УДК 004.94:510.3 DOI 10.31548/energiya2022.06.104

ЗАСТОСУВАННЯ ТЕОРІЇ НЕЧІТКИХ МНОЖИН ДРУГОГО ТИПУ ДО ПРОГНОЗУВАННЯ КАТЕГОРІАЛЬНИХ ЧАСОВИХ РЯДІВ: МАТЕМАТИЧНА МОДЕЛЬ ТА АЛГОРИТМ РЕАЛІЗАЦІЇ

О. С. Тимчук, кандидат технічних наук, доцент

E-mail: oleg.s.tymchuk@gmail.com

А. І. Пилипенко, кандидат технічних наук, доцент

E-mail: annapylypenko4@gmail.com

О. Г. Іванченко, кандидат технічних наук, асистент

E-mail: ivanchenko.oleksii@gmail.com

Київський національний університет імені Тараса Шевченка

Анотація. *Розглянуто проблему прогнозування категоріальних часових рядів. Такі ряди мають широке практичне застосування майже у всіх сферах діяльності, де застосовуються судження та експертні оцінки. Аналіз сучасних досліджень показав, що залишається недостатньо вивченою проблема врахування нечіткості, яка притаманна словесним оцінкам.*

Мета роботи полягає у розробці моделі нечіткого часового ряду, яка дозволить виконувати обчислення зі словами.

У статті застосовується теорія нечітких множин другого типу, в якій приймається, що ступінь приналежності елемента універсальної множини до нечіткої підмножини є також нечіткою на відрізку $[0,1]$. *Модель нечіткого часового ряду другого типу надає результат у вигляді гранульованого терму, який описується словом і дискретною інтервальною нечіткою множиною другого типу.*

На базі запропонованої моделі розроблено нечіткий алгоритм прогнозування часових рядів, який складається з п'яти кроків: визначення моделі слів; фазіфікація значень часового ряду; визначення нечітких відносин; нечітке прогнозування; дефазіфікація результатів прогнозування.

Висока якість запропонованої прогнозної моделі підтверджена трьома оціночними характеристиками: середня абсолютна помилка прогнозу; середньоквадратична помилка прогнозу; середня відносна помилка прогнозу. У подальшому запропонована модель може бути розвинута для розв'язання практичних задач обчислень зі словами при прийнятті рішень.

Ключові слова: ***часові ряди, категоріальні змінні, нечіткі множини другого типу, невизначеність, обчислення зі словами, нечітке прогнозування***

Актуальність. Під категоріальним часовим рядом (Categorical Times Series, CTS) будемо розуміти такий часовий ряд, у якому спостереження в кожен момент

часу мають категоріальні значення (номінальні або порядкові). Очевидно, що такі ряди на практиці досить поширені, хоча в літературі їм приділяють набагато меншу увагу, ніж рядам з числовими неперервними змінними. Вже класичним прикладом CTS став набір даних зі станом сну у новонароджених дітей [1]. Дитячий невролог оцінював електроенцефалограму малюка кожну хвилину упродовж приблизно двох годин. Невролог класифікував стан сну немовляти як одне з наступних: qt - тихий сон (quiet sleep, trace alternant); qh - тихий сон (quiet sleep, high voltage); tr - перехідний сон (transitional sleep); al - активний сон (active sleep, low voltage); ax - активний сон (active sleep, high voltage); aw – прокинувся (awake). Не менш відомим прикладом CTS є часовий ряд з геномною ДНК, що міститься в хромосомах і є частиною генома [2]. Опубліковано ряд статей, в яких увагу приділяють пошуку залежностей у ДНК та тестуванню відмінностей у трендах між кодуючими та некодуючими ділянками ДНК. Джерелом CTS можуть бути як пристрої, так і експертні судження. Наприклад, з web-серверу отримують IP-адреси, web-адреси, коди регіонів; у результаті медичного обстеження фіксують діагнози; при розпізнаванні мови опрацьовують послідовності букв та слів тощо.

Аналіз останніх досліджень та публікацій. Протягом останніх 20 років минулого століття було запропоновано ряд різних підходів до моделювання CTS. Здебільшого ці моделі базувалися на ланцюговій моделі Маркова і дискретній моделі ARMA. Особлива увага приділялася теорії узагальнених лінійних моделей, авторами якої є П. МакКаллаг і Дж.Нелдер [McCullagh, Peter; Nelder, John (1989). Generalized Linear Models]. Параметричні методи аналізу часових рядів передбачають, що в основі даних лежить стаціонарний процес. CTS вважається стаціонарним, якщо граничний розподіл даних є постійним протягом періоду часу, за який вони були зібрані, і кореляція між послідовними значеннями є функцією лише від їх відстані один від одного, а не від їхнього положення в ряду. Однак є багато прикладів категоріальних рядів, які не відповідають цьому визначенню стаціонарності. Ґрунтовний аналіз нестаціонарності CTS наведено в роботі Хайнц Кауфманн [Kaufmann, Heinz (1987). Regression Models for Nonstationary Categorical Time Series: Asymptotic Estimation Theory].

В останніх публікаціях автори продовжують вивчати проблему стаціонарності та максимізації правдоподібності CTS, автокореляційні функції для номінальних та порядкових даних, пропонується розширення базових дискретних моделей авторегресії. Також для прогнозування CTS використовують класифікаційно-регресійні дерева, які можуть працювати як з неперервними, так і з категоріальними даними. Перевагами цих методів є те, що вони є простими у розумінні та інтерпретації, але є проблема перенавчання дерева, надмірна чутливість до вихідних даних, навіть невелика зміна в даних може істотно змінити структуру дерева.

Аналіз сучасних досліджень показав, що залишається недостатньо вивченою проблема врахування в повній мірі невизначеності, яка притаманна судженням та експертним оцінкам з тієї причини, що слова для різних експертів позначають різне. Точність та надійність прогнозування розроблених моделей ще потребую доопрацювання та покращення.

Мета дослідження - розробка моделі нечіткого часового ряду, яка дозволить виконувати обчислення зі словами. На підставі розробленої моделі запропонувати нечіткий алгоритм прогнозування категоріальних часових рядів та оцінити якість прогнозування.

Матеріали і методи дослідження. У теорії нечітких множин ступінь приналежності елемента універсальної множини до нечіткої підмножини може бути будь-яким дійсним числом, що лежить між 0 і 1. У цій статті застосовується теорія нечітких множин другого типу (T2 FS), в якій приймається, що і сама ця ступінь приналежності не визначається для кожного елемента універсальної множини однозначно, а допускається певна розмитість. Іншими словами, ступінь приналежності є нечіткою і належить відрізку $[0,1]$. У загальному вигляді модель нечіткого часового ряду другого типу (T2 FTS – type-2 fuzzy time series) може бути представлена так:

$$v^k(t) = v^j(t-1) \circ R(t-n), \quad v^k(t) \in V, v^j(t-1) \in V, \forall k,j \in \{1,\dots,m\}, \tag{1}$$

де $v^k(t)$ – прогнозоване значення, подане у вигляді гранульованого терму, $v^j(t-1)$ – крайнє значення в T2 FTS, V – словник, m – кількість гранульованих

термів у словнику, $R(t-n), n=\overline{1,N}$ – відношення, яке описує нечіткі відносини між $V(t)$ і $V(t-n)$, N – кількість значень часового ряду, $^\circ$ – композиційне правило виведення Заде (MAX MIN).

Гранульований терм описується словом і дискретною інтервальною нечіткою множиною другого типу (DIT2FS).

$$V = \langle v_j \rangle, \qquad j \in \{1, \dots, m\},$$

$$v_j = \langle T_j, \tilde{Y}_j \rangle,$$

де v_j – гранульований терм,

T_j – слово,

$\tilde{Y}_j$ – DIT2FS, описувальне слово.

Запропонована модель (1) дозволила розробити нечіткий алгоритм прогнозування часових рядів, який ґрунтується на базових принципах теорії нечітких множин другого типу [3], обчислень зі словами [4] та підходу до прогнозування нечітких часових рядів [5]. Розглянемо п’ять основних кроків запропонованого алгоритму.

Крок 1: Визначення моделі слів. При інтерпретації часового ряду та прогнозованого значення людям властиво переходити від числових значень до категоріальних, наприклад: мало, трохи, помірно, добре, багато. У розробленому алгоритмі використовуються моделі слів (codebook), запропоновані J.M. Menel [6]. Для опису значення допускаються codebooks, які складаються з 32, 16, 15, 11, 6, 5 і 3 слів (гранульованих термів). Вибір codebook залежить від рівня категоріальної деталізації. Кожен гранульований терм представляється як пара – слово і DIT2FS, що визначено на універсальній множині $X[0,10]$.

Крок 2: Фазіфікація значень часового ряду. Значення часового ряду та універсальна множина обраного codebook приводяться до нормованого інтервалу [0,1] за формулою

$$x^N = \frac{x - x_{min}}{x_{max} - x_{min}},$$

де x^N – нормоване значення, x – реальне значення, x_{min}, x_{max} – мінімальне та максимальне значення відповідно.

Для часового ряду:

$$x_{min} = X_{min} - dx_1,$$
$$x_{max} = X_{max} - dx_2,$$

де X_{min}, X_{max} – мінімальне та максимальне значення часового ряду, dx_1, dx_2 – можливі відхилення у меншу та більшу сторони.

Для codebook: $x_{min} = 0$, $x_{max} = 10$.

Кожному нормованому значенню ставиться у відповідність активований гранульований терм із codebook, який визначається за формулою:

$$MAX(y_j), j = \overline{1,M},$$
$$y_j = \frac{lmf(\tilde{Y}_j) + umf(\tilde{Y}_j)}{2},$$

де y_j – ступінь належності нормованого значення часового ряду до j-го гранульованого терму з codebook, $lmf(\tilde{Y}_j), umf(\tilde{Y}_j)$ – нижня та верхня функції приналежності DIT2FS, яка описує j-й гранульований терм.

Крок 3: Визначення нечітких відносин. Нечіткі відносини визначаються як нечіткі відносини другого порядку Chen's [7].

Крок 4: Нечітке прогнозування. Нечітке прогнозування визначається за допомогою композиційного правила виведення Заде (MAX MIN).

Крок 5: Дефазіфікація результатів прогнозування. Результатом прогнозування може бути як числове значення, так і слово з codebook. Для отримання результату в числовому вигляді потрібно здійснити зниження типу $\tilde{Y}(t)$. Зниження типу може бути виконано за допомогою алгоритму EKM для DIT2FS [8]. Для отримання результату у вигляді слова з codebook, необхідно виконати порівняння на подібність $\tilde{Y}(t)$ з DIT2FS, які описують слова codebook. Для порівняння може бути використана міра подібності Jaccard для DIT2FS.

Результати досліджень та їх обговорення. Для оцінки запропонованої моделі необхідно перевірити, що ряд помилок прогнозу є білим шумом з нульовим середнім значенням. Для тестування гіпотези про рівність середнього значення ряду помилок нулю використано двосторонній t-тест. Отриманий результат (t-statistic = 7.35, two-sided p-value = 2e-10), свідчить, що ми не можемо відхилити нульову

гіпотезу про рівність середнього значення нулю. Наступним кроком застосовано критерії Льюнга-Бокса і Бокса-Пірса, що використовуються для тестування відмінності від нуля групи авторегресивних коефіцієнтів часового ряду. Перші п'ять значень тестової статистики і p-значення за обома критеріями представлено в таблиці. Зі збільшенням номеру лагу значення lb_stat і bp_stat зростають, а lb_pvalue і bp_pvalue спадають. Такі результати свідчать про те, що гіпотеза про випадковість залишків не відкидається, і цей процес являє «білий шум». Отже, вся інформація у часовому ряді була використана моделлю для прогнозування, а отримані помилки - це випадкові коливання, які не можуть бути змодельовані.

Фрагмент результату тестування гіпотез Льюнга-Бокса і Бокса-Пірса для часового ряду помилок прогнозування (перші п'ять лагів)

	Критерій Льюнга-Бокса		Критерій Бокса-Пірса	
Лаг	lb_stat	lb_pvalue	bp_stat	bp_pvalue
1	35.989546	1.983790e-09	34.605333	4.037986e-09
2	62.838709	2.263330e-14	60.077615	9.001432e-14
3	76.685452	1.576953e-16	73.036747	9.545780e-16
4	81.935594	6.773731e-17	77.883032	4.890340e-16
5	85.528583	5.832197e-17	81.153573	4.813480e-16

Для розуміння того, який відсоток спостережень описує дана модель, розраховано коефіцієнт детермінації $R^2 = 0{,}96$. Коефіцієнт детермінації близький до одиниці, і це підтверджує судження про вірність запропонованої моделі.

Основними оціночними характеристиками якості прогнозної моделі є наведені нижче показники:

1) Середня абсолютна помилка прогнозу (mean absolute error, MAE)

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\varepsilon_t| = 1365{,}22.$$

2) Середньоквадратична помилка прогнозу (root mean squared error, RMSE)

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}\varepsilon_t^2}{n-1}} = 306760{,}25.$$

3) Середній абсолютний відсоток помилок або середня відносна помилка прогнозу (mean absolute percentage error, MAPE)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|\varepsilon_t|}{y_t} \times 100\% = 8{,}39\%.$$

Ці показники в подальшому можуть бути використані для порівняння моделей прогнозування і вибору найкращої з них.

Висновки і перспективи. Отже, розроблено прогнозну математичну модель нечіткого часового ряду, в основу якої покладено базові принципи теорії нечітких множин другого типу та обчислення зі словами. На базі розробленої моделі запропоновано нечіткий алгоритм прогнозування категоріальних часових рядів. Статистична перевірка ряду помилок прогнозу на відповідність білому шуму з нульовим середнім значенням підтвердила високу якість прогнозу. У подальшому запропонована модель може бути розвинута для розв'язання практичних задач обчислень зі словами при прийнятті рішень.

Список використаних джерел

1. Konstantinos Fokianos. Benjamin Kedem. Regression Theory for Categorical Time Series. Statist. Sci. 18 (3) 357 - 376, August 2003. https://doi.org/10.1214/ss/1076102425

2. Monnie McGee, Ian Harris. Coping with Nonstationarity in Categorical Time Series", Journal of Probability and Statistics, vol. 2012, Article ID 417393, 9 pages, 2012. https://doi.org/10.1155/2012/417393

3. Song, Q., Chissom, B.S. Fuzzy time series and its models. Fuzzy Sets and Systems 54, 269–277 (1993) https://doi.org/10.1016/0165-0114(93)90372-O

4. Mendel, J.M., John, R.I.B. Type-2 Fuzzy Sets Made Simple. IEEE Transactions on Fuzzy Systems, 10 (2), 117-127 (2002).

5. Petrenko, T., Tymchuk, O. Package library and toolbox for discrete interval type-2 fuzzy logic systems. In: the 18th International Conference on Soft Computing, pp. 233-238, MENDEL, Brno, Czech Republic (2012).

6. Mendel, J.M., Wu, D. Perceptual Computing: Aiding People in Making Subjective Judgments. 1st edn. Wiley-IEEE Press (2010).

7. Chen, S.M. Forecasting enrollments based on fuzzy time series. Fuzzy Sets Syst. 81 (3), 311–319 (1996).

8. Wu, D. and Mendel, J. M. Enhanced Karnik-Mendel Algorithms for Interval Type-2 Fuzzy Sets and Systems, Fuzzy Information Processing Society, 2007. NAFIPS '07. Annual Meeting of the North American, 2007, pp. 184 – 189.

References

1. Konstantinos Fokianos. Benjamin Kedem (2003). Regression Theory for Categorical Time Series. Statist. Sci., 18 (3), 357 - 376, https://doi.org/10.1214/ss/1076102425

2. Monnie McGee, Ian Harris (2012). Coping with Nonstationarity in Categorical Time Series", Journal of Probability and Statistics, Article ID 417393, 9. https://doi.org/10.1155/2012/417393

3. Song, Q., Chissom, B. S. (1993). Fuzzy time series and its models. Fuzzy Sets and Systems, 54, 269–277 https://doi.org/10.1016/0165-0114(93)90372-O

4. Mendel, J. M., John, R. I. B. (2002). Type-2 Fuzzy Sets Made Simple. IEEE Transactions on Fuzzy Systems, 10 (2), 117-127.

5. Petrenko, T., Tymchuk, O. (2012). Package library and toolbox for discrete interval type-2 fuzzy logic systems. In: the 18th International Conference on Soft Computing, MENDEL, Brno, Czech Republic, 233-238.

6. Mendel, J. M., Wu, D. (2010). Perceptual Computing: Aiding People in Making Subjective Judgments. 1st edn. Wiley-IEEE Press.

7. Chen, S. M. (1996). Forecasting enrollments based on fuzzy time series. Fuzzy Sets Syst., 81 (3), 311–319.

8. Wu, D., Mendel, J. M., Enhanced Karnik-Mendel (2007). Algorithms for Interval Type-2 Fuzzy Sets and Systems, Fuzzy Information Processing Society, NAFIPS '07. Annual Meeting of the North American, 184 – 189.

APPLICATION OF THE THEORY OF TYPE-2 FUZZY SETS TO THE FORECASTING OF CATEGORICAL TIME SERIES: A MATHEMATICAL MODEL AND ALGORITHM

O. Tymchuk, A. Pylypenko, O. Ivanchenko

Abstract. *In this paper is considered the problem of forecasting categorical time series. Such series have a wide practical application in almost all spheres where judgments and expert evaluations are used. The analysis of modern research shows that the problem of taking into account the linguistic uncertainty remains insufficiently studied.*

The purpose of this research is to design a time series model based on type-2 fuzzy sets theory that will allow to perform computing with words.

The type-2 fuzzy time series model gives the result in the form of a granular term, which is described by a word and a discrete interval type-2 fuzzy set.

Based on the proposed model, the fuzzy algorithm for forecasting time series has been developed, which consists of five steps: word model definition; fuzzification of time series values; fuzzy relations definition; fuzzy forecasting; defuzzification.

The high quality of the proposed forecast model is confirmed by three evaluation characteristics: Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE).

Key words: ***time series, categorical data, type-2 fuzzy set, uncertainty, computing with words, fuzzy prediction***