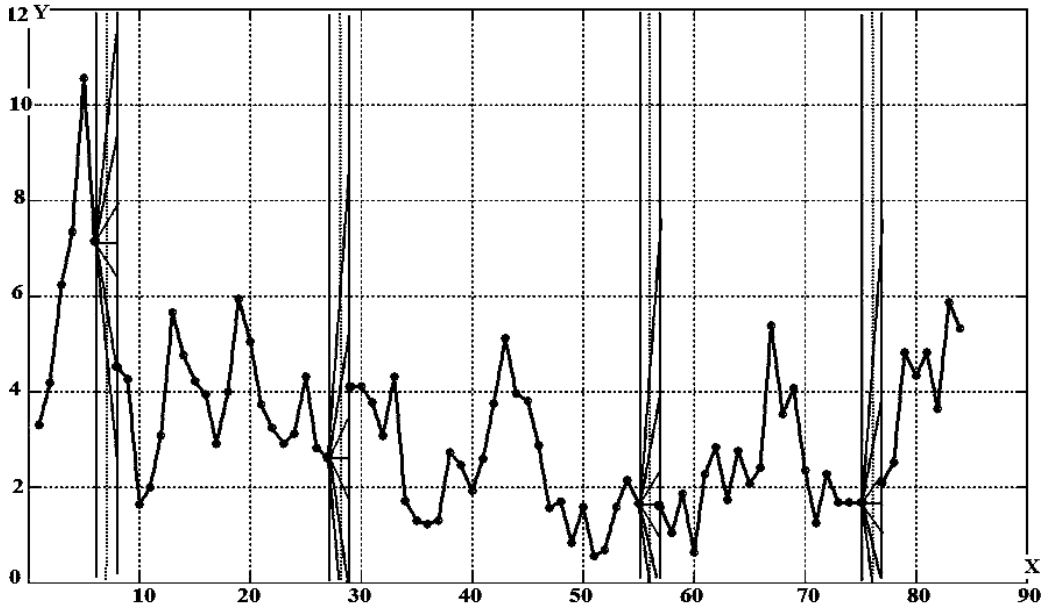
В нечітких системах відношення входу та виходу подається у вигляді правил ЯКЩО-ТОДІ. В нейронних мережах зв'язок подається не в чіткій формі, а в формі закодованих параметрів мережі. Нейро-нечіткі моделі поєднують прозорість нечіткої системи з можливістю навчання за допомогою нейронної мережі.

Якщо-тоді правила для системи нечіткого виводу можуть бути подані у такому вигляді:

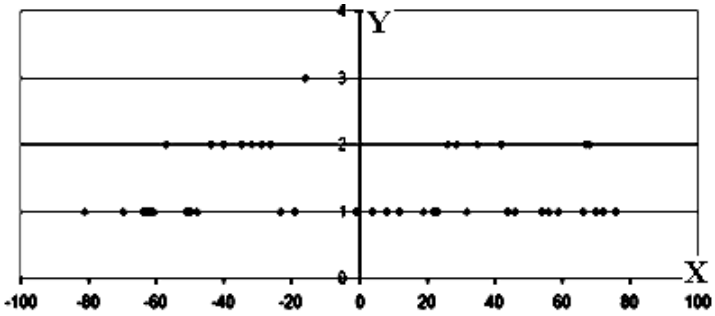
$$\text{IF } x_1 \text{ IS } L_1 \ \&\ x_2 \text{ IS } L_2 \ \&$$

$$\&\ x_n \text{ IS } L_n \text{ THEN } y = f(x), \qquad (1)$$

де $X = [x_1, x_2, ...x_n]^T$, $L_1, L_2, ..., L_n$ – нечіткі множини вхідних спостережень; y – функція, що залежить від вхідних параметрів $x_1, x_2, ...x_n$.



а



б

Рис. 3. Результати лінгвістичної апроксимації трендів медичних часових рядів:
а - тестовий набір даних з пропусками; б - мінливість трендів; Y – кількість трендів, X – кут нахилу

В статті розглядається нейро-нечітка модель типу Такагі-Сугено, яка використовується для встановлення залежності вихідних параметрів від вхідних, що придатна для вирішення проблеми прогнозування. Ця модель може бути подана у вигляді нечітких правил і припущень, побудованих на цих правилах. Правила можуть бути подані в такій формі [6]:

$$\text{IF } \ldots, \text{THEN } y_{k+n} = f(y_k, \ldots, y_{k+n-mn}), \qquad (2)$$

де m – число змінних; n – затримка.

Прогнозування часового ряду з використанням моделі Брауна

Для прогнозування епідемічної ситуації необхідні такі моделі, що були б здатні відображати динамічні властивості часового ряду, що змінюються у часі і дають можливість враховувати інформаційну цінність його членів. При прогнозуванні звичайно висувається гіпотеза про те, що основні взаємозв'язки і тенденції збережуться на період прогнозу або їх можна обґрунтувати і врахувати напрямок їхніх змін у розглянутій перспективі. На практиці найбільш широко застосовується адаптивний метод Брауна, здатний прогнозувати періодичні коливальні процеси. В адаптивних методах прогнозування використовуються часові характеристики параметрів, взаємозв'язки послідовних членів часових рядів. Відповідні моделі перебудовуються з урахуванням нової інформації.

Нами проаналізовано різні методи прогнозування інфекційної захворюваності за допомогою математичних моделей, що описують закономірності епідемічного процесу для різних умов його розвитку. Найбільш достовірною виявилася верифікація прогнозу, основаного на методі експонентного згладжування (модель Брауна).

Розглянемо модель Брауна, коли тренд описується параболою другого порядку. Експонентні середні:

$$\begin{aligned} S'_t &= \alpha x_t + (1-\alpha) S'_{t-1}; \\ S''_t &= \alpha S'_t + (1-\alpha) S''_{t-1}; \\ S'''_t &= \alpha S''_t + (1-\alpha) S'''_{t-1}, \end{aligned} \qquad (3)$$

де t – поточний час; x_t – фактичне значення часового ряду в час t; α – згладжуючий параметр, $\alpha = \text{const}$, $0 < \alpha < 1$; S'_t, S''_t, S'''_t – експоненціальне

середнє першого, другого та третього ступенів відповідно в час t.

Обчислення початкових величин відбувається за допомогою таких формул:

$$S_0' = a_0 - \frac{\beta}{\alpha}a_1 + \frac{\beta(2-\alpha)}{2\alpha^2}a_2;$$
$$S_0'' = a_0 - \frac{2\beta}{\alpha}a_1 + \frac{\beta(3-2\alpha)}{\alpha^2}a_2; \quad (4)$$
$$S_0''' = a_0 - \frac{3\beta}{\alpha}a_1 + \frac{3\beta(4-3\alpha)}{2\alpha^2}a_2,$$

де $\beta = 1-\alpha$.

Щоб виразити значення коефіцієнтів a_0, a_1, a_2 необхідно брати коефіцієнти рівняння лінії тренду, отримані методом найменших квадратів. Прогноз на момент часу t розраховуємо за формулою

$$\hat{x}_t = \bar{a}_0 + \bar{a}_1 t + \frac{1}{2}\cdot\bar{a}_2 t^2. \quad (5)$$

Оцінки коефіцієнтів параболічної залежності для тренду:

$$\bar{a}_0 = 3S_t' - 3S_t'' + S''';$$
$$\bar{a}_1 = \frac{\alpha}{2\beta^2}[(6-5\alpha)S_t' - (10-8\alpha)S_t'' + (4-3\alpha)S_t''';$$
$$\bar{a}_2 = \frac{\alpha}{\beta^2}[S_t' - 2S_t'' + S_t'''].$$

Для здійснення прогнозу методом Брауна первісний часовий ряд було реконструйовано за таким правилом:

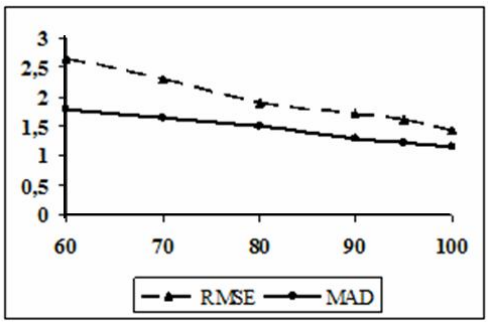
$$X = \begin{pmatrix} X^1 \\ X^2 \\ \cdots \\ X^{12} \end{pmatrix} = \begin{pmatrix} x_i^1 & x_{i+L}^1 & x_{i+2L}^1 \cdots x_{i+nL}^1 \\ x_i^2 & x_{i+L}^2 & x_{i+2L}^2 \cdots x_{i+nL}^2 \\ \cdots & & \\ x_i^{12} & x_{i+L}^{12} & x_{i+2L}^{12} \cdots x_{i+nL}^{12} \end{pmatrix},$$

де X_i^n – стан системи в час i; n – кількість років; L – запізнення або зміщення. Для здійснення прогнозу часового ряду використовується така формула:

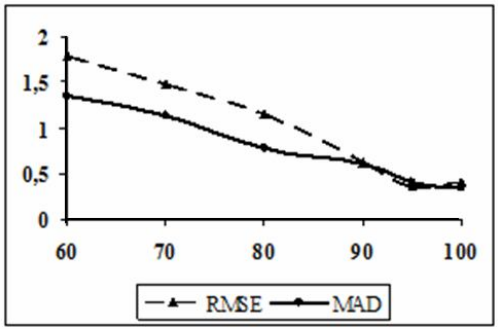
$$X \to \begin{pmatrix} x_{i+(n+1)L}^1 \\ x_{i+(n+1)L}^2 \\ \cdots \\ x_{i+(n+1)L}^{12} \end{pmatrix}. \quad (6)$$

2. Результати

Для проведення численних досліджень часовий ряд спостережень за захворюваністю було проаналізовано на наявність пропусків. Було розглянуто кілька вибірок, кожна з котрих становила більше двох тисяч даних. У результаті було побудовано гістограму розподілення пропусків в даних. Далі, з первісного часового ряду у випадковому порядку було виключено 5%, 10%, 20%, 30% та 40% даних (рис. 5). Спрогнозовані дані отримано як результат методу, представленого в даній роботі та порівняно з фактичними значеннями



а



б

Рис. 5. Залежність похибки прогнозування від кількості вхідних даних: a – результати моделі Брауна; b – результати нейро-нечіткої моделі Такагі-Сугено

Практичним результатом поданої роботи є програмна система. Архітектуру системи подано на рис. 6.

Розроблений пакет програм дозволяє автоматизувати роботу лікаря епідеміолога. Ретроспективний та оперативний аналізи виконуються за допомогою VBA після отримання звітів СУБД Кларіон. Оперативний аналіз містить методи, за якими обчислюють генерацію послідовностей, інтенсивність і приріст показників захворюваності [7]. У процесі ретроспективного аналізу виконуються аналіз сезонних коливань, тенденцій, а також лінгвістична апроксимація. Потім, базуючись на отриманих даних, здійснюється прогнозування за допомогою моделей Сугено (рис. 7) та Брауна (рис. 8).

Висновки

В роботі представлена проблема аналізу та прогнозування часових рядів, що містять пропуски. Наведено порівняння нейро-нечіткого методу Такагі-Сугено та адаптивної моделі Брауна. Першим кроком було запропоновано попередню обробку часового ряду методом Скланскі–Гонзалез, що дозволило уникнути проблему невизначеності на період прогнозування та покращити ефективність результатів прогнозування

Рис. 6. Функціональна модель системи

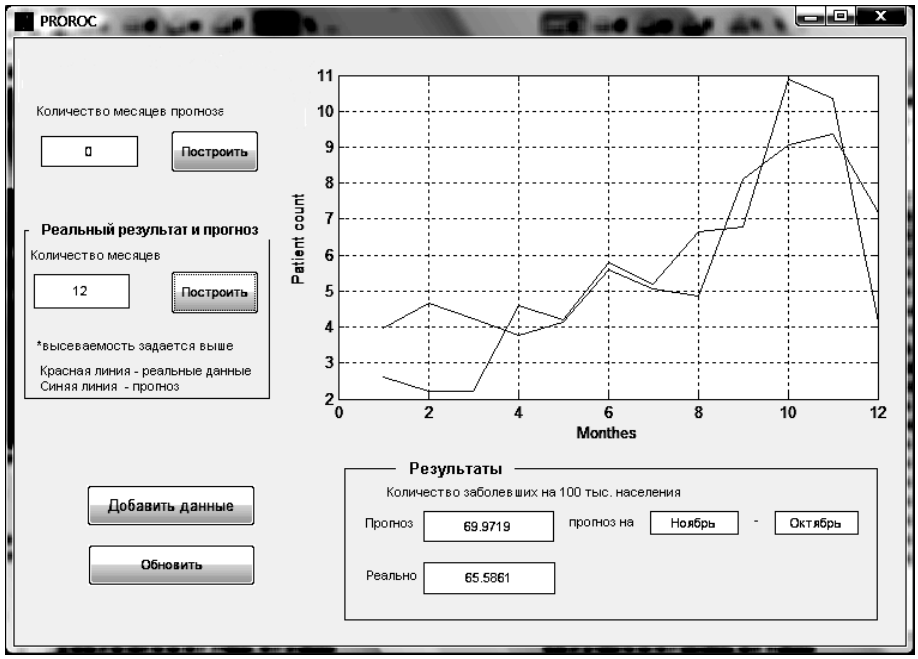


Рис. 7. Прогнозування часового ряду на базі моделі Сугено

Прогнозування рівнів захворюваності та їх реалізація на **2008** **рік**

Кількість років: N = 8 ОБЧИСЛИТИ

Фактичні дані

1	2	3	4	5	6	7	8	9	10	11	12
Січень	Лютий	Березень	Квітень	Травень	Червень	Липень	Серпень	Вересень	Жовтень	Листопад	Грудень
2.20	4.60	4.20	5.80	5.20	6.6	6.8	10.9	10.3	4.1	3.3	3.3

Прогноз

4.0615	3.9217	2.7222	5.6239	4.3800	5.7076	7.2750	8.8182	7.6820	5.5494	4.1578	4.0071

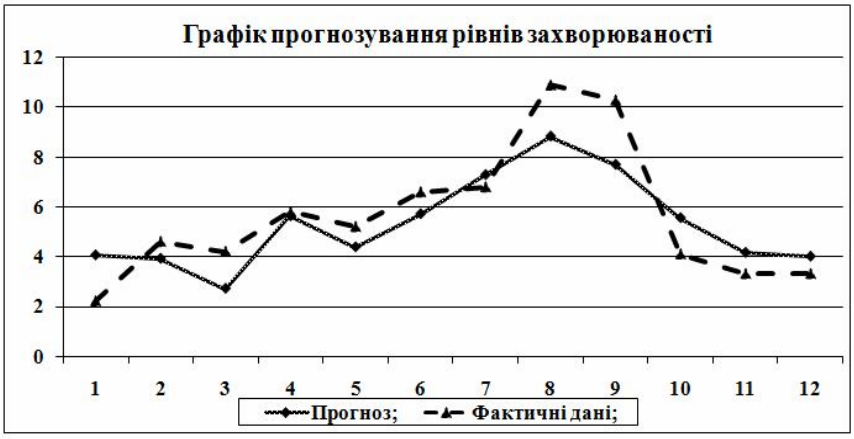


Рис. 8. Прогнозування часового ряду на базі моделі Браун

Література

1. Gail M.H. Encyclopedia of Epidemiologic Methods (Wiley Reference Series in Biostatistics) / M.H. Gail, J.B. Gail. – Wiley: 1 edition, 2000. – 978 p.

2. Allison P.D. Missing Data (Quantitative Applications in the Social Sciences) / P.D. Allison. – Sage Publications, Inc: 1 edition, 2001. – 104 p.

3. Kacprzyk J. Linguistic summarization of time series using a fuzzy quantifier driven aggregation / J. Kacprzyk, A. Wilbik, S. Zadrozny // Fuzzy Sets and Systems. – 2008. – №12 (159). – P. 1485 – 1499.

4. Sklansky J. Fast polygonal approximation of digitized curves / J. Sklansky, V. Gonzalez // Pattern Recognition. – 1980. – №12 (5). – P. 327–331.

5. Хайндман Р. Бібліотека часових рядів [Електронний ресурс]: містить дані о 800 часових рядах з різних галузь науки / Р. Хайндман. – Режим доступу : www.robjhyndman.com/TSDL.

6. Sokolov A.Y. Fuzzy prediction of short-time series / A.Y. Sokolov, T.V. Korchak, O.A. Plyasunova // Proceedings of West Fuzzy Colloquium EUSFLAT'2008. – Zittau: IPM. – 2008. – P. 215-222.

7. Epidemiology and medical statistics / R.R. Calyampudi, C.R. Rao, J.P. Miller, C. R. Dabeeru. – Elsevier Science, North Holland, 2008. – 852 p.

Надійшла до редакції 15.09.2010

Рецензент: д-р техн. наук, проф. каф. інформатики М.Л. Угрюмов, Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ», Харків, Україна.

ОБРАБОТКА НЕОПРЕДЕЛЕННОСТИ В МЕДИЦИНСКИХ ВРЕМЕННЫХ РЯДАХ ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ПРОГНОЗИРОВАНИЯ

Т.В. Корчак, О.С. Радивоненко, А.Н. Скаковская

В данной работе рассмотрены задачи анализа и прогнозирования временных рядов с пропущенными данными. Был модифицирован и расширен метод Скланси и Гонзалеза с целью применения лингвистического обобщения трендов для анализа медицинских временных рядов в условиях неопределенности, такие, как выбросы, отсутствующие значения. Был предложен подход удобного представления задач обработки временных рядов. Подробно изложены характерные признаки медицинских данных. Метод Скланси и Гонзалеза был использован для определения характеристик временных рядов по динамике изменения, продолжительности и изменчивости с последующей предварительной обработкой временных рядов. Была представлена функциональная система первичной обработки временного ряда и модель Брауна для прогнозирования временных рядов медицинских данных.

Ключевые слова: прогнозирование, эпидемиологические данные, адаптивные методы, нейро-нечеткие модели, динамика заболеваемости.

MEDICAL TIME SERIES MANIPULATION UNDER UNCERTAINTY CONDITIONS FOR IMPROVEMENT OF FORECASTING EFFICIENCY

T.V. Korchak, O.S. Radyvonenko, A.N. Skakovskaya

In this paper problem of analysis and forecasting of time series with missing data is considered. The previous works were reformulated and extended towards applying the linguistic summarization of trends for medical time series analysis under uncertainties as outliers, missing values. A convenient presentation of problems of the time series processing was proposed and focused on some characteristic featured to medical data. Sklansky and Gonzalez method was used to characterize the time series by the dynamics of change, duration, and variability with subsequent time series preprocessing. An application to the preprocessing and Brawn model forecasting of medical time series was presented.

Key words: forecasting, epidemiologic data, adaptive method, neuro-fuzzy models, morbidity dynamics.

Корчак Тетяна Вікторівна – асистент каф. інформатики, Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ», Харків, Україна, e-mail: kotavi@i.ua.

Радивоненко Ольга Сергіївна – канд. техн. наук, доцент каф. інформатики, Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ», Харків, Україна, e-mail: oradivonenko@gmail.com.

Скаковська Алла Миколаївна – канд. техн. наук, старший викладач каф. інформатики, Сумський державний університет, Суми, Україна.