МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ
ТА СПОРТУ УКРАЇНИ
Дніпропетровський національний університет
ім. Олеся Гончара

АКТУАЛЬНІ ПРОБЛЕМИ АВТОМАТИЗАЦІЇ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

ТОМ 16

Збірник наукових праць

Дніпропетровськ
Видавництво
Дніпропетровського національного університету
2012

УДК 519.689:519.254
ББК 32.973.202
А 43

Описані задачі автоматизованої обробки даних, інформаційна технологія сучасних обчислювальних процедур більш адекватного відображення експериментальних даних у системах моніторингу. Досліджені обчислювальні процедури сплайн-перетворень, які застосовані у системах ГІС та прийняття рішень, прикладні задачі інформаційних технологій.

Для науковців, аспірантів та фахівців у галузі інформатики.

А43 **Актуальні проблеми автоматизації та інформаційних технологій :** зб. наук. пр. / наук. ред. О.Г. Байбуз – Д. : Вид-во Дніпропетр. нац. ун-ту, 2012. – Т.16. – 176 с.
ISBN 978-966-551-362-9

Описаны задачи автоматизированной обработки данных, информационная технология современных вычислительных процедур более адекватного отображения экспериментальных данных в системах мониторинга. Исследованы вычислительные процедуры сплайн-преобразований, примененных в системах ГИС и принятия решений, прикладные задачи информационных технологий.

Для научных сотрудников, аспирантов и специалистов в области информатики.

УДК 519.689:519.254
ББК 32.973.202

Рекомендовано вченою радою
Дніпропетровського національного університету ім. Олеся Гончара

УДК 519.254:519.600:519.674

Ю.М. Архангельська, Ю.В. Самарець

Дніпропетровський національний університет ім. Олеся Гончара

АПРОКСИМАЦІЯ НЕПЕРЕРВНИХ ВИПАДКОВИХ ПРОЦЕСІВ МАРКОВСЬКИМИ У СИСТЕМІ ГІДРОХІМІЧНОГО МОНІТОРИНГУ

На підставі методів теорії марковських процесів, розроблено математичну модель поведінки концентрації гідрохімічних показників питної води.

Ключові слова: *гідрохімічний моніторинг, математична модель, питна вода.*

На основе методов теории марковских процессов разработана математическая модель поведения концентрации гидрохимических показателей питьевой воды.

Ключевые слова: *гидрохимический моніторинг, математическая модель, питьева вода.*

On the basis of method of theory of markov processes, the mathematical model of conduct of concentration of hydrochemical indexes of drinking-water is developed.

Key words: *territorial hydrochemical monitoring, mathematical model, drinking-water.*

Постановка проблеми у загальному вигляді та її зв'язок з важливими науковими дослідженнями. В Україні, зокрема на території Придніпров'я, в умовах розвиненої промислової та сільськогосподарської діяльності постає завдання своєчасного та якісного аналізу стану питної води (ПВ) техногенно навантажених регіонів. Одним з основних завдань під час проведення аналізу даних гідрохімічного моніторингу питної води (ГХМ ПВ) та оцінки факторів навколишнього середовища для запобігання їх шкідливому впливу на стан здоров'я населення є аналіз зміни гідрохімічних показників (ГХП) у певних межах**.** Вирішення поставленого завдання бачиться у використанні методів математичного моделювання.

Аналіз останніх досліджень та публікацій, в яких розпочато вирішення даної проблеми. У науковій літературі багато

теоретичних обґрунтувань і практичних підходів до побудові аналітичних моделей екосистем. Найбільш адекватним математичним апаратом побудови і аналізу аналітичних моделей є теорія диференціальних рівнянь [1] та теорія біфуркацій [2; 3]. Також особливе місце займають стохастичні моделі та імітаційні моделі екосистем [4 – 13]. За своєю природою концентрація ГХП ПВ є стохастичним неперервним процесом. З іншого боку, концентрації речовин змінюються у певних межах, перехід між якими відбувається у випадкові моменти часу. В зв'язку з вказаними властивостями пропонується аналізувати поведінку концентрації показників ПВ за моделями з класу марковських. Застосування методів теорії марковських процесів знайшло себе у розв'язанні задач екологічного моніторингу, однак відсоток робіт, де марковські моделі використовуються в задачах ГХМ, дуже малий. Основу таких моделей було подано у [6].

У зв'язку з цим постає задача в необхідності розробки математичної моделі поведінки концентрації ГХП ПВ та її аналіз.

Постановка задачі. Задача оцінки реакції гідрохімічного середовища на техногенне навантаження з ієрархічним діленням його компонентів, ймовірності знаходження гідрохімічних показників питної води (ГХП ПВ) у заданих станах та ризику знаходження стану ПВ поза зоною норми з виділенням ГХП ПВ та аналізом їх впливу на стан ПВ може бути вирішена за допомогою теорії марковських процесів, розбиттям множини $g(t)$ реалізацій випадкового неперервного процесу $g(t)=\{g_t, t\in T\}=(g_1, g_2,...g_m)$ на n' підмножин: $g\in k \Leftrightarrow g\in\left(g_k^*, g_{k+1}^*\right]$, де g_k^* – елементи розбиття $\Delta_h = \left\{g_k^*, k=1,2,...n'+1\right\}$; t – дата відбору проб води; g_t – значення концентрації гідрохімічної речовини в t-й момент часу (рис.1).

Основний матеріал. При відстеженні накопичення хімічних елементів у ПВ було виділено три стани:

$$\begin{cases} 0:\ g(t) < ФЗК, \\ 1:\ ФЗК \le g(t) \le ГДК, \\ 2:\ g(t) > ГДК. \end{cases}$$

де ФЗК – фонове значення концентрації гідрохімічної речовини, ГДК – гранично-допустима концентрація гідрохімічної речовини.

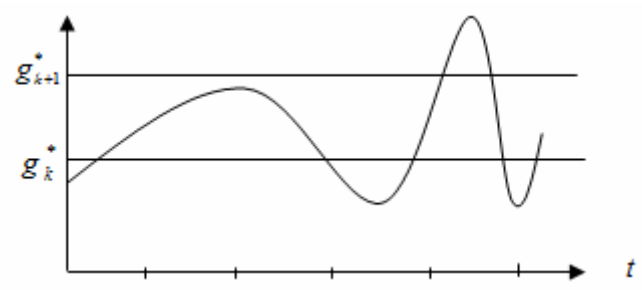


Рис.1. Реалізація неперервного випадкового процесу у вигляді дискретно-неперервного процесу

Спеціалістами санітарно-епідеміологічної станції (СЕС) три стани було охарактеризовано як: «0» – фоновий стан води; «1» – вода для побутово-питних нужд; «2» – вода для побутових нужд.

Нехай $P_k(t)$ – ймовірність знаходження процесу в стані k ,тоді, виходячи з природи неперервного випадкового гідрохімічного процесу $g(t)=\{g_t, t\in T\}=(g_1, g_2, ...g_M)$, побудовано математичну модель процесу поведінки концентрації хімічних показників, граф якої надано на рис.2.

Відповідно заданих станів проведена оцінка інтенсивностей потоку вимог протягом часу Δt для знаходження вірогідних кусково-сталих інтенсивностей згідно [6]. Оцінка інтенсивностей пов'язана з процедурою перевірки гіпотези H_o:$\lambda_i=\lambda_{i+1}$ та обчисленням статистичної характеристики

$$t=\frac{\lambda_{i+1}-\lambda_i}{\sqrt{(n_i-1)\lambda_{i+1}^2+(n_{i+1}-1)\lambda_i^2}}\sqrt{\frac{n_i n_{i+1}(n_i+n_{i+1}-2)}{n_i+n_{i+1}}},$$

яка має t – розподіл з кількістю степенів вільності $\nu=n_i+n_{i+1}-2$.

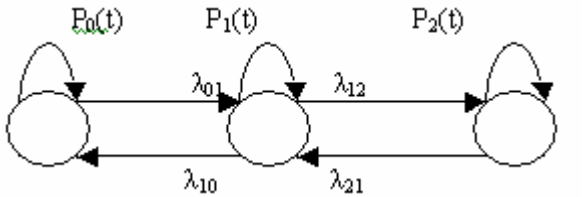


Рис.2. Граф переходів між станами

Якщо $|t|\le t_{\alpha/2,\nu}$, то гіпотеза H_o є вірною та здійснюється перерахування значення λ_i.

$$\lambda_i=\frac{n_i+n_{i+1}}{(N-N_i)(\Delta t_i+\Delta t_{i+1})}, \text{ де } N_i=\sum_{j=1}^{i+1} n_j .$$

За результатом оцінки інтенсивностей потоків вимог відповідно заданих зон виділено $\lambda_{ij}=const, i,j=\overline{0,2}$.

Для визначення ймовірностей перебування в різних зонах при $\lambda_{ij}=const; i,j=\overline{0,2}$ розв'язано систему диференційних рівнянь вигляду:

$$\begin{cases}\dfrac{dP_0(t)}{dt}=-\lambda_{01}P_0(t)+\lambda_{10}P_1(t),\\ \dfrac{dP_1(t)}{dt}=-(\lambda_{10}+\lambda_{12})P_1(t)+\lambda_{01}P_0(t),\\ \dfrac{dP_2(t)}{dt}=\lambda_{12}P_1(t)-\lambda_{21}P_2(t),\\ P_0(t)+P_1(t)+P_2(t)=1,\end{cases}$$

$$P_0(t_0)=1, P_1(t_0)=0, P_2(t_0)=0$$

та отримано розв'язок для динамічного та стаціонарного режимів.

Розв'язком стаціонарного режиму є:

$$P_0=\frac{\lambda_{21}\lambda_{10}}{\lambda_{01}\lambda_{21}+\lambda_{01}\lambda_{12}+\lambda_{10}\lambda_{21}},\ P_1=\frac{\lambda_{01}}{\lambda_{10}}P_0,\ P_2=\frac{\lambda_{01}\lambda_{12}}{\lambda_{21}\lambda_{10}}P_0 .$$

Для знаходження розв'язку в динамічному режимі застосовувалось перетворення Лапласа. У цьому випадку розв'язок має вигляд:

$$\begin{cases}P_0(t)=\dfrac{\lambda_{21}\lambda_{10}}{\lambda_{01}\lambda_{21}+\lambda_{01}\lambda_{12}+\lambda_{10}\lambda_{21}}+\dfrac{(\lambda_{12}+\lambda_{10}+S_1)(\lambda_{21}+S_1)-\lambda_{12}\lambda_{21}}{\Delta^1(S_1)}\exp^{(S_1t)}+\\ +\dfrac{(\lambda_{12}+\lambda_{10}+S_2)(\lambda_{21}+S_2)-\lambda_{12}\lambda_{21}}{\Delta^1(S_2)}\exp^{(S_2t)},\\ P_1(t)=\dfrac{\lambda_{01}\lambda_{21}}{\lambda_{01}\lambda_{21}+\lambda_{01}\lambda_{12}+\lambda_{10}\lambda_{21}}+\dfrac{\lambda_{01}(\lambda_{21}+S_1)}{\Delta^1(S_1)}\exp^{(S_1t)}+\dfrac{\lambda_{01}(\lambda_{21}+S_2)}{\Delta^1(S_2)}\exp^{(S_2t)},\\ P_2(t)=\dfrac{\lambda_{01}\lambda_{12}}{\lambda_{01}\lambda_{21}+\lambda_{01}\lambda_{12}+\lambda_{10}\lambda_{21}}+\dfrac{\lambda_{01}\lambda_{12}}{\Delta^1(S_1)}\exp^{(S_1t)}+\dfrac{\lambda_{01}\lambda_{12}}{\Delta^1(S_2)}\exp^{(S_2t)},\end{cases}$$

де $\Delta = S^3+S^2(\lambda_{12}+\lambda_{01}+\lambda_{21}+\lambda_{10})+S(\lambda_{01}\lambda_{21}+\lambda_{01}\lambda_{12}+\lambda_{10}\lambda_{21})$,

$\Delta^1 = 3S^2 + 2S(\lambda_{12} + \lambda_{01} + \lambda_{21} + \lambda_{10}) + (\lambda_{01}\lambda_{21} + \lambda_{01}\lambda_{12} + \lambda_{10}\lambda_{21})$,

S_1, S_2 – корені рівняння Δ .

Наступним кроком аналізу марківської системи є визначення моменту, в який процес переходить з динамічного режиму в стаціонарний. Для цього, задаючись похибкою, знаходять границі стаціонарності щодо зон:

$$P_0^{н,в} = P_0 \mp \alpha/2\, P_0,$$
$$P_1^{н,в} = P_1 \mp \alpha/2\, P_1,$$
$$P_2^{н,в} = P_2 \mp \alpha/2\, P_2.$$

Для знаходження моментів t_0, t_1, t_2 з яких починається стаціонарний режим стосовно $P_0(t), P_1(t), P_2(t)$, розв'язано рівняння вигляду:

$$P_0^{в} = \frac{\lambda_{21}\lambda_{10}}{\lambda_{01}\lambda_{21} + \lambda_{01}\lambda_{12} + \lambda_{10}\lambda_{21}} + \frac{(\lambda_{12} + \lambda_{10} + S_1)(\lambda_{21} + S_1) - \lambda_{12}\lambda_{21}}{\Delta^1(S_1)} \exp^{(S_1 t_0)} +$$
$$+ \frac{(\lambda_{12} + \lambda_{10} + S_2)(\lambda_{21} + S_2) - \lambda_{12}\lambda_{21}}{\Delta^1(S_2)} \exp^{(S_2 t_0)},$$

$$P_1^{н} = \frac{\lambda_{01}\lambda_{21}}{\lambda_{01}\lambda_{21} + \lambda_{01}\lambda_{12} + \lambda_{10}\lambda_{21}} + \frac{\lambda_{01}(\lambda_{21} + S_1)}{\Delta^1(S_1)} \exp^{(S_1 t_0)} + \frac{\lambda_{01}(\lambda_{21} + S_2)}{\Delta^1(S_2)} \exp^{(S_2 t_0)},$$

$$P_2^{н} = \frac{\lambda_{01}\lambda_{12}}{\lambda_{01}\lambda_{21} + \lambda_{01}\lambda_{12} + \lambda_{10}\lambda_{21}} + \frac{\lambda_{01}\lambda_{12}}{\Delta^1(S_1)} \exp^{(S_1 t_0)} + + \frac{\lambda_{01}\lambda_{12}}{\Delta^1(S_2)} \exp^{(S_2 t_0)}.$$

Результати досліджень. Апробація розробленої моделі поведінки концентрації ГХП ПВ виконана за даними гідрохімічного моніторингу питної води Дніпропетровського району Дніпропетровської області за 14 точками водозабору на період з 2003 – 2010 рр:

- лівобережна частина Дніпропетровського району: № 1, 2, 3 с. Чумаки; смт. Ювілейне; № 1, 2 с. Партизанське; с. В. Маївка; с. Степне; с. Зоря; м. Підгороднє, вул. Робоча (№1); м. Підгороднє, вул. Енергетиків (№2); м. Підгороднє, вул. Геологів (№3);
- правобережна частина Дніпропетровського району: с. Новомиколаївка, с. Сурсько-Литовське.

Оцінка ймовірності знаходження ГХП ПВ у заданих станах та ризику знаходження стану ПВ поза зоною норми проведено на прикладі показників ПВ, що подано у табл. 1.

Таблиця 1

Нормативні значення основних гідрохімічних показників

Назва показника питної води	Одиниця виміру	ГДК	Фонова концентрація
Сухий залишок	мл/дм^3	1000	650
NH_4^+	мл/дм^3	2	0,3
NO_2^-	мл/дм^3	3,3	0,8
NO_3^-	мл/дм^3	45	2
Fe	мл/дм^3	0,3	0,1
Окисність	мл/дм^3	4	2,6

За рахунок великої кількості обчислень результати досліджень представлено на прикладі с. Чумаки водозабір № 1 на прикладі ГХП ПВ Fe період збору інформації з червня 2003 року по березень 2009 року.

Була отримана оцінка сталих функцій інтенсивностей для ГХП ПВ Fe та інших ГХП ПВ за водозабором №1 с. Чумаки (табл. 2).

Таблиця 2

Інтенсивності переходів між станами

Гідрохімічний показник ПВ	Інтенсивності (1/міс.)			
	λ_{01}	λ_{12}	λ_{21}	λ_{10}
Fe	0,2899	0,2041	0,5813	0,1454
Cu *	–	–	–	–
NH_4^+	0,1928	0,2384	0,4798	0,1790
NO_3^-	0,0159	0	0	0,0154
NO_2^- *	–	–	–	–
Загальна жорсткість	0,1819	0,1264	0,1416	0,3095

Закінчення табл. 2

Окисність	0,2312	0,1214	0,1166	0,2244
*Сухий залишок**	–	–	–	–

«*» – гідрохімічний показник знаходиться в одному стані.

Результатом динамічного та стаціонарних режимів є функції знаходження концентрації ГХП ПВ у різних станах. Остання функція $P_2(t)$ уявляє собою функцію ризику знаходження ГХП ПВ поза зоною норми:

$$\begin{cases} P_0(t) = 0{,}3534 + 0{,}5728\exp^{(-0{,}4155t)} + 0{,}0739\exp^{(-0{,}6816t)}, \\ P_1(t) = 0{,}5018 - 0{,}29965\exp^{(-0{,}4155t)} - 0{,}20192\exp^{(-0{,}6816t)}, \\ P_2(t) = 0{,}1448 - 0{,}27314\exp^{(-0{,}4155t)} + 0{,}13792\exp^{(-0{,}6816t)}. \end{cases}$$

Для стаціонарного режиму: $P_0 = 0{,}3534$, $P_1 = 0{,}5018$, $P_2 = 0{,}1448$.

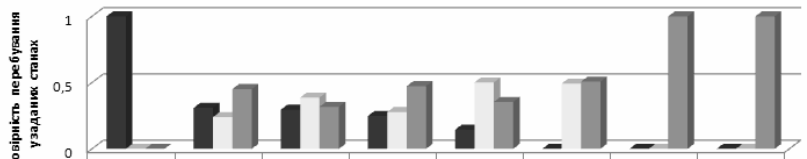
Аналогічним чином знайдено функції знаходження концентрації ГХП ПВ у різних станах для інших ГХП ПВ для динамічного та стаціонарного режимів. Наступним кроком аналізу марковської системи є визначення моменту, в який процес переходить з динамічного режиму до стаціонарного. Так, задавшись похибкою $\alpha = 0.05$, знайдено границі стаціонарності щодо станів для кожного з ГХП ПВ. Для знаходження моментів t_0, t_1, t_2, з яких починається стаціонарний режим для Fe стосовно $P_0(t)$, $P_1(t)$, $P_2(t)$, розв'язано рівняння вигляду:

$$\begin{cases} P^{в}{}_0(t) = 0{,}3534 + 0{,}5728\exp^{(-0{,}4155t)} + 0{,}0739\exp^{(-0{,}6816t)}, \\ P^{н}{}_1(t) = 0{,}5018 - 0{,}29965\exp^{(-0{,}4155t)} - 0{,}20192\exp^{(-0{,}6816t)}, \\ P^{н}{}_2(t) = 0{,}1448 - 0{,}27314\exp^{(-0{,}4155t)} + 0{,}13792\exp^{(-0{,}6816t)}. \end{cases}$$

Для хімічного елемента Fe, t_0, t_1, t_2, відповідно дорівнюють $t_0 = 8{,}8$ місяців, $t_1 = 6{,}4$ місяців, $t_2 = 9{,}14$ місяців. У результаті проведеної апроксимації можна зробити висновок, що час переходу системи до стаціонарного режиму становить 9,14 місяців. Аналогічним чином знайдено час переходу до стаціонарного режиму за іншими ТГХМ ПВ Дніпропетровського району. За даними ГХМ ПВ отримано, що система переходить у стаціонарний режим, час якого

склав у середньому біля 6 років та при похибці $\alpha = 0.05$ відбувся на початку 2008 року.

На рис. 3 представлено ймовірність знаходження ГХП ПВ у «0»-му, «1»-му станах та функція ризику знаходження ГХП ПВ поза зоною норми («2»-й стан) у стаціонарному режимі для водозабору № 1 с. Чумаки.



	Сухий залишок	Окисність	NH4+	Загальна жорсткість	Fe	NO3-	Cu	NO2-
P2	1	0,3092	0,2954	0,2484	0,1445	0	0	0
P1	0	0,2396	0,3874	0,2782	0,5016	0,4927	0	0
P0	0	0,4512	0,3172	0,4734	0,3539	0,5073	1	1

Рис. 3. Ймовірності знаходження ГХП ПВ у заданих станах водозабору №1 с. Чумаки

Висновки. Отримані ймовірності та функція ризику як для динамічного, так і для стаціонарного режимів за математичної моделлю поведінки концентрації ГХП ПВ дозволили виявити речовини, що мають найбільший вплив на стан ПВ та зробити висновок щодо рекомендацій по зменшенню техногенного навантаження за вказаним показниками. Так, речовина сухий залишок на період дослідження знаходиться завжди поза зоною норми ПВ та у першу чергу потребує контролю. Речовини, які під впливом техногенного навантаження перейшли з фонового стану і можуть знаходитись у стані перевищення ГДК за водозабором № 1 с. Чумаки є: окисність (31 %), NH_4^+ (30 %), загальна жорсткість (25 %), Fe (14 %). Хоча NO_3^- і знаходиться у стані норми відносно ГДК речовини у ПВ, його поведінку слід контролювати, тому що вказаний ГХП ПВ схильний до впливу техногенного навантаження. Речовини NO_2^- та Cu за період дослідження залишились у фоновому стані.

У подальшому ймовірності знаходження ГХП ПВ у заданих станах, що отримані за математичною моделлю будуть використані у технології оцінки техногенного впливу на стан підземних питних вод.

Бібліографічні посилання

1. **Эрроусмит Д.** Обыкновенные дифференциальные уравнения. Качественная теория с приложениями / Д. Эрроусмит, К. Плейс – М., 1986. – 243 с.

2. **Свирежев Ю. М.** Вито Вольтерра и современная математическая экология / Ю. М. Свирежев, В. Вольтерра // Математическая теория борьбы за существование. – М., 1976. – С. 245–286.

3. **Свирежев Ю. М.** Нелинейные волны, диссипативные структуры и катастрофы в экологии / Ю. М. Свирежев – М., 1987. – 368 с.

4. **Винберг Г.Г.** Математическая модель водной экосистемы. / Г.Г. Винберг, С.И. Анисимов // Фотосинтезирующие системы высокой продуктивности. – М., 1966. С. 213 – 223.

5. **Абросов Н.С.** Экологические и генетические закономерности сосуществования и коэволюции видов. / Н.С. Абросов, А.Г. Боголюбов – Новосибирск, 1988. – 333 с.

6. **Байбуз О.Г.** Оперативний аналіз в системах моніторингу. / О.Г. Байбуз, О.П. Приставка, С.В. Земляна – Д., 2005. – 92 с.

7. **Базыкин А.Д.** Математическая биофизика взаимодействующих популяций / А.Д. Базыкин – М., 1985. – 180 с.

8. **Крестин С.В.** Об одном механизме "цветения воды" в водохранилище равнинного типа / С.В. Крестин, Г.С. Розенберг // Биофизика. – 1996. – Т. 41. Вып. 3. – С. 650–654.

9. **Крестин С.В.** Двухмерная модель "цветения воды" в водохранилище равнинного типа / С.В. Крестин, Г.С. Розенберг – Изв. СамНЦ РАН. – 2002. – Т. 4, № 2. – С. 276–279.

10. **Розенберг Г.С.** Модели в фитоценологии. / Г.С. Розенберг – М., 1984. – 256 с.

11. **Федоров В.Д.** Экология / В.Д. Федоров, Т.Г. Гильманов – М., 1980. – 464 с.

12. **Флейшман Б.С.** Системные методы в экологии / Б.С. Флейшман // Статистические методы анализа почв, растительности и их связи. – Уфа, 1978. – С. 7–10.

13. **Горстко А.Б.** Имитационная система “Азовское море” – инструмент анализа и прогнозирования / А.Б. Горстко, Л.В. Эпштейн // Математическое моделирование водных экологических систем. – Иркутск, 1978. С. 47–58.

Надійшла до редколегії 20.09.2012

УДК 519.254

О. Г. Байбуз, М. Г. Сидорова

Дніпропетровський національний університет ім. Олеся Гончара

ГРУПУВАННЯ ПУНКТІВ ГІДРОХІМІЧНОГО СПОСТЕРЕЖЕННЯ ЗА СХОЖІСТЮ ХІМІЧНОГО СКЛАДУ ВОДИ З УРАХУВАННЯМ ЧАСОВИХ ЗМІН

Запропоновано інформаційну технологію визначення груп схожих об’єктів за сукупністю досліджуваних ознак, враховуючи їх зміни у часі. Розроблено обчислювальні схеми та програмне забезпечення для аналізу даних гідрохімічного моніторингу.

Ключові слова: *кластерний аналіз, гідрохімічний моніторинг, часові ряди, інформаційна технологія.*

Предложена информационная технология определения групп похожих объектов по совокупности исследуемых признаков, учитывая их временные изменения. Разработаны вычислительные схемы и программное обеспечение для анализа данных гидрохимического мониторинга.

Ключевые слова: *кластерный анализ, гидрохимический мониторинг, временные ряды, информационная технология*.

Information technology for determining groups of similar objects on the set of the studied features taking into account their changes over time has been proposed. Computational schemes and software for data analysis of hydrochemical monitoring has been developed.

Key words: *cluster analysis, hydrochemical monitoring, time series, information technology.*

Вступ. За процесом видобутку та збагачення залізних руд, унаслідок значного техногенного навантаження змінюються гідрохімічні процеси водних об’єктів. Тому актуальним є проведення гідрохімічного моніторингу з метою збереження, поліпшення і стабілізації якості поверхневих вод для забезпечення оптимальних умов функціонування екосистем та підвищення ефективності природно-господарського комплексу. Особливо складним є гідрохімічний моніторинг водних об’єктів у районах з підвищеним техногенним навантаженням.

Одне з основних місць у системі гідрохімічного моніторингу

займає обґрунтування пунктів спостережень, об'ємів і періодичності гідрохімічних випробувань. Це може бути вирішено на підставі гідрохімічного районування територій. Використовуючи районування, можна до певної міри уніфікувати водоохоронні заходи в межах виділених груп та районів. Визначивши пріоритетні водоохоронні заходи для одного району, планувати і впроваджувати їх для всієї виділеної групи [1].

Постановка задачі. Метою роботи є визначення групи пунктів спостереження, що характеризуються схожим хімічним складом води р. Інгулець поблизу ВАТ «Центрального гірничо-збагачувального комбінату» за досліджуваними компонентами для правильного планування природоохоронних заходів та керування якістю вод річки. Проби води відбиралися у 5 пунктах спостереження: селище Тернівка, створи балок: Мала Лозоватка, Велика Лозоватка, Завертана, північна частина Карачунівського водосховища. Аналіз проводився за вмістом головних іонів у воді річки Інгулець: HCO_3^-, Cl^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, Na^+ та мінералізацією протягом наступних років: 1993–1995, 1997, 2001, 2003, 2005–2007. Представимо досліджувані дані у вигляді наступної структури $X = \{x_{ijt}\}, i = \overline{1,N}, j = \overline{1,p}, t = \overline{1,T}$. Тобто маємо N об'єктів, які характеризуються p ознаками, значення яких змінюються протягом T моментів часу, x_{ijt} – значення j-го показника i-го об'єкта в момент часу t. Необхідно виділити групи об'єктів, схожих між собою за усіма досліджуваними ознаками у заданому періоді спостережень.

Таким чином, метою роботи є розробка технології, яка дозволить виділити групи об'єктів, схожих між собою за усіма досліджуваними ознаками, що змінюються у часі та врахувати часові зміни досліджуваних показників.

Основний матеріал. Методи кластерного аналізу, що дозволяють виявити групи схожих між собою об'єктів не можна явним виглядом застосувати до вирішення поставленої задачі, оскільки в якості вхідної інформації вони використовують матрицю «об'єкти-ознаки» $X = \{x_{ij}\}, i = \overline{1,N}, j = \overline{1,p}$, де x_{ij} – конкретне значення, а не часовий ряд, як у нашому випадку. Тому в даній роботі пропонується технологія виділення груп об'єктів, схожих між собою за набором ознак, які змінюються у часі, що ґрунтується на методах колективної кластеризації та складається з наступних етапів: визначення груп об'єктів для кожного моменту часу t, побудова узгодженої матриці подібності, отримання узагальнюючого розв'язку задачі.

Визначення груп об'єктів. Представимо вихідні дані у вигляді групи двовимірних матриць $X^{(t)} = \{x_{ij}^{(t)}\}, i = \overline{1,N}, j = \overline{1,p}, t = \overline{1,T}$. Окремо до кожної з них застосовуємо відомі методи кластерного аналізу [2–6] для визначення структури даних у певний момент часу. Тобто, отримаємо набір угрупувань $G_t = \{g_1^{(t)}, g_2^{(t)}, \ldots, g_k^{(t)}\} : t = \overline{1,T}$, де g_i^t – список об'єктів, що потрапили до i-го кластера в t-му угрупуванні, k – кількість кластерів. Важливо проводити оцінку якості та вибір найбільш вірогідного результату кластеризації, щоб кожне з угрупувань G_t найкраще відповідало структурі досліджуваних даних.

Побудова узгодженої матриці подібності. На основі отриманого набору угрупувань будуємо узгоджену матрицю подібності об'єктів $S = \{s_{ij}; i, j = \overline{1,N}\}$, де N – кількість об'єктів, s_{ij} – частота віднесення i-го та j-го об'єктів до одного кластера.

Процес побудови можна представити у вигляді наступного алгоритму:

1. Створюємо матрицю $S = \{s_{ij} = 0; i, j = \overline{1,N}\}$.

2. Розглядаємо по черзі угрупування $G_t : t = \overline{1,T}$. Якщо i-й та j-й об'єкти відносяться до одного кластера в момент t, то s_{ij} збільшуємо на одиницю.

3. Зводимо елементи матриці подібності до одиничної шкали $S = \{s_{ij} = \frac{s_{ij}}{T}; i, j = \overline{1,N}\}$. Чим ближче значення s_{ij} до одиниці, тим вища ймовірність віднесення об'єктів і та j до одного кластеру.

Отримання узагальнюючого колективного розв'язку. Для отримання підсумкового узагальнюючого розв'язку необхідно виділити групи об'єктів на основі обчисленої матриці подібності. Для цього застосовуємо графовий алгоритм найкоротшого незамкненого шляху. В якості матриці близькості використовуємо матрицю $S' = \{s'_{ij} = 1 - s_{ij}; i, j = \overline{1,N}\}$. Тобто чим більше подібні об'єкти i та j за матрицею S, тим менша відстань між ними у матриці S'.

Розглянемо результати запропонованої технології застосованої до даних гідрохімічного моніторингу, що проводиться Криворізькою геолого-гідрогеологічною партією по р. Інгулець (Кривбас).

За постановкою завдання маємо дев'ять багатовимірних вибірок «пункти спостереження – досліджувані компоненти», кожна з яких відповідає певній даті відбору проб води з річки. За допомогою

методів кластерного аналізу (ієрархічних, К-середніх, графового, Forel) та технології оцінки якості [7–9] та підвищення стійкості результатів [6; 10] отримуємо угрупування схожих між собою об'єктів для кожного окремого моменту часу. Для зведення даних до єдиного масштабу попередньо проводимо стандартизацію.

На рис. 1, 2 представлені розбиття, що відповідають даним 2001 та 2007 років. За станом води відібраних проб пункти спостереження розподілилися на 2 групи наступним чином: у 2001 році до першого класу увійшли с. Тернівка, Мала Лозоватка, до другого – Велика Лозоватка, Завертана, північна частина Карачунівського водосховища; у 2007 році в перший кластер виділено с.Тернівка, другий містить усі інші об'єкти дослідження.

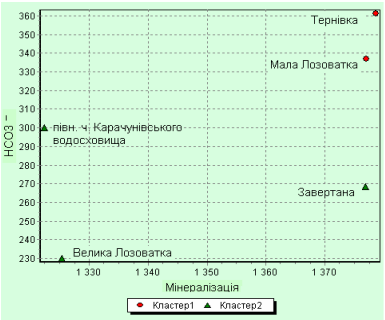


Рис. 1. Результати кластеризації за даними 2001 р.

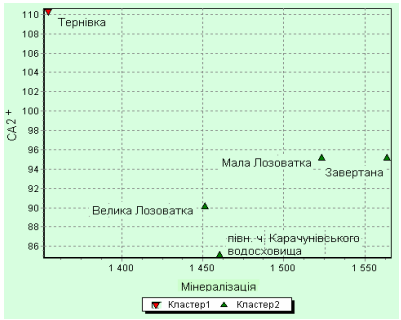


Рис. 2. Результати кластеризації за даними 2007 р.

Такий підхід визначає угрупування пунктів спостереження на певну дату, що дозволяє аналізувати зміни схожості об'єктів за часом.

Для визначення об'єктів схожих між собою на всьому часовому проміжку спостереження за всіма досліджуваними показниками одночасно з метою відображення загальної картини перебігу певних гідрохімічних процесів у воді річки застосовуємо запропоновану технологію часової кластеризації. За результатами аналізу було виділено дві групи об'єктів: перша складається з пункту спостереження у с. Тернівка, друга містить всі інші об'єкти дослідження.

Проведемо кластерний аналіз за даними представленими у вигляді матриці «пункти спостереження – значення фіксованого показника у часі», тобто визначимо пункти спостережень, які є схожими між собою на досліджуваному часовому проміжку за одним з показників. Отримані результати (при виборі будь-якого показника) співпадають з результатами продемонстрованими запропонованою технологією

часової кластеризації, що підтверджує її адекватність. На рис. 3 – 4 представлено діаграми розсіювання об'єктів за значеннями показників HCO_3^- (рис.3) та Ca^{2+} (рис.4) у часовому діапазоні.

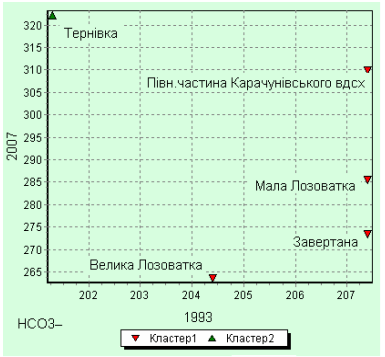


Рис. 3. Результати кластеризації за значеннями показника HCO_3^-

Рис. 4. Результати кластеризації за значеннями показника Ca^{2+}

Отримане поділення на кластери, за думкою фахівців з предметної області, відповідає дійсній гідрологічній та гідрохімічній ситуації на досліджуваній ділянці р. Інгулець. До Карачунівського водосховища подається вода каналом «Дніпро – Інгулець». Найбільший вплив цього каналу відзначається на верхній ділянці, що вивчається, а саме в районі селища Тернівка. Нижче за течією, вплив дніпровської води на формування хімічного складу води у річці Інгулець менший, більший вплив надає гірничо-збагачувальний комбінат (фільтраційні втрати з гідротехнічних споруд комбінату, пиління хвостосховища та інші).

Висновки. У даній роботі запропоновано технологію виділення груп об'єктів, схожих між собою за набором ознак, що змінюються за часом. Розроблено обчислювальні схеми та створено програмне забезпечення, це дозволило вирішити прикладну задачу визначення груп пунктів спостереження, що характеризуються схожим хімічним складом води у р. Інгулець за досліджуваними компонентами для правильного планування природоохоронних заходів та керування якістю вод річки. Запропонована технологія може бути застосована і в інших предметних галузях для виділення груп схожих об'єктів з урахуванням часових змін досліджуваних ознак.

Бібліографічні посилання

1. **Шерстюк Н. П.** Особливості гідрохімічних процесів у техногенних та природних водних об'єктах Кривбасу / Н. П. Шерстюк, В. К. Хільчевський. – Д., 2012. – 263 с.

2. **Мандель И. Д.** Кластерный анализ / И. Д. Мандель. – М., 1988. – 176 с.

3. **Айвазян С. А.** Классификация многомерных наблюдений / С. А. Айвазян, З. И. Бежаева, О. В. Староверов. – М., 1974. – 240 с.

4. **Загоруйко Н. Г.** Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. – Новосибирск, 1999. – 270 с.

5. **Jain A. K.** Data Clustering: A Review / A. K. Jain, M. N. Murty, P. J. Flunn // ACM Computing Surveys, Vol. 31. – № 3. – September 1999. – P.265–323.

6. **Бериков В. С.** Современные тенденции в кластерном анализе / В. С. Бериков, Г. С. Лбов // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. – 26 с.

7. **Приставка О. П.** Підтримка прийняття рішень в задачах кластерного аналізу / О. П. Приставка, М. Г. Сидорова // Актуальні проблеми автоматизації та інформаційних технологій : зб. наук. праць. – 2011. – Т.15. – С.117–125.

8. **Гусарова Л.** Проверка обоснованности кластерного решения / Л. Гусарова, И. Яцкив // Proceedings of International Conference RelStat'03. – Vol. 2.

9. **Halkidi M.** Clustering validity assessment: Finding the optimal partitioning of a data set/ M. Halkidi, M.Vazirgiannis // Proceedings of ICDM, 2001. – P. 187–194.

10. **Бирюков А. С.** Решение задач кластерного анализа коллективами алгоритмов / А.С. Бирюков, В.В. Рязанов, А.С. Шмаков // Журн. Вычислит. математ. и математ. физики. – 2008. – Т. 48, № 1. – С. 176–192.

Надійшла до редколегії 12.06.2012

УДК 519.248

С.А. Ватковский, Т.Г. Емельяненко

Днепропетровский национальный университет им. Олеся Гончара

НЕЧЕТКИЕ МОДЕЛИ МАРКОВСКИХ ПРОЦЕССОВ В ТЕОРИИ НАДЕЖНОСТИ

Досліджено застосування нечітких Марківських процесів у задачах теорії надійності.

Ключові слова*: теорія надійності, Марківські процеси, нечіткі множини.*

Исследована применимость нечетких Марковских процессов в задачах теории надежности.

Ключевые слова: *теория надежности, Марковские процессы, нечёткие множества.*

It was researched the applicability of fuzzy Markov processes in problems of reliability theory.

Key words: *reliability theory, Markov processes, fuzzy sets.*

Введение и постановка проблемы. Надежность технических систем приобретает все большую значимость в связи с автоматизацией процессов. Различные технические системы могут использоваться в средах, где человеческое присутствие невозможно, или же где требуется большая деликатность и точность, например, в медицинских системах. Отказы в этих системах, вероятно, будут иметь очень серьезные последствия. В этих непредсказуемых рабочих средах отказы более распространены и их тяжелее предугадать.

Увеличивающееся требование производить более надежные системы пробудило интерес к нескольким теориям, используемым в проектировании отказоустойчивости. Такие методы в теории надежности направлены на оценку эффективности новых проектов. Инструменты анализа надежности, такие как деревья отказа и модели Маркова необходимы, чтобы дать представление о том, что преимущества отказоустойчивого проекта ощутимы и стоят затраченных усилий. К сожалению, компонентная интенсивность отказов, используемая в этих расчетах, очень часто зависит от конфигурации и среды, и, таким образом, известна только приблизительно на этапе конструирования. Нечеткие методы оценки

реализуют оценивание явления единичного характера, для которых отсутствуют вероятностные характеристики.

В связи с этим возникает необходимость разработки информационной технологии для создания отказоустойчивых систем.

Анализ последних исследований и публикаций. Стандартные подходы для обеспечения надежности основываются на вероятностной модели, которая часто является несоответствующей для задач такого рода [2]. Теория вероятности часто – сложный и не интуитивный подход, результат которого трудно анализировать. Так же, вероятностный анализ обычно требует больше информации о системе, чем известно, например, средние интенсивности отказов или распределения интенсивности отказов [2]. Обычно, это приводит к неправдоподобным предположениям об исходных данных. Вероятностная парадигма также имеет много ограничений при применении для выборок небольшого объема [4].

Нечеткая логика предлагает альтернативу вероятностной парадигме,– возможность, которая более адекватна для надежности в автоматизированном окружении [2; 4]. Теория возможности применима для расчетов количественной надежности, которая сохраняет неопределенность в исходных данных. Также этот подход применятся для небольших выборок без каких-либо ограничений [4], что делает его подходящим для анализа специализированных систем, а также их прототипов, которые являются важными аспектами в надежности.

Нечеткая логика также повышает надежность исследования через понятие инспекции и восстановления. Стандартные модели надежности обычно имеют дело с двоичным представлением отказа: система работает или отказала. Более гибкая и реалистическая модель позволяет системам медленно ухудшаться, так же как и отказывать мгновенно. Инспекция и восстановление учитывают частичные системные отказы [2].

Среди подходов, используемых для анализа надежности, только деревья отказов и их разновидности адаптированы к нечеткой логике в полной мере. Однако деревья отказов несколько ограничены в применении. Частичные отказы, покрытие, ремонтопригодные системы и другие важные проблемы надежности хорошо не охвачены методом анализа деревьев отказов [3]. Альтернативой данному подходу является использование нечетких Марковских процессов. Однако проведённый обзор нечетких их моделей показал, что они в задачах надежности достаточно не исследовались.

Формулировка цели статьи. Исследовать возможность применения нечетких моделей Марковских процессов в проектировании отказоустойчивых систем.

Основной материал исследования. Компонентную интенсивность отказов иногда очень трудно вычислить точно во время процесса проектирования. Нельзя легко определить факторы окружающей среды и взаимодействия между компонентами прежде, чем созданы несколько прототипов [6]. Полученные четкие значения задают оценку порядка погрешности при использовании методов, основанных на теории вероятностей [2]. Эта проблема существенно уменьшает преимущество, полученное при использовании моделей Марковских процессов для лучшего представления системы.

Нечетким моделям Марковских процессов необходимы оценки возможности нечеткого события. Это добавляет дополнительный элемент нечеткости к системе, с которой нужно иметь дело.

Для решения этой проблемы в большинстве работиспользуются нечеткий интеграл. Нечеткий интеграл – математическое действие над нечетким множеством и нечеткой мерой, выдающей четкий результат. Они позволяют определить возможность нечеткого события, подобного интегрированию четкого события.

Чтобы использовать нечеткие интегралы, необходимо ввести понятие нечеткой меры [8]. Пусть $X = \{x_1, x_2 ..., x_n\}$. Нечеткой мерой g на X является функция, отображающая мощность множества X, P(x) (X и все возможные подмножества X) на интервал [0; 1]

$$g: P(x) \rightarrow [0, 1]$$

У нечеткой меры есть следующие дополнительные свойства:

$$g(\varnothing) = 0$$

$$g(X) = 1$$

$$\forall A, B \subseteq X, если\ A \subseteq B, то\ g(A) \le g(B).$$

Нечеткая мера может рассматриваться как аналог стандартной вероятностной меры, с заменённым традиционным требованием аддитивности на условие монотонности [8].

Нечеткий интеграл Сугено нечеткого множества h(x) на нечеткой мере g определяется следующим образом [8]:

$$S_g(h) = \int h \circ g = \sup_{0 \le \alpha \le 1} (\alpha, g(H_\alpha)), где\ H_\alpha = \{x \in X \mid h(x) \ge \alpha\},$$

где H_α это α – отсечение h.

Второй, часто используемый, нечеткий интеграл – это интеграл Шоке [8]. Используя нотацию, разработанную для интеграла Сугено, интеграл Шоке представляется виде

$$E_g(h) = \int h(H_\alpha) d\alpha.$$

Интеграл Сугено несколько легче использовать при анализе и он чаще упоминается в литературе, в то время как у интеграла Шоке есть определенные теоретические преимущества [8].

Однако, при использовании нечетких моделей Марковских процессов, приходится иметь дело с нечеткими подмножествами X. Согласно стандартному нечеткому подходу необходимо взять максимум пересечения нечеткого множества события и нечеткого множества возможности, $\max_{x \in X}(h(x) \wedge f(x))$ (где $h(x)$ является нечетким событием), следующим образом (рис. 1).

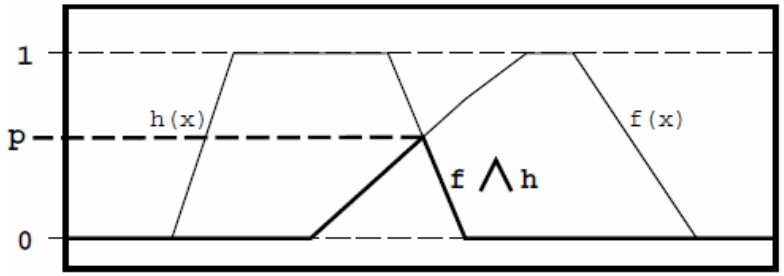


Рис. 1. Нахождение возможности p нечеткого события h при распределении возможности f

Формат модели. Для задач надежности и анализа популяций интенсивности переходов будут нечеткими для нечетких моделей Марковских процессов. Они предоставляют возможность, что данные вероятности корректны.

Модель Марковского процесса – способ определения поведения системы при использовании информации об определенных вероятностях событий в пределах системы. Однако, в задачах надежности, часто значения вероятностей не известны, таким образом, необходимо выдвинуть некоторые предположения.

Приведенный подход оценивает нижние и верхние границы рассматриваемых вероятностей, и использует их для определения трапецеидальной функции принадлежности. Такая оценка легко выполнима для большинства систем, кроме того она обладает свойствами ясности, прозрачности и несложной модифицируемости. Для простоты математического вывода, будем использовать

трапецеидальную функцию принадлежности возможностей, используя строгие границы для 0-отсечения, и оптимистические границы для 1-отсечения.

Вывод результатов для нечеткой модели Маркова является трехмерным, с осями вероятности, возможности и времени. Однако, можно уменьшить размерность до двух, изображая только углы, или контрольные точки распределения возможности, как видно на рис. 2.

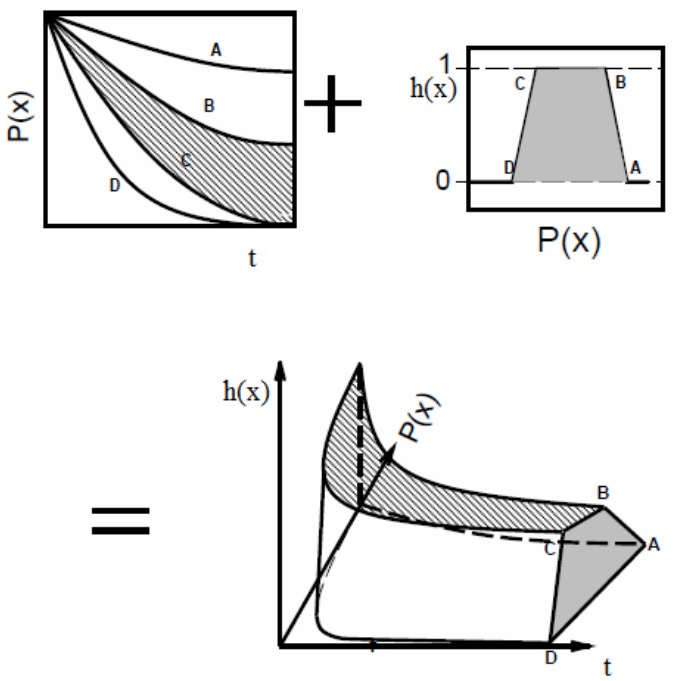


Рис.2. Выходной формат для нечеткой модели Маркова

На рис. 3 представлена модель Марковского процесса для моделирования процесса старения оборудования.

В этой модели работающее состояние поделено на k ухудшенных состояний (D_1, D_2, …,D_k). Инспекция (I) проводится до главного (ММ) или меньшего (М) ремонта. Интенсивности перехода из одного состояния в другое обозначены λ, μ. F_0 и F_1 – это состояния случайного и ремонтного отказа соответственно.

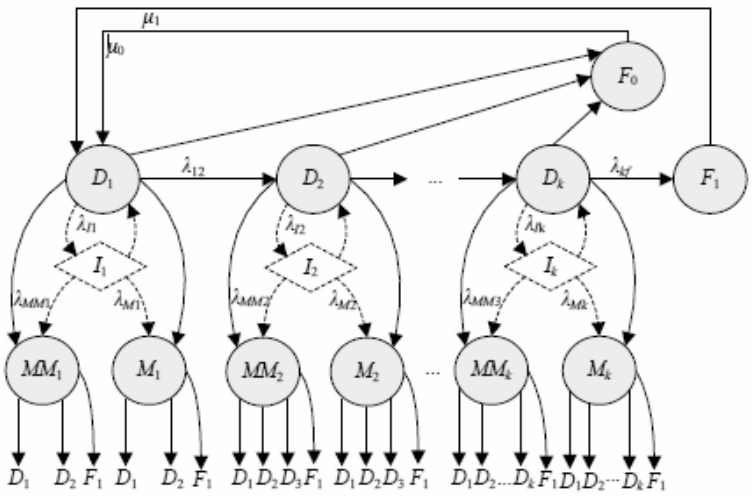


Рис. 3. Диаграмма состояний для стареющего оборудования с инспекцией и ремонтом

Необходимые условия для модели. Есть несколько важных необходимых условий, которые должны быть выполнены для применения нечеткой модели Марковских процессов.

Первое условие – нечеткая модель должна быть лучше в некотором смысле, чем обычная модель Маркова.

Другим важным фактором является сложность. Нечеткая модель Маркова, вероятно, будет более сложной, чем четкая модель, поскольку она использует нечеткую плотность возможности, когда четкая модель использует четкие вероятности. Для обычной модели Марковских процессов эти вероятности – константы, что дает минимальную сложность. Для нечетких моделей, используя трапецеидальные функции принадлежности, мы должны будем иметь дело с четырьмя значениями для каждого значения из четкой модели. Взаимодействие между этими значениями может повысить потенциальную сложность. В худшем случае, полный нечеткий набор возможности нужно будет рассмотреть вместо этих четырех точек. Это, вероятно, сильно увеличит сложность модели.

Конечным критерием, по которому будет оценена нечеткая модель Маркова, является адекватность. Модель, которая дает нецелесообразный, не интуитивный, или чрезмерно сложный вывод, вряд ли будет хорошей моделью.

Один из возможных критериев нечеткой адекватности –метод определения адекватности нечеткого множества. Для наших целей у адекватного нечеткого множества есть $H_\gamma \subseteq H_\beta, \forall \gamma \le \beta$, где H_α – это α-отсечение нечеткого вывода. Другими словами, мы требуем, чтобы более низкие возможности всегда покрывали больший интервал, чем более высокие возможности. Кроме того, хотя это не строгий критерий, предпочтительно, чтобы нечеткие множества были непрерывны и гладки, для простоты истолкования и дальнейшего математического вывода.

Другой критерий адекватности – «вероятностная правильность». Основное требование здесь – то, что не существует оценки возможности за пределами интервала [0,1]. Так как эти вероятности невозможны, недопустимо иметь не нулевые возможности для них.

Нечеткая Марковская модель α-отсечений. Одна из возможных фаззификаций модели Маркова – использование методов, подобных нечеткому дереву отказов. При их использовании достаточно распространить экстремальные значения α-отсечения по всему дереву отказов, как будто оно является четким, полученные экстремальные точки берутся как соответствующие α-отсечения для выходного распределения возможности.

К сожалению, этот метод не пригоден для сложных нечетких моделей Маркова, поскольку это допустимо только для тривиальных систем. Как контр пример, приведем систему с тремя состояниями показанную на рис. 4.

Рис. 4. Модель Маркова с тремя состояниями

Изначально система работает в «Рабочем» состоянии, но со временем система перейдет в среднее «Поврежденное», так как из первого существует только один переход. Однако,система из этого состояния когда-нибудь перейдет в последнее «Отказавшее» состояние. Период времени, за который система перейдет из

начального в конечное состояние, и общая форма траектории зависит от значения интенсивностей переходов. Когда мы распространяем экстремаль значений α-отсечения по этой модели, мы быстро обнаруживаем проблему, показанную на рис.5.

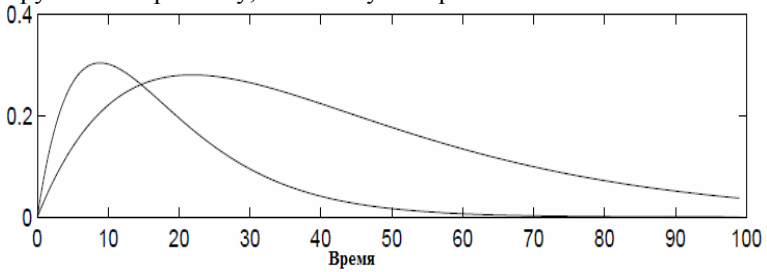


Рис. 5. Сбой экстремумов

Нечеткие модели Маркова, основанные на принципе расширения. Обобщение любых четких операторов к нечетким на нечетких числах может быть выполнено через принцип расширения. Модель решается, как будто она четкая, используя символьные константы для вероятностей возникновения повреждения. К полученным уравнениям результата применяется нечеткая логика, путем замены нечеткими возможностями вероятностных констант и нечеткими операциями четких операций.

Рассмотрим подход использующий принцип расширения для перехода от обычной модели Марковского процесса к нечеткой модели.

Для данной модели Марковского процесса с матрицей нечетких интенсивностей A_λ с доверительным уровнем α и допустимыми границами интенсивностей переходов $[\lambda_\alpha^-, \lambda_\alpha^+]$, интервалы для вероятностей пребывания в состояниях могут быть вычислены по следующим формулам:

$$P_{k,\alpha}^- = \min\{P_k(\lambda) \mid \forall\lambda : \lambda_\alpha^- \le \lambda \le \lambda_\alpha^+\}$$

$$P_{k,\alpha}^+ = \max\{P_k(\lambda) \mid \forall\lambda : \lambda_\alpha^- \le \lambda \le \lambda_\alpha^+\}.$$

Генерация нечетких моделей Маркова через выборку интенсивностей отказов. При исследовании нечеткой модели Марковского процесса с использованием α-отсечения предполагалось приближение, для которого и была решена задача для экстремальных значений отсечений. Естественно рассмотреть приближение, при котором рассматриваются также все промежуточные значения.

Если мы знаем определенный интервал, в котором находится интенсивность отказов, мы можем определить возможное поведение системы, исследуя свойство моделей, следующих из каждого возможного значения на этом интервале. Так, как не известно, какое из этих значений является корректным, можно просто рассмотреть их все. Если это возможно, то такой подход может решить поставленную задачу.

Данный подход является достаточно сложным. Так как интервал содержит бесконечное число точек, и необходимо бесконечное число моделей Марковских процессов для решения задачи. Это невозможно, но если предположить, что модель обладает достаточной гладкостью, такой, что значения выборки интенсивностей отказов дадут подобное поведение системы, то можно рассматривать не весь интервал значений, а только такую выборку интенсивностей отказов. На рис. 6 приведен пример того, как может работать этот метод. Сложность этого подхода все еще высока, но решение задачи теперь доступно.

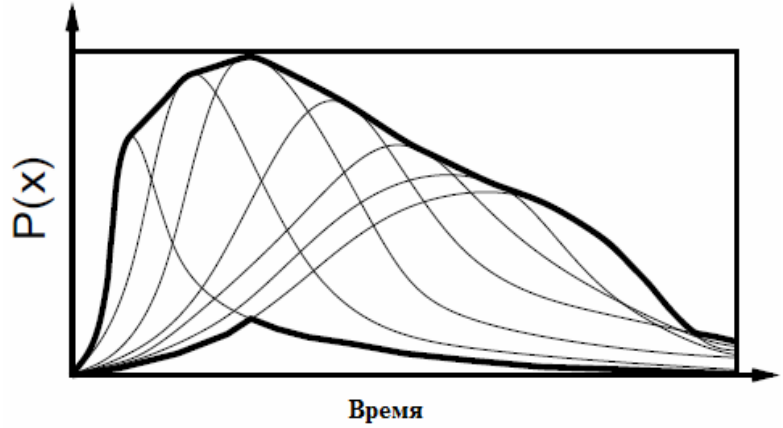


Рис. 6. Нечеткая модель Марковского процесса через выборочный метод закрытия

Несмотря на то, что этот подход основан на переборе, он удовлетворяет всем поставленным техническим условиям для нечеткой модели Марковского процесса за исключением одного – сложность. Выборка интенсивностей отказов требует решения большого количества четких моделей Маркова, для решения единственной нечеткой модели Маркова. Например, если взять N выборок на интервале, и модель имеет M нечетких интенсивностей отказов, то необходимо решить N^M четких моделей Марковских

процессов. Очевидно, что сложность вычисления будет очень быстро расти.

Подход выборки интенсивностей отказов, описанный выше, является методом, который на данный момент используетсядля вычисления нечетких моделей Маркова. Несмотря на проблему сложности, можно показать, что этот метод не теряет важную информацию, не дает невозможный результат или бесполезный вывод. Таким образом, исходная проблема обнаружения нечеткой модели Маркова развилась в проблему упрощения и реализации закрытия, выбирающего нечеткую модель Марковского процесса.

В сложной системе со многими различными частями, вероятно, такой подход будет иметь дело со многими нечеткими интенсивностями отказов, чего более чем достаточно, чтобы сделать нечеткую модель Марковского процесса непрактичной. Однако, исследуя характеристики отказов любой комплексной системы, вполне вероятно, что мы сможем сгруппировать ее части в подсистемы, что увеличит понимание системы, поскольку человеческий разум ограничивается количеством независимых переменных, которые можно рассматривать одновременно.

Можно использовать структуру устройства, для упрощения нечетких моделей Марковских процессов. Все, что следует сделать, так это обнаружить способ группировки интенсивностей отказов отдельных компонентов в единственную компонентную интенсивность отказов. Нечеткие деревья отказов идеальны для этой цели. Их легко реализовать, в их основе лежит нечеткая математика, и они специально предназначены для определения интенсивностей отказов для наборов компонент. Нечеткие модели Марковских процессов, которые используют нечеткие деревья отказа для упрощения, являются мощным инструментом надежности.

Выводы. Нечеткая модель Марковских процессов является адекватным методом для анализа отказоустойчивости систем. Достаточно сложно обеспечить правдивость информации, поддерживая нечеткость, меняющуюся от ситуации. Этот метод хорошо работает вместе с нечеткими деревьями отказа, известным нечетким инструментом надежности.

Слабость нечеткой модели Маркова – это вычислительная сложность. Размерность модели линейно зависит от числа рассматриваемых нечетких распределений возможности. В настоящий момент, только тривиальные модели, или модели с простыми нечеткими деревьями отказа или компонентной группировкой, разрешимы за разумное время.

Дальнейшие исследования будут направлены на сокращение вычислительной сложности модели; изучение возможных методов упрощения модели; разработку информационной технологии, основанной на нечетких Марковских процессах, для исследования характеристик надежности различных технических систем.

Библиографические ссылки

1. **Marius Bazu**. A Combined Fuzzy–Logic & Physics–of–Failure Approach to Reliability Prediction / Marius Bazu // IEEE – Transactions on Reliability, – 1995.

2. **John B. Bowles**. Applying Fuzzy Logic to the Design Process. In Proceedings Annual Reliability and Maintainability / John B. Bowles. // Symposium Tutorial Notes, – 1996.

3. **Mark A. Boyd**. What Markov Modeling Can Do for You. In Proceedings Annual Reliability and Maintainability / Mark A. Boyd // Symposium Tutorial Notes. – 1996.

4. **Kai-Yaun Cai**. System Failure Engineering and Fuzzy Methodology An Introductory Overview. – Cai Kai-Yaun //Fuzzy Sets and Systems. – 1995.

5. **Wen Chun-Yaun**. Fuzzy variables as a basis for a theory of fuzzy Reliability in the Possibility contest /Wen Chun-YaunCai Kai-Yuan, Zhang Ming-Lian // Fuzzy Sets and Systems, 1991.

6. **B. A. Carteret**. Light Duty Utility Arm System Applications for Tank Waste Remediation / B. A. Carteret // Technical report, Westinghouse Hanford Company. – 1994.

7. **Geo Cunlie**. Light Duty Utility Arm for Undeground Storage Tank Inspection and Characterisation / Geo Cunlie // Technical report, Spar Aerospace Ltd. – 1994.

8. **Luis M. de Campos** Characterization and Comparison of Sugeno and Choquet Integrals / Luis M. de Campos // Fuzzy Sets and Systems. – 1992.

Надійшла до редколегії 20.10.2012

УДК 519.682:336.763

В. Б. Говоруха, О. Ю. Лебідь

Академія митної служби України м. Дніпропетровськ

ПРОБЛЕМА ФОРМАЛЬНОГО ВІДОБРАЖЕННЯ ЗМІСТУ ТЕКСТУ

Розглянуто проблему відображення змісту тексту на основі семантичного аналізу. Проаналізовано засоби використання методів аналізу текстів для подальшого пошуку інформації та класифікації документів.

Ключові слова: *семантичний аналіз, система розуміння тексту, пошукова система, класифікація документів, змістове відображення тексту.*

Рассмотрена проблема представления смыслового содержания текста на основе семантического анализа. Проанализированы способы использования методов анализа текста для дальнейшего поиска информации и классификации документов.

Ключевые слова: *семантический анализ, система понимания текста, поисковая система, классификация документов, смысловое представление текста.*

The basic problem representation of the semantic content of text based on semantic analysis is considered. The different methods of text analysis methods for further information retrieval and document classification.

Key words: *semantic analysis, system understanding of the text, the search engine, classified documents, the semantic representation of text.*

Вступ. Розуміння текстів на природній мові та відповідний машинний переклад є одним з напрямків розвитку теорії штучного інтелекту. Останні дослідження в цьому напрямку показали, що для вирішення проблеми недостатньо створити великі сховища словників для перекладу з однієї мови на іншу та відповідні правила перекладу. Тому згодом було створено мову-посередник, яка, у свою чергу, перетворилася на семантичну модель відображення змісту текстів, що перекладаються.

Як відомо, природна мова – це універсальна знакова система, що служить для обміну інформацією між людьми. Оскільки документи на вході документально-пошукових інформаційних систем записані на природній мові, слушно було б поставити питання, а чи не можна

використовувати природну мову як основний засіб відображення інформації під час усього циклу функціонування документально-інформаційних пошукових систем? Відповідь буде позитивною, якщо йдеться про ті інформаційно-пошукові системи, в яких відповідність між запитом і документом встановлює людина. Проте в сучасних документально-пошукових інформаційних системах цю операцію виконує комп'ютер, тому виключається використання природної мови як основного засобу відображення інформації. Це пояснюється істотними особливостями природної мови з погляду машинної технології обробки інформації, такими як різноманіття засобів передачі змісту тексту, семантична неоднозначність, синонімія та багатозначність [1; 2].

Тому для організації пошуку та класифікації документів потрібен етап попереднього аналізу текстів документів. Одним із перспективних напрямів у системах пошуку і класифікації документів виступає семантичний аналіз.

Метою даної роботи є аналіз засобів використання методів семантичного аналізу текстів для класифікації та пошуку інформації.

Результати дослідження. Семантичний аналіз – процес виявлення змісту слів і словосполучень у реченні. Він забезпечує нормалізацію синтаксичної структури речень, розпізнавання термінів, їх класифікацію за семантичними ознаками, з урахуванням синонімічних і гіпонемічних (загальне – часткове) класів, виявлення визначень термінів.

Тематику будь-якого терміна або тексту можна відобразити комбінацією базових семантичних категорій, що асоціюються з ним, і кількість яких уже значно менша, ніж кількість слів. Таким чином, семантичний опис використовує замість слів укрупнені поняття – категорії, кожна з яких характеризується своїм набором термінів. Тому семантичне відображення змісту текстів супроводжується істотним стиском інформації. Стиск інформації при переході від лексичного до семантичного опису документів відбувається за рахунок використання певних знань про структуру мови.

Терміни «семантичний аналіз» і «машинне розуміння тексту» вважаються еквівалентними. За основу в даній роботі взято методи текстології отримання знань, що використовуються під час розробки і ручного наповнення баз знань експертних систем. У разі такого підходу процедури «розуміння» і «здобування знань» є ідентичними, а результат їх виконання формалізується у вигляді деякої семантичної структури. Аналогічно машинне розуміння розглядається у вигляді процесу формування семантичного образу для аналізованого тексту на

природній мові, що виконується системами розуміння тексту (СРТ) (рис. 1).

У СРТ виділено лінгвосемантичне і програмне забезпечення. Перше використовується для опису моделі наочної області і являє собою лінгвістичний і семантичний словники, в термінах яких СРТ формує образ тексту. Програмне забезпечення реалізує відповідні методи аналізу. Роботу СРТ можна поділити на два етапи: лінгвістична обробка і семантична інтерпретація, що виконуються відповідно лінгвістичним і семантичним модулями СРТ.

Лінгвістичний модуль об'єднує етапи безпосередньої обробки природної мови. На цих етапах з використанням словників лінгвістичного забезпечення відбувається первинна формалізація речень вхідного тексту. На етапі графоматичного аналізу виділяються текстові одиниці, такі як слова, речення і абзаци. Крім того, на цьому етапі виключаються незначущі слова і складні конструкції, такі як вступні речення. На етапі морфологічного аналізу визначаються граматичні значення слів, такі як частина мови, рід, число тощо. На етапі синтаксичного аналізу визначається синтаксична структура речення. Найчастіше для опису синтаксису використовується V-мова.

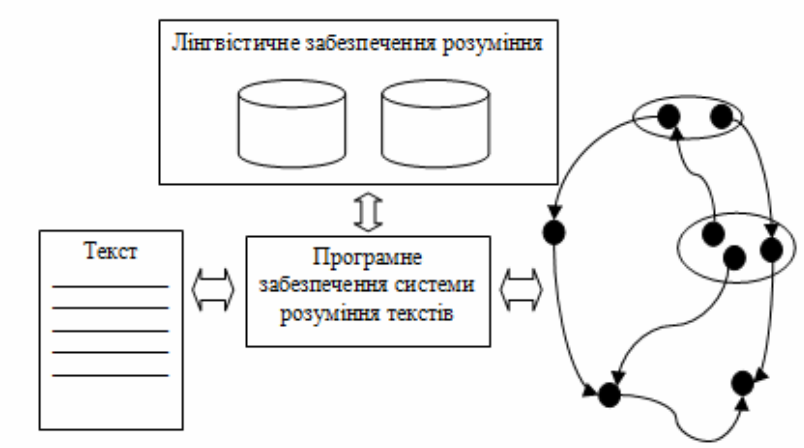


Рис. 1. Функціонування системи розуміння тексту

Семантичний модуль виконує змістовну обробку тексту вхідних даних, отриманими лінгвістичним модулем. Даний вид обробки називається інтерпретацією, оскільки згідно із закладеною в словниках семантичного забезпечення моделлю наочної області визначається формальний сенс окремих формул V-мови. Ця процедура виконується

на етапі семантичного аналізу. На етапі міжфразового семантичного аналізу об'єднуються семантичні зображення окремих речень в єдину семантичну мережу, що описує сенс усього тексту.

Моделі семантичного аналізу почали розвиватися у зв’язку з активним розвитком комп’ютерної обробки текстів. Особливо великі досягнення в цій галузі пов’язані з питаннями пошуку інформації в глобальній мережі Інтернет. Тому логічно розглядати моделі семантичного аналізу в контексті пошуку інформації, а також кластеризації та класифікації документів.

Розглянемо докладніше моделі пошуку. Перший підхід у них базується на теорії множин, другий – на векторній алгебрі, а третій – на теорії вірогідності. Усі ці підходи досить ефективні на практиці, проте в канонічному вигляді всі вони мають спільний недолік, який випливає з припущення, що контент документа, його основний зміст визначається множиною ключових слів – термінів і понять, які входять до нього. Звичайно ж, такий підхід частково веде до втрати змістовних відтінків документів, проте дозволяє виконувати швидкий пошук і групування документів за формальними ознаками. Нині ці підходи найпоширеніші. Слід також зазначити, що існують інші методи, наприклад семантичні, в рамках яких робляться спроби виявити зміст за рахунок аналізу граматики тексту, використання баз знань і тезаурусів, які відображають семантичні зв'язки між окремими словами та їх групами. Очевидно, що такі підходи вимагають істотних витрат на підтримку баз знань і тезаурусів для кожної мови, тематики і виду документів, галузь їх застосування – професійні аналітичні системи.

Булева модель, що базується на теорії множин, є класичною і найбільш поширеною моделлю зображення інформації. Її популярність пов'язана, перш за все, з простотою реалізації, що дозволяє індексувати і виконувати пошук у масивах документів великого обсягу. В даний час популярне об'єднання булевої моделі з векторно-просторовою моделлю алгебри зображення даних, що забезпечує, з одного боку, швидкий пошук з використанням операторів математичної логіки, а з іншого – якісне ранжування документів, що базується на вагах ключових слів, які входять до них. У рамках булевої моделі документи і запити зображуються у вигляді множини морфемних основ ключових слів (терми).

У булевій моделі запитом користувача слугує логічний вираз, в якому ключові слова (терми запиту) пов’язуються операторами з теорії множин і відповідними логічними операторами AND, OR та NOT. У різних пошукових системах, що використовують булеву

модель, зокрема в Інтернеті, користувачі при формуванні запитів можуть просто перераховувати ключові слова, не вказуючи в явному вигляді логічних операцій. Найчастіше при цьому передбачається, що всі ключові слова з'єднуються логічною операцією AND, у цих випадках до результатів пошуку включаються тільки ті документи, які містять одночасно всі ключові слова запиту. У системах, в яких пропуск між словами прирівнюється до оператора OR, у результати пошуку включаються документи, до яких входить хоч би одне з ключових слів запиту. При використанні булевої моделі база даних включає індекс, організований у вигляді масиву, в якому для кожного терма зі словника бази даних міститься список документів, в яких цей терм зустрічається. В індексі можуть зберігатися також частоти даного терма в кожному документі, що дозволяє сортувати список за убуванням частоти. Класична база даних, що відповідає булевій моделі, організована так, щоб за кожним термом можна було швидко дістати доступ до відповідного списку документів. Крім того, структура масиву забезпечує його швидку модифікацію при включенні в базу даних нових документів. У зв'язку з цими вимогами масив часто реалізується у вигляді B-дерева. Існує декілька підходів до формування архітектури пошукових систем, що відповідають булевій моделі і які знайшли своє втілення в реальних системах. Однією з найбільш вдалих реалізацій структури бази даних інформаційно- пошукової системи на мейнфреймах фірми IBM визнано модель даних системи STAIRS (Storage and Information Retrieval System), яка завдяки початковим вдалим архітектурним рішенням досі продовжує розвиватися. База даних інформаційно-пошукових систем цієї традиційної архітектури складається з таких основних таблиць [3]:

- текстова – містить текстову частину всіх документів;
- таблиці покажчиків текстів – включає покажчики місцезнаходження документів у текстовій таблиці, а також поля форматів усіх документів;
- словник – містить всі унікальні слова, що зустрічаються в полях документів, тобто ті слова, за якими може здійснюватися пошук. Слова можуть об'єднуватися в синонімічні ланцюжки;
- масив термів – містить списки номерів документів і координати окремих слів у полях документів.

Процеси, що відбувалися при пошуку інформації в базі даних STAIRS, сьогодні реалізуються засобами сучасних СУБД і інформаційно-пошукових систем документального типу. Пошук

терміна в базі даних здійснюється таким чином. Відбувається звернення до словникової таблиці, за якою визначається, чи входить слово до складу словника бази даних, і якщо входить, то визначається посилання на ланцюжок появи цього слова в документах. Відбувається звернення до масиву термів, за якими визначаються координати всіх входжень терма в текстову таблицю бази даних. За номером документа відбувається звернення до запису таблиці покажчиків текстів. Кожен запис цього файла відповідає одному документу в базі даних. За номером документа відбувається пряме звернення до фрагмента текстової таблиці – документа і подальше його виведення. У разі коли обробляється не один термін, а деяка їх комбінація, в результаті відпрацювання пошуку за кожним терміном запиту формується масив записів, що відповідає входженню цього терміна в базу даних.

Висновки. Актуальним напрямком розвитку систем машинного перекладу, систем пошуку інформації та класифікації документів є використання методів семантичного аналізу.

У статті проаналізовано деякі з методів семантичного аналізу текстів для подальшої класифікації та пошуку інформації. Надалі планується розробка методів та алгоритмів для вирішення зазначених актуальних питань на основі методів семантичного аналізу та подальше їх упровадження в інформаційних системах.

Бібліографічні посилання

1. Нильсон Н. Принципы искусственного интеллекта / Н. Нильсон – М., 1985. – 374 с.

2. Рубашкин В. Ш. Представление и анализ смысла в интеллектуальных информационных системах / В.Ш. Рубашкин – М., 1989. – 258 с.

3. Ландэ Д. В. Определение тематической направленности запросов путем анализа набора рейтинговых источников / Д. В. Ландэ, С. М. Брайчевский // Открытые информационные и компьютерные интегрированные технологии. – Харьков, 2005. – Вып. 29. – С. 169–174.

Надійшла до редколегії 20.07.2012

УДК 004.912:519.766.4

А.А. Дегтярьов, Т.А. Зайцева
Дніпропетровський національний університет ім. Олеся Гончара

СИМПЛІСТИЧНИЙ МЕТОД ТЕМАТИЧНОЇ ІДЕНТИФІКАЦІЇ НАТУРАЛЬНОМОВНОГО ТЕКСТУ НА ОСНОВІ РОЗПОДІЛУ КЛЮЧОВИХ ТЕРМІВ У ТЕКСТІ

Пропонується спосіб оцінки ступеня належності тексту до певної, наперед заданої тематики. Тематичний індекс розраховується як дійсне число та може бути використаний для порівняння з деяким пороговим значенням при визначенні тематичної сутності тексту. Підхід базується лише на припущенні, що задана множина характеристичних слів для оцінюваного тематичого напрямку. Запропонований метод має просту реалізацію, яка підходить як для розв'язку задач, де висока точність оцінки не є пріоритетною вимогою, так і для розв'язку більш складних задач, де даний метод може бути використаний у контексті попередньої оцінки тексту у багатоетапному процесі тематичної ідентифікації.

Ключові слова: *обробка натуральних мов, тематична ідентифікація, розпізнавання тематичного напрямку, тематичний індекс, оцінка тематичної належності тексту, автоматична каталогізація текстів.*

Предлагается метод оценки степени принадлежности текста на натуральном языке некоторой, наперед заданной тематике. Тематический индекс рассчитывается как действительное число и может быть использован для сравнения с некоторым пороговым значением при определении тематической сущности текста. Подход основывается только на предположении о том, что задано множество характеристических слов для оцениваемой тематики. Метод подразумевает простую реализацию, подходящую как для решения задач, где высокая точность оценки не является доминирующим требованием, так и для решения более сложных задач, где предлагаемый метод может быть использован в контексте предварительной оценки текста в многоэтапном процессе тематической идентификации.

Ключевые слова: *обработка естественных языков, тематическая идентификация, распознавание тематического направления, тематический индекс, оценка тематической принадлежности текста, автоматическая каталогизация текстов.*

A simplistic method of topic identification is proposed. The topic index is calculated as a real number, and can be used for comparing with a threshold when validating the topic identity of the content. The approach is based on the assumption of having a set of characteristic words for the estimated topic and yields a simple implementation which is suitable both for the tasks where the high precision is not required and for solving more complex problems, where it can be used in the context of preliminary topic evaluation in a multi-level topic identification process.

Key words: *natural language processing, topic identification, topic recognition, topic index, estimating the relevancy of a text in the context of a given topic, automatic text categorization.*

Постановка проблеми. Проблема визначення тематичної спрямованості отриманого на вхід довільного тексту постає у задачах автоматичної каталогізації ресурсів. Зокрема, ця задача може виникати у різноманітних інтернет-додатках, де вона тісно пов'язана із задачею валідації контента на відповідність певним вимогам, що висуваються до ресурсу. Наприклад, задача тематичної ідентифікації стає актуальною при динамічному наповненні індексної бази певної галузевої пошукової системи, коли необхідно визначити, чи необхідно заданий документ включати до індексу або відкинути його як такий, що не відповідає профілю системи. Також, задача тематичної валідації може застосовуватись при перевірці коректності нового повідомлення на форумі і визначення відповідності змісту повідомлення загальній тематиці форуму, на основі чого системою буде прийняте рішення про публікацію або відкладення його для подальшої перевірки.

Аналіз останніх досліджень і публікацій. Існує ряд підходів до тематичної ідентифікації тексту [4; 5], в межах кожного з яких робиться акцент на певному аспекті тематичної складової текстового потоку. У той час, як практично кожен з них базується на концепції наборів ключових слів, що характеризують певний тематичний напрямок, методи подання цих наборів і встановлення зв'язку між ними та вхідним текстом у контексті його аналізу істотно відрізняються [6; 7]. Варто зазначити, що у сфері тематичної ідентифікації тексту слід розділяти задачі, які сфокусовані на побудові наборів ключових слів для заданих тематичних напрямків, та задачі, які сфокусовані саме на визначенні ступеня належності вхідного тексту певному тематичному напрямку, який задається вже визначеним набором ключових слів.

Методи тематичної класифікації тексту можна розділити на дві групи: лінгвістичні [5] та статистичні [1; 4]. Крім того, існують розроблені методи, що спираються на певні лінгвостатистичні

особливості тексту, зумовлені його авторством [3], або специфікою предметної області [2]. У кожній з цих груп існують досить точні методи, але їх реалізація з метою практичного застосування може вимагати певного об'єму ресурсів, який виявляється необґрунтовано великим при застосуванні для вирішення відносно простих задач. У цьому контексті виникає необхідність у розробці досить простого в реалізації методу, який не потребував би великих витрат як машинних (підготовка та зберігання великого об'єму лінгвістичних даних, час виконання, апаратні ресурси), так і людських (людиногодини розробника програмного забезпечення) ресурсів.

Постановка задачі. Нехай $C = \{w_1, w_2, \ldots, w_n\}$ – вхідний контент, заданий у вигляді скінченої послідовності слів w_i, $i = \overline{1, n}$ певної мови L. Нехай також $D(T) = \{t_1, t_2, \ldots, t_k\}$ – заданий тематичний словник теми T. Тематичним словником теми T будемо називати скінчену множину таких слів t_i, $i = \overline{1, k}$ мови L, які є характеристичними для теми T, тобто поява кожного з них у тексті підвищує імовірність того, що даний текст належить до тематичного напрямку T. Формально це може означати, що питома вага тематичного слова в тексті, про який заздалегідь відомо, що він належить до тематичного напрямку T, є у середньому вищою за питому вагу інших слів, що з'являються у тексті. Тематичність слова може також визначатись тим, що його частота у тематичних текстах вища за частоту, з якою це слово зазвичай з'являється в мові L. Наприклад, у контексті натуральних мов, значне відхилення частоти слова у певному наборі тематичних текстів від значення, що визначено для нього в мовному корпусі, може бути підставою для внесення його до тематичного словника цього тематичного напрямку. Слід сказати, що задача формування наборів характеристичних слів є окремою за своїм характером, і в задачі визначення належності вхідного текстового потоку до певного тематичного напрямку ми будемо вважати тематичний словник для цього напрямку вже сформованим та заданим заздалегідь.

Сформулюємо задачу визначення індексу належності тексту до заданого тематичного напрямку наступним чином: необхідно визначити таку функцію $I: (C, D(T)) \to [0;1]$, яка ставить у відповідність вхідному текстовому потоку певне чисельне значення індексу належності до тематики T за заданим тематичним словником $D(T)$.

Основний матеріал. Концептуальний підхід до розрахунку тематичного індексу полягає у поданні вихідної величини як добутку складових, які також належать проміжку [0; 1], та відображають ступень вираженості тієї чи іншої характеристики щодо властивостей розташування тематичних слів за контентом. Таке розбиття на компоненти дозволяє звести обчислення досить абстрактної величини до обчислення цілком конкретних характеристик текстового потоку, на основі чого можна також робити висновки щодо певних особливостей контента, що розглядається, та проводити додатковий аналіз з метою виявлення інших закономірностей, що можуть представляти інтерес для дослідження в цілому.

Характеристики, що обираються для обчислення сумарного тематичного індексу, повинні бути в певному сенсі незалежними одна від одної, та виявляти різні аспекти вхідного текстового потоку з боку особливостей появи тематичних слів у ньому. Наприклад, очевидно, що обрання лише однієї характеристики, що обчислюється як питома вага тематичних слів у контенті, не може давати вичерпної картини щодо тематики усього контента, бо тематичні слова можуть бути сконцентровані лише у кількох реченнях, де коротко йдеться про предметну область, для якої ці слова є характерними, у той час як весь він у цілому належить до іншої предметної області. У такому випадку, очевидно, треба також враховувати і рівномірність появи слів по всьому текстовому потоку, щоб виключити сильний вплив випадкової появи кількох з них лише в одному фрагменті.

У цьому контексті, в якості характеристик для розрахунку сумарного тематичного індексу текстового потоку, було обрано наступні показники:

- Кількісна оцінка ваги тематичних слів по відношенню до всього контента.
- Оцінка рівномірності розташування тематичних слів у конвенті.
- Оцінка ступеня унікальності набору тематичних слів, що зустрічаються у контенті (оцінка різноманітності представлених на контенті тематичних слів).

Не зважаючи на однобічність, яка властива кількісній оцінці питомої ваги тематичних слів, вона повинна впливати на сумарне значення індексу, бо більш висока частота появи тематичних слів у тексті непрямо підвищує імовірність належності цього тексту до відповідного тематичного напрямку. При обчисленні цієї величини, важливо зазначити, що як правило, при прямому підрахунку, питома вага тематичних слів є досить малою величиною, з погляду на те, що фактично кожна натуральна мова містить велику кількість службових

слів та конструкцій, які не несуть певної тематичної завантаженості, але своєю появою у тексті значно примножують сумарну кількість слів у ньому. Тому, задля поліпшення обчислень, та відходу від замалих величин, має сенс ввести певний істотний поріг питомої ваги тематичних слів, та розраховувати значення першої характеристики відносно цього порогу. Значення порогу повинно бути водночас досить великим, щоб тексти, що не є тематичними, давали значення цієї оцінки близьким до нуля, та досить малим, щоб більшість тематичних текстів отримували значення цієї оцінки близьким до одиниці. Виходячи з цього, першу оцінку пропонується обчислювати наступним чином

$$q(D(T);C)=\begin{cases}\dfrac{N'}{k_q N}, & \dfrac{N'}{N}<k_q\\ 1, & \dfrac{N'}{N}\geq k_q\end{cases}, \qquad (1)$$

де $N-$ загальна кількість слів у контенті C, $N'-$ кількість тематичних слів у контенті C (або кількість елементів у послідовності $Q=\{v:(v\in C)\wedge(v\in D)\}$), k_q – параметр, який визначає істотний поріг питомої ваги тематичних слів. Значення порогу, у загальному випадку, залежить від мови, та визначається особливостями тематичних текстів, поданих в якості бази для автоматичного навчання системи. Наприклад, значення порогу може бути обчислено як середнє значення питомої ваги тематичних слів у цих текстах, про які заздалегідь відомо, що вони належать певній тематиці. Емпірічно було встановлено, що у більшості випадків оптимальним значенням є $k_q=0.15$.

Значення другої обраної характеристики, а саме оцінки рівномірності розташування тематичних слів у контенті, пропонується обчислювати базуючись на підході, описаному у [7] для підрахунку рівномірності розташування тегів частин мови у реченні. Після адаптації у даному контексті, підсумкова формула виглядає наступним чином:

$$d(D(T);C)=1-\mu W(D(T);C), \qquad (2)$$

де $W(D(T);C)$ – величина, яку назвемо дисперсією тематичних слів із словника $D(T)$ у контенті C, та яка обчислюється за наступною формулою

$$W(D(T);C)=\frac{1}{N'+1}\sum_{i=0}^{N'}\left(d_i-\frac{N-N'}{N'+1}\right)^2. \qquad (3)$$

У загальному випадку, ця величина потребує нормалізації до інтервалу [0; 1] задля використання як множника при підрахунку тематичного індексу. Коефіцієнт нормалізації μ виконує цю функцію, та обчислюється наступним чином

$$\mu=\frac{1}{N'}\left(\frac{N'+1}{N-N'}\right)^2. \qquad (4)$$

У формулах (3) та (4) величини N та N' визначені так само, як в (1), а d_i - відстань у словах між двома послідовними (i-ою та $(i+1)$-ою) появами тематичних слів у контенті, причому d_1 – це кількість слів від початку тексту до першої появи тематичного слова, а $d_{N'+1}$ – кількість слів від появи останнього тематичного слова в контенті до його кінця. Тривіальний випадок $N'=0$ не розглядається, бо в такому випадку сумарний тематичний індекс можна вважати нульовим, без обчислення жодної із характеристик. Формула (2) випливає з наступного твердження: чим менша сума квадратів різниць відстаней між послідовними появами тематичних слів у контенті та величиною, що виражає відстань між тематичними словами при ідеальній рівномірності, тим рівномірнішим є розташування слів за контентом, і тим більшим повинне бути значення другої характеристики. Коефіцієнт нормалізації обчислюється виходячи з найгіршого з боку рівномірності випадку: якщо всі тематичні слова розташовані послідовно на початку або в кінці контента, то вектор відстаней має вигляд $\left(\underbrace{0,0,0,...,0}_{N'},(N-N')\right)$, а міра рівномірності вважається рівною нулю, тобто коефіцієнт μ фактично визначається з рівняння $d(D(T);C)=0$.

Третьою характеристикою, яку пропонується обчислювати та включати до складу компонент сумарного тематичного індексу, є міра унікальності тематичних слів у наборі, що з'являється у контенті. Ця характеристика виражає те, наскільки різноманітним є набір тематичних слів на контенті. Урахування цієї міри дозволяє виключити сильний вплив випадкової, але частої появи лише кількох тематичних слів у нетематичному контенті, що може виникати, наприклад, при невирішених омонімічних колізіях у тематичному

словнику. Чисельне значення цієї характеристики пропонується обчислювати наступним чином

$$u(D(T);C)=\begin{cases}\dfrac{N''}{k_u N'}, & \dfrac{N''}{N'}<k_u\\ 1, & \dfrac{N''}{N'}\geq k_u\end{cases}, \qquad (5)$$

де N та N' визначені так само, як в (1), N'' – кількість унікальних тематичних слів у контенті C (тобто кількість всіх слів контента, які належать тематичному словнику $D(T)$, без урахування дублікатів), а k_u – істотний поріг унікальності, який введено з таких міркувань, як і істотний поріг питомої ваги у (1). Значення порогу може коливатись та є менш чутливим з погляду впливу на об'єктивне формування сумарного тематичного індексу. Значення порогу залежить від предметної області, характеру тематичного словника та його потужності, а також мови контента, що оцінюється. За результатами проведених експериментів для різних крайових випадків, було встановлено, що адекватне значення цього параметра лежить у проміжку [0.3; 0.6]. Точне значення порогу повинно обиратись залежно від обсягу тематичного словника, а також виходячи із специфіки предметної області та особливостей уживання ключових термів у типовому контенті, що належить даній тематиці.

Після обчислення кожної з наведених характеристик, сумарний тематичний індекс контенту обчислюється за наступною формулою

$$I(D(T);C)=q(D(T);C)*d(D(T);C)*u(D(T);C). \qquad (6)$$

Слід зазначити, що у певних випадках, з метою оптимізації часу та пам'яті, що витрачаються при подальших обчисленнях з використанням значення сумарного тематичного індексу, має сенс зводити обчислене значення до цілочисельного інтервалу [0; 255] та використовувати його у подальшому.

Проілюструємо розрахунок тематичного індексу за допомогою запропонованого методу на прикладі першого абзацу частини 1 роботи [5]. В якості тематичного напрямку, належність до якого підлягає аналізу, будемо розглядати обробку натуральних мов (англ. Natural Language Processing), а в якості тематичного словника оберемо множину специфічних для даної тематики слів. Базовий набір тематичних слів може бути складений на основі енциклопедичної статті, яка дає широке означення обробки натуральних мов як області штучного інтелекту (прикладом може бути стаття з вікіпедії). Після

первісного аналізу вхідного тексту можна скласти наступний набір слів, які з високою імовірністю будуть присутніми у кожному тематичному словнику за даною тематикою:

$$D(T)=\{"linguistics","syntactic","word",\\ "model","grammar","text","sentence",\\ "graph","relation","subject","meaning","verb"\} \qquad (7)$$

Для проведення аналізу корисно побудувати карту розташування слів за контентом. Зазначимо, що для підвищення точності статистичного аналізу, до загальної карти має сенс включати лише ті слова, які можуть істотно впливати на загальний зміст тексту, і, відповідно, виключати всі службові елементи. До службових елементів, у першу чергу, слід віднести всі цифрові слова (роки, номери тощо), та деякі службові частини мови (сполучники, займенники, прийменники, частки та артиклі). Така фільтрація є простою у реалізації, так як списки слів, що належать до другорядних частин мови, досить малі, повністю відомі, та практично позбавлені необхідності розв'язання омонімічних колізій. Після фільтрації, можна побудувати наступну карту розташування слів (тематичні слова виділені спеціальним чином):

Карта розподілу тематичних слів по контенту

Syntactic	dependency	representations	long	history	descriptive	theoretical	**linguistics**
Many	formal	**models**	advanced	most	notably	**Word**	**Grammar**
Hudson	**Meaning**	**Text**	Theory	Melchuk	Functional	Generative	Description
Sgall	Hajicov	Panevov	Constraint	dependency	**Grammar**	Maruyama	Common
Common	theories	notion	directed	**syntactic**	dependencies	**words**	**sentence**
Example	given	Figure	**sentence**	hearing	scheduled	issue	today
Extracted	Penn	Treebank	Marcus	Santorini	Marcinkiewicz	dependency	**graph**
Sentence	represents	**word**	**syntactic**	modifiers	labeled	directed	arcs
Arc	label	comes	finite	set	representing	possible	**syntactic**
Roles	Returning	Example	figure	see	multiple	instances	labeled
dependency	**relations**	finite	**verb**	hearing	labeled	SBJ	indicating
Hearing	head	**syntactic**	**subject**	finite	**verb**	artificial	**word**
Inserted	beginning	**sentence**	always	serve	single	root	**graph**
Primarily	means	simplify	computation				

За даною картою розташування загальна кількість слів, що підлягає аналізу, дорівнює $N=108$, кількість тематичних слів становить $N'=25$, а кількість унікальних слів у тематичному наборі становить $N''=12$. Проводячи розрахунок розглянутих характеристик за формулами (1), (2), (5), отримуємо наступні результати:

$q(D(T);C)=1$, так як $N/N'=0.2315>k_q=0.15$;

$$d(D(T);C) = 1 - \frac{1}{25} * \left(\frac{26}{108-25}\right)^2 * \frac{1}{26} * \sum_{i=1}^{26}\left(d_i - \frac{108-25}{26}\right)^2 = 0.9502$$

при векторі відстаней визначеному з карти як $(0, 6, 2, 3, 0, 1, 0, 10, 6, 1, 0, 3, 11, 0, 1, 0, 11, 9, 1, 6, 0, 1, 1, 2, 4, 4)$;

$u(D(T);C)$, так як $N / N' = 12/25 = 0.48 > k_u = 0.45$.

Сумарна оцінка ступеня належності поданого тексту до розглянутої тематики дорівнює:

$$I(D(T);C) = q(D(T);C) * d(D(T);C) * u(D(T);C) = 1 * 0.9502 * 1 = 0.9502.$$

Обчислене значення слід розглядати як досить високе та ідентифікувати текст як такий, що належить тематиці, представленій розглянутим тематичним словником.

На практиці важливим вдосконаленням такого підходу є врахування можливості присутності на контенті слів, що належать до іншої тематики або є небажаними. У такому разі тематичний індекс контенту являє собою алгебраїчну суму значень, обчислених за вищенаведеним принципом, взятих із відповідним знаком, при цьому множина тем розділяється відповідно на дві підмножини. Також важливим вдосконаленням є розділення вхідного потоку на основний текст та підзаголовки, що можливо практично для кожного популярного формату. При такому уточненні сумарний тематичний індекс контенту розраховується як середнє зважене окремо розрахованих індексів за основним текстом та підзаголовками, причому підзаголовки можуть отримувати більшу вагу.

Після розрахунку значення тематичного індексу, необхідно прийняти рішення про валідацію контента як такого, що підходить або не підходить вимогам, висунутим у конкретній задачі. Для цього заздалегідь доцільно задати значення порогу тематичного індексу, при подоланні якого відповідний ресурс визнається валідним.

Висновки. У результаті проведеного дослідження запропоновано метод оцінки належності вхідного натуральномовного тексту до певного тематичного напрямку, заданого набором нормалізованих ключових термів. У контексті запропонованого методу побудовано функцію розрахунку тематичного індексу (7), яка представляє собою добуток нормалізованих множників, що визначаються за формулами (1), (2), (5), та семантично представляють основні характеристики розподілу ключових термів у вхідному текстовому потоці (кількість, рівномірність та унікальність).

Запропонований підхід до розв'язку задачі тематичної валідації тексту є досить універсальним, і може бути використаний у силу своєї швидкісності практично у всіх проектах, де швидкість є пріоритетною вимогою. Серед таких задач варто відзначити побудову системи інформаційного пошуку, зокрема системи автоматичного супроводження процесу огляду літератури під час виконання роботи студентами, фахівцями, аспірантами та науковими співробітниками, в ході якого виконується автоматична каталогізація літератури та уточнення набору тематик у процесі пошуку певного питання.

Бібліографічні посилання

1. **Steyvers M., Griffiths T.** Probabilistic topic models. / Latent Semantic Analysis: A Road to Meaning – University of California, Irvine, 2007. – P. 1–15.
2. **Andrzejewski D., Zhu X., Craven M., Recht B.** A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation using First-Order Logic // IJCAI – 2011.
3. **Dai M. A., Storkey A. J.** The Grouped Author-Topic Model for Unsupervised Entity Resolution. // ICANN – 2011.
4. **Singhal A., Mitra M., Buckley C.** Learning routing queries in a query zone. // In Proc. of the SIGIR'97. – 1997. – P. 25–32.
5. **Hatzivassiloglou V., Gravano L., Maganti A**. An investigation of linguistic features and clustering algorithms for topical document clustering. // In Proc. of the SIGIR'2000. – 2000.
6. **Salton G., Allan J., Singhal A.** Automatic text decomposition and structuring. // Information Processing & Management. – 1996. – 32(2) – p. 127–138.
7. **Salton G., Singhal A., Mitra M., and Buckley C.** Automatic text decomposition and summarization. // Information Processing & Management. – 1997. – 33(2) p. 193–208.
8. **McDonald R., Nivre J.** Analyzing and Integrating Dependency Parsers. // Computational Linguistics. – vol. 37. – 2011.
9. **Дегтярьов А. А.** Контекстно-зважена тематична валідація тексту довільного характеру. /А. А. Дегтярьов // Системний аналіз та інформаційні технології: Матеріали 12-ї міжнародної науково-технічної конференції SAIT 2010, Київ, 25-29 травня, 2010 р. – К., 2010. – 544 с.
10. **Дегтярёв А. А.** Лингвостатистическая модель оценки репрезентативности текстового фрагмента с применением адаптационных механизмов определения границ предложения. / А. А. Дегтярёв, Т. А. Зайцева // Питання прикладної математики і математичного моделювання. – Дніпропетровськ, 2012. – С. 94–102.

Надійшла до редколегії 11.07.2012

УДК 519.816:519.237.8

Т.Г. Ємел'яненко

Дніпропетровський національний університет ім. Олеся Гончара

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПОБУДОВИ ПРОГНОЗІВ З ВИКОРИСТАННЯМ МАТЕМАТИЧНОГО АПАРАТУ НЕЧІТКОЇ ЛОГІКИ ДЛЯ ФІНАНСОВИХ ЧАСОВИХ РЯДІВ

Запропоновано використовувати нечітку логіку з метою побудови прогнозних моделей фінансових часових рядів. Виконано порівняльний аналіз використання методу прогнозування на основі теорії японських свічок та на основі нечітких логічних відношеннях вищих порядків.

Ключові слова: *нечітка логіка, нечіткий часовий ряд, прогнозування, фінансові часові ряди.*

Предложено использовать нечеткую логику с целью построения прогнозных моделей финансовых временных рядов. Выполнен сравнительный анализ использования метода прогнозирования на основе теории японских свечей и на основе нечетких логических отношений высших порядков.

Ключевые слова: *нечеткая логика, нечеткий временной ряд, прогнозирование, финансовые временные ряды.*

It is proposed to use fuzzy logics for making prognosis models of financial time series. It has made comparison between Japanese candlestick theory and method based on high-order fuzzy logical relationships.

Key word: *fuzzy logics, fuzzy time series, forecasting, financial time series.*

Вступ. Фінансові прогнози, такі як передбачення варіації цін на акції або прогнозування флуктуації майбутнього індексу, є однією з важливих задач під час розробки програмних продуктів для комерційних застосувань. Важливість цієї задачі викликала розробку великої кількості концепцій та технологій у фундаментальному та технічному аналізі. Існує велика кількість програмних продуктів, що дозволяють будувати прогнози, як правило, в них реалізовані такі методи: штучні нейронні мережі, нечіткі нейронні мережі, генетичні алгоритми, класифікаційні та регресійні дерева, наївний класифікатор Байєса, метод опорних векторів, нечіткі часові ряди тощо. Актуальним є використання інтелектуальних методів, які трансформують

фінансові дані у нечіткі правила та візуальні моделі, які покликані заповнити недоліки перерахованих вище традиційних методів фінансового прогнозування. Використання предметно-орієнтованих знань може покращити результати прогнозування нечітких часових рядів. Однією з переваг такого підходу є те, що отримані результати стають зрозумілими та піддаються тлумаченню. Ефективність такої системи залежить від якості запропонованих знань і може бути покращена за рахунок оновлення бази правил.

Аналіз останніх досліджень і публікацій. Існує значна кількість робіт присвячених розгляду задачі прогнозування з використанням нечіткого підходу. У роботах Q. Song, B.S. Chissom [1–3] запропоновано концепцію нечіткого часового ряду, яка базується на теорії нечітких множин. Задачам фінансового прогнозування з використанням нечіткого підходу присвячені роботи S. M. Chen and N. Y. Wang [4], H. K. Yu and K. H. Huarng [5], H. J. Teoh, T. L. Chen, C. H. Cheng, and H. H. Chu [6] та багатьох інших. Представлена робота присвячена розробці інформаційної технології, яка дозволить виконувати порівняльний аналіз різних підходів з метою підтримки прийняття рішень під час вибору методу прогнозування фінансових даних різної природи (котирування валюти, курси акцій) за декількома критеріями.

Постановка задачі. Задано часовий ряд $\left\{x_t^{(1)}, x_t^{(2)}, x_t^{(3)}, x_t^{(4)}; t = \overline{1, N}\right\}$, який є відображенням динаміки зміни фінансового показника і компоненти кожного елемента якого означають денну ціну відкриття, закриття, найвищу та найнижчу ціни відповідно, N – кількість спостережень. Необхідно побудувати прогноз, тобто побудувати продовження часового ряду у вигляді

$$x_t, \quad t = \overline{1, N + l},$$

де l – довжина прогнозу.

Основний матеріал. Запропоновано розробити інформаційну технологію, що базується на використанні комбінації методу прогнозування на основі теорії японських свічок та нечітких логічних відношеннях вищих порядків.

Використання свічкового патерна для прогнозування фінансових часових рядів складається з таких кроків [7].

Крок 1. Підготовка даних для набору свічкових патернів. На цьому кроці мають бути підготовлені ціни відкриття, закриття, найвища та найнижча. Процент варіації між двома цінами закриття у моменти часу t та $t + n$ обчислюється за формулою

$$\frac{x_{t+n}^{(2)} - x_t^{(2)}}{x_t^{(2)}} \cdot 100 \qquad (1)$$

для отримання наступного тренда. Базуючись на варіації, отриманій з (1), ми можемо знайти найменший $I_{\min}$ та найбільший приріст $I_{\max}$. Таким чином, визначаємо область розгляду $UoD = [I_{\min} - D_1; I_{\max} + D_2]$, де D_1 та D_2 – відповідні додатні числа. Наприклад, якщо $I_{\min}$ = – 5.83 і $I_{\max}$ = 7.66, ми можемо взяти D_1 = 0.17 та D_2 = 0.34, область UoD = [–6,8].

Крок 2. Поділ області розгляду на інтервали $u_1, u_2, ..., u_m$. Беручи UoD з кроку 1 в якості прикладу, ми можемо поділити цю область на 7 інтервалів: u_1 = [–6, –4], u_2 = [–4, –2],…, u_7 = [6,8].

Крок 3. Визначення нечітких множин на області розгляду UoD. Цей крок задає лінгвістичні змінні, які представлені нечіткими множинами, для опису степені варіації між даними у момент часу t t та $t+n$. Якщо ми задамо кількість нечітких множин, рівну 7, вони можуть мати, наприклад, такий сенс:

A1 = (LARGE_DECREASE),
A2 = (NORMAL_DECREASE),
A3 = (SMALL_DECREASE),
A4 = (SMALL_INCREASE),
A5 = (NORMAL_INCREASE),
A6 = (LARGE_INCREASE),
A7 = (EXTREME_INCREASE).

Крок 4. Фаззифікація значень варіації, отриманих на кроці 1. Якщо значення варіації дорівнює $v \in u_i$ і це значення представлене нечіткою множиною A_j, максимум функції приналежності для якої досягається на множині u_i, фаззифікуємо v як A_j.

Крок 5. Обчислення свічкових патернів. На цьому кроці для розрахунку параметрів свічкових патернів використовуються визначення для свічкового патерна, наведені вище. Дані стосовно цін відкриття, закриття, найвищої та найнижчої трансформуються у свічки.

Крок 6. Очищення отриманих патернів. Оскільки свічкові моделі одержані на основі необроблених даних, було б бажано знайти більш важливі атрибути для впливу на наступний тренд для даної моделі. Це є задачею класифікації. Оскільки дані представлені символічними

патернами, основаними на наступному тренді, для класифікації отриманих нечітких свічкових патернів може бути застосований алгоритм C4.5. У результаті роботи цього алгоритму ми фільтруємо атрибути, які є менш важливими для наступного тренда.

Крок 7. Вибір патернів для прогнозування. На цьому кроці ми знаходимо, коли патерн з'являється і з якою імовірністю реальний наступний тренд відноситься до кожної з передбачених нечітких множин. Ця імовірність може бути обчислена за формулою Байєса

$$p(A_i \mid p_j) = \frac{p(p_j \mid A_i) \cdot p(A_i)}{p(p_j \mid A_i) \cdot p(A_i) + p(p_j \mid A_{\bar{i}}) \cdot p(A_{\bar{i}})}, \qquad (2)$$

де p – імовірність подій; A_i – подія, яка визначається тим, що наступний тренд відноситься до нечіткої множини i; p_j – подія, яка визначається появою патерна, що ідентифікується; $A_{\bar{i}}$ – подія, яка визначається тим, що наступний тренд не відноситься до нечіткої множини i. Апріорні ймовірності $p(A_i)$ можуть бути отримані з попередніх експериментів і $p(A_{\bar{i}}) = 1 - p(A_i)$. Апостеріорна ймовірність $p(A_i \mid p_j)$ оцінюється більш точно за рахунок більшої кількості експериментальних даних. Вираз для вибору патерна може бути скорочений таким чином:

$$T = \frac{count(p_j \cap A_i)}{count(p_j \cap A_i) + count(p_j \cap A_{\bar{i}})} = \frac{count(p_j \cap A_i)}{count(p_j)}, \qquad (3)$$

де T – поріг для вибору патерна, терм $count$ є кількістю появ патерна p_j з наступним трендом A_i, і p_j із наступним трендом $A_{\bar{i}}$.

Наприклад, нехай правило вибору має такий вигляд: «Якщо $T > 0.5$, вибираємо патерн». Припустимо, що необхідно обчислити поріг T для очищеного патерна p_a для з'ясування необхідності його вибору. Наступний тренд для p_a – A_4. Шляхом розгляду всіх виділених на кроці 5 патернів визначимо, що серед них є три патерни p_x, p_y та p_z, які можуть бути виведені з патерна p_a. Для p_x та p_y наступний тренд – A_4, а для p_z – A_2. Тоді можемо отримати T за такою формулою

$$T = 2/(1+2) = 0{,}67,\ 0{,}67 > 0{,}5$$

Цей патерн буде обраний для прогнозування.

Крок 8. Прогнозування наступного тренду. Використовуємо процедуру, описану в кроці 5 для представлення тестових даних у вигляді тестових свічкових патернів, порівнюємо їх із обраними на кроці 7 патернами і застосовуємо такі правила для прогнозу наступного тренда для тестових патернів:

Правило 1. Якщо тестовий патерн не може бути знайдений порівнянням із обраними, тоді встановлюємо значення варіації наступного тренда рівним 0.

Правило 2. Якщо існує точно один обраний патерн, з якого може бути виведений тестовий і наступний тренд для цього патерна – A_k, тоді в якості варіації наступного тренда для тестового патерна використовуємо центр ваги m_k нечіткої множини A_k.

Правило 3. Якщо підходять для виведення декілька з обраних патернів, використовуємо середнє арифметичне центрів ваги для обчислення наступної варіації для тестового патерна

$$v_{following} = \frac{\sum_{i=1}^{k} m_i}{k}. \quad (4)$$

Нарешті, прогнозні дані можуть бути отримані за такою формулою

$$x_{t+1} = x_t^{(2)} + x_t^{(2)} \cdot v_{following}. \quad (5)$$

Далі наведені основні етапи прогнозування з використанням нечітких логічних відношень вищих порядків.

Нехай U – це область розгляду, де $U = \{u_1, ..., u_n\}$. Нечітка множина A_i області розгляду може бути визначена таким чином

$$A_i = f_{A_i}(u_1)/u_1 + f_{A_i}(u_2)/u_2 + ... + f_{A_i}(u_n)/u_n,$$

де f_{A_i} – це функція приналежності нечіткої множини A_i, $f_{A_i} : U \to [0;1]$, $f_{A_i}(u_j)$ є степенем приналежності u_j нечіткій множині A_i, $f_{A_i}(u_j) \in [0;1]$ і $1 \le j \le n$.

Нехай $Y_t, t = 1,2,...$ – це область розгляду, яка є підмножиною множини R. Припустимо, що $f_i(t), i = 1,2,...$ визначені в області розгляду Y_t, а $F(t)$ є колекцією $f_i(t), i = 1,2,...$. Тоді $F(t)$ називається нечітким часовим рядом ряду $Y_t, t = 1,2,...$.

Якщо існує нечітке відношення $R(t-1,t)$, таке що $F(t) = F(t-1) \circ R(t-1,t)$, де символ «$\circ$» являє собою max-min оператор, кажемо, що $F(t-1)$ спричиняє $F(t)$.

Нехай $F(t-1) = A_i$ і $F(t) = A_j$, де A_i і A_j є нечіткими множинами, тоді нечітке логічне відношення між $F(t-1)$ і $F(t)$ може бути визначене так: $A_i \to A_j$, де A_i і A_j називаються відповідно лівою та правою частинами нечіткого логічного відношення.

Нечіткі логічні відношення, які мають спільну ліву частину, можуть бути об'єднані в групи. Наприклад, припустимо, що існують наступні нечіткі логічні відношення:

$$A_i \to A_{ja},$$
$$A_i \to A_{jb},$$
$$\ldots,$$
$$A_i \to A_{jm}.$$

Тоді ці логічні відношення можуть бути згруповані так:

$$A_i \to A_{ja}, A_{jb}, ..., A_{jm}.$$

Нехай $F(t)$ – це нечіткий часовий ряд. Якщо $F(t)$ спричиняється $F(t-1)$, $F(t-2)$, …, $F(t-n)$, тоді цей взаємозв'язок може бути представлений «нечітким логічним відношеннями *n*-го порядку»:

$$F(t-n), \ldots, F(t-2), F(t-1) \to F(t).$$

Якщо $F(t-n) = A_{in}$, $F(t-2) = A_{i2}$, $F(t-1) = A_{i1}$, $F(t) = A_j$ де A_{in}, …, A_{i2}, A_{i1}, A_j – нечіткі множини, тоді нечітке логічне відношення *n*-го порядку можна подати у такому вигляді:

$$A_{in}, \ldots, A_{i2}, A_{i1} \to A_j,$$

де A_{in}, …, A_{i2}, A_{i1} називають попередніми нечіткими множинами нечіткого логічного відношення *n*-го порядку; A_{in}, …, A_{i2}, A_{i1} і A_j називаються відповідно лівою та правою сторонами відношення.

Якщо існують нечіткі логічні відношення *n*-го порядку, які мають спільну ліву частину:

$$A_{in}, \ldots, A_{i2}, A_{i1} \to A_{ja},$$

$$A_{in}, \ldots, A_{i2}, A_{i1} \to A_{jb},$$
$$\ldots,$$
$$A_{in}, \ldots, A_{i2}, A_{i1} \to A_{jm},$$

тоді ці відношення формують групу нечітких логічних відношень

$$A_{in}, \ldots, A_{i2}, A_{i1} \to A_{ja}, A_{jb}, \ldots, A_{jm}.$$

Алгоритм побудови прогнозу із використанням нечітких логічних відношень вищого порядку складається з таких кроків [8].

Крок 1. Визначення області розгляду U, $U = [D_{\min} - D_1; D_{\max} + D_2]$ і поділ її на інтервали однакової довжини, де $D_{\min}$ та $D_{\max}$ – мінімальне та максимальне значення історичних даних, а D_1 та D_2 – відповідні додатні числа. Наприклад, якщо мінімальний та максимальний індекси цін на акції складають відповідно 5474.79 та 8608.91, ми можемо взяти D_1 = 74.79 та D_2 = 91.09, область U = [5400, 8700]. Необхідно також визначити довжину кожного з інтервалів і розділити область розгляду на таку кількість інтервалів, щоб довжина кожного складала задану величину. Інтервали можуть бути подані таким чином:

$$u_i = [D_{\min} + (i-1) \cdot l; D_{\min} + i \cdot l], i = 1, 2, \ldots, k$$

де l – довжина одного інтервалу, k – загальна кількість інтервалів.

Крок 2. Визначення представлених нечіткими множинами лінгвістичних змінних A_i:

$$A_1 = \frac{1}{u_1} + \frac{0.5}{u_2} + \frac{0}{u_3} + \frac{0}{u_4} + \ldots + \frac{0}{u_{n-1}} + \frac{0}{u_n},$$
$$A_2 = \frac{0.5}{u_1} + \frac{1}{u_2} + \frac{0.5}{u_3} + \frac{0}{u_4} + \ldots + \frac{0}{u_{n-1}} + \frac{0}{u_n},$$
$$A_3 = \frac{0}{u_1} + \frac{0.5}{u_2} + \frac{1}{u_3} + \frac{0.5}{u_4} + \ldots + \frac{0}{u_{n-1}} + \frac{0}{u_n},$$
$$\ldots$$
$$A_{n-1} = \frac{0}{u_1} + \frac{0}{u_2} + \frac{0}{u_3} + \frac{0}{u_4} + \ldots + \frac{1}{u_{n-1}} + \frac{0.5}{u_n},$$
$$A_n = \frac{0}{u_1} + \frac{0}{u_2} + \frac{0}{u_3} + \frac{0}{u_4} + \ldots + \frac{0.5}{u_{n-1}} + \frac{1}{u_n},$$

де A_1, A_2, …, A_n – лінгвістичні змінні, представлені нечіткими множинами.

Крок 3. Фаззифікація історичних даних із використанням нечітких змінних, визначених на кроці 2. Якщо значення належить до інтервалу u_i і максимальне значення функції приналежності нечіткої множини A_i набувається на цьому інтервалі, тоді значення фаззифікується як A_i, де $1 \le i \le n$.

Крок 4. Побудова нечітких логічних відношень *n*-го порядку на основі фаззифікованих історичних даних. Наприклад, якщо в *i*-ий день індекс фаззифікувався як A_{31}, у $(i+1)$-ий – як A_{32}, а у $(i+2)$-ий – як A_{41}. Тоді ми можемо отримати таке нечітке логічне відношення 2-го порядку

$$A_{31}, A_{32} \to A_{41}.$$

Таким чином, отримуємо всі нечіткі логічні відношення з фаззифікованих історичних даних.

Крок 5. Трансформація кожного нечіткого логічного відношення *n*-го порядку $A_{X_1}, A_{X_2}, A_{X_3}, \ldots, A_{X_i}, \ldots, A_{X_n} \to A_{X_r}$ в таку форму:

$$A_{X_1}, A_{X_1+V(X_1)}, A_{X_1+V(X_1)+V(X_2)}, \ldots, A_{X_1+V(X_1)+V(X_2)+\ldots+V(X_i)},$$
$$\ldots,$$
$$A_{X_1+V(X_1)+V(X_2)+\ldots+V(X_i)+\ldots+V(X_m)} \to A_{X_1+V(X_1)+V(X_2)+\ldots+V(X_i)+\ldots+V(X_m)+V(X_n)},$$

де $V(X_1)$, $V(X_2)$, …, $V(X_n)$ – цілі числа. Наприклад, якщо нечітке логічне відношення другого порядку має вигляд $A_{31}, A_{32} \to A_{41}$ трансформуємо його в $A_{X_1}, A_{X_1+V(X_1)} \to A_{X_1+V(X_1)+V(X_2)}$, де індекси X_1, X_2 та X_3 нечітких множин A_{31}, A_{32} та A_{41} відповідно дорівнюють 31, 32 та 41, тобто $X_1 = 31$, $X_2 = 32$ та $X_3 = 41$. Таким чином, ми отримуємо $V(X_1) = X_2 - X_1 = 32 - 31 = 1$ і $V(X_2) = X_3 - (X_1 + V(X_1)) = X_3 - X_2 = 41 - 32 = 9$. Отже, нечітке логічне відношення другого порядку $A_{31}, A_{32} \to A_{41}$ представляється як $A_{31}, A_{32} \to A_{31+1+9}$. У такий спосіб ми трансформуємо всі нечіткі логічні відношення другого порядку.

Крок 6. Згрупуємо всі трансформовані нечіткі логічні відношення *n*-го порядку, отримані на попередньому кроці, які мають спільну ліву частину, в групи нечітких логічних відношень. Наприклад,

припустимо, що ми одержали наступні трансформовані нечіткі логічні відношення третього порядку:

$$A_{a_1},\ A_{a_1+V(a_1)},\ A_{a_1+V(a_1)+V(a_2)} \rightarrow A_{a_1+V(a_1)+V(a_2)+V(a_3)},$$
$$A_{b_1},\ A_{b_1+V(b_1)},\ A_{b_1+V(b_1)+V(b_2)} \rightarrow A_{b_1+V(b_1)+V(b_2)+V(b_3)},$$
$$\ldots,$$
$$A_{k_1},\ A_{k_1+V(k_1)},\ A_{k_1+V(k_1)+V(k_2)} \rightarrow A_{k_1+V(k_1)+V(k_2)+V(k_3)},$$

де $V(a_1)=V(b_1)=...=V(k_1)$ і $V(a_2)=V(b_2)=...=V(k_2)$. Далі, ці нечіткі логічні відношення третього порядку можуть бути згруповані і представлені в такому вигляді:

$$A_X,\quad A_{X+V(y_1)},\quad A_{X+V(y_1)+V(y_2)} \rightarrow \quad A_{X+V(y_1)+V(y_2)+V(a_3)},$$
$$A_{X+V(y_1)+V(y_2)+V(a_3)},\ A_{X+V(y_1)+V(y_2)+V(b_3)},\ \ldots,\ A_{X+V(y_1)+V(y_2)+V(k_3)},$$

де $X=a_1,b_1,...,k_1$, $V(y_1)=V(a_1)=V(b_1)=...=V(k_1)$ і $V(y_2)=V(a_2)=V(b_2)=...=V(k_2)$. Наприклад, припустимо, що ми маємо такі нечіткі логічні відношення другого порядку: A_{31}, $A_{31+1} \rightarrow A_{31+1+9}$, A_{43}, $A_{43+1} \rightarrow A_{43+1-4}$ і A_{37}, $A_{37+1} \rightarrow A_{37+1-4}$. Тоді трансформовані нечіткі логічні відношення матимуть наступний вигляд: A_x, $A_{x+1} \rightarrow A_{x+1+9}$, A_x, $A_{x+1} \rightarrow A_{x+1-4}$, і A_x, $A_{x+1} \rightarrow A_{x+1-4}$. Вони, в свою чергу, можуть бути об'єднані в одну групу, оскільки мають спільну ліву частину «A_x, A_{x+1}»: A_x, $A_{x+1} \rightarrow A_{x+1+9}$, A_{x+1-4}, A_{x+1-4}.

Крок 7. Вибір групи трансформованих нечітких логічних відношень *n*-го порядку для побудови прогнозу. Нехай $F(t-n)=A_{in}$, $F(t-(n-1))=A_{i(n-1)}$, …, $F(t-2)=A_{i2}$, $F(t-1)=A_{i1}$ і нехай ми хочемо спрогнозувати $F(t)$, де A_{i1}, A_{i2}, …, A_{in} – нечіткі множини. Згідно з отриманими на попередньому кроці групами нечітких логічних відношень *n*-го порядку, оберемо відповідну групу для побудови знаходження прогнозного значення. Якщо було обрано таку групу

$$A_X,\ A_{X+V(y_1)},\ \cdots,$$
$$A_{X+V(y_1)+V(y_2)+V(y_{n-1})} \rightarrow A_{X+V(y_1)+...+V(y_{n-1})+V(a_n)},$$
$$A_{X+V(y_1)+...+V(y_{n-1})+V(b_n)},\ldots,A_{X+V(y_1)+...+V(y_{n-1})+V(k_n)},$$

де $A_{in}=A_X$, $A_{i(n-1)}=A_{X+V(y_1)},\ldots$, $A_{i1}=A_{X+V(y_1)+V(y_2)+...+V(y_{n-1})}$, тоді замінимо X індексом *in* нечіткої множини A_{in} для отримання нечітких множин $A_{in+V(y_1)+...+V(y_{n-1})+V(a_n)}$, $A_{in+V(y_1)+...+V(y_{n-1})+V(b_n)}$, …, $A_{in+V(y_1)+...+V(y_{n-1})+V(k_n)}$ для прогнозування.

Нехай $A_{j1}=A_{in+V(y_1)+...+V(y_{n-1})+V(a_n)}$, $A_{j2}=A_{in+V(y_1)+...+V(y_{n-1})+V(b_n)}$, …, $A_{jk}=A_{in+V(y_1)+...+V(y_{n-1})+V(k_n)}$. Тоді прогнозне значення FV для дня *t* обчислюється за такою формулою

$$FV=\frac{\sum_{i=1}^{k} m_{ji}}{k}, \qquad (6)$$

де максимальні значення функцій приналежності нечітких множин A_{j1}, A_{j2}, …, A_{jk} набувається відповідно на інтервалах u_{j1}, u_{j2}, …, u_{jk}, а m_{j1}, m_{j2}, …, m_{jk} – центри ваги даних нечітких множин.

Апробація запропонованої інформаційної технології виконана на декількох незалежних наборах фінансових даних (курсу акцій банку «Bank Of America Corp»; курсу євро по відношенню до долара; фунта стерлінгів до долара; комплексного показника по акціям підприємств України за індексом UX-C). Побудовано прогнози за допомогою методу, основаного на нечітких свічкових патернах; за допомогою нечітких логічних відношень другого; другого та третього; другого, третього та четвертого порядків. Дані результати порівнювалися із результатами, отриманими застосуванням нечіткого підходу Мамдані після налаштування параметрів функцій приналежності. 75 % кожного з наборів даних використовувалося для навчання системи, всі інші дані – для тестування. У таблицях 1 – 4 наводяться обчислені значення середньоквадратичної, середньої та максимальної помилки для взятих наборів даних.

Таблиця 1

Дані «Bank Of America Corp» за період липень – грудень 2011р.

Метод	Середньо-квадратична помилка	Середня помилка	Максималь-на помилка
Відношення першого порядку	0.09341	0.23260	0.64369
Метод японської свічки	0.02437	0.12042	0.40280
Нечіткі відношення 2-го порядку	0.02073	0.11319	0.33894
Нечіткі відношення 2-го та 3-го порядків	0.02741	0.13175	0.36969
Нечіткі відношення 2-го, 3-го та 4-го порядків	0.03711	0.15783	0.37368

Таблиця 2

Дані «Курс Forex EUR USD» за період листопад 2011р. – березень 2012р.

Метод	Середньо-квадратич-на помилка	Середня помилка	Максималь-на помилка
Відношення першого порядку	0.00010	0.00695	0.03154
Метод японської свічки	0.00012	0.00852	0.02310
Нечіткі відношення 2-го порядку	0.00005	0.00561	0.01492
Нечіткі відношення 2-го та 3-го порядків	0.00006	0.00614	0.01759
Нечіткі відношення 2-го, 3-го порядків та 4-го порядків	0.00007	0.00686	0.01733

Таблиця 3

Дані «Курс Forex GBP EUR» за період листопад 2011р. – квітень 2012р.

Метод	Середньо-квадратична помилка	Середня помилка	Максимальна помилка
Відношення першого порядку	0.00001	0.00250	0.00662
Метод японської свічки	0.00005	0.00650	0.01371
Нечіткі відношення 2-го порядку	0.00001	0.00251	0.00998
Нечіткі відношення 2-го порядку після налаштування параметрів функцій приналежності	0.00001	0.00223	0.00759
Нечіткі відношення 2-го та 3-го порядків	0.00001	0.00260	0.00755
Відношення 2-го, 3-го та 4-го порядків	0.00001	0.00275	0.00752

Таблиця 4

Дані «MTX» за період січень – березень 2012р.

Метод	Середньо-квадратична помилка	Середня помилка	Максимальна помилка
Відношення першого порядку	4519.20150	54.81747	120.74102
Метод японської свічки	7305.44031	64.31292	191.48429
Нечіткі відношення 2-го порядку	6330.71301	61.35305	213.66600
Нечіткі відношення 2-го порядку після налаштування параметрів функцій приналежності	3486.17473	49.66126	143.57980
Нечіткі відношення 2-го та 3-го порядків	5492.42331	60.72588	167.53933
Нечіткі відношення 2-го, 3-го та 4-го порядків	5392.50373	59.11080	164.35955

Таблиця 5

Дані «UX-C 2009-2012» за період 2009-2012 р.

Метод	Середньо-квадратична помилка	Середня помилка	Максималь-на помилка
Відношення першого порядку	3044.90540	46.69723	130.48601
Метод японської свічки	1485.36913	29.06150	110.62950
Нечіткі відношення 2-го порядку	1195.54247	26.65964	109.29556
Нечіткі відношення 2-го порядку після налаштування параметрів функцій приналежності	996.85698	22.74599	106.80162
Нечіткі відношення 2-го та 3-го порядків	1109.52496	25.54228	112.68330
Нечіткі відношення 2-го, 3-го та 4-го порядків	1312.23357	28.06804	111.69616

Із наведених вище результатів видно, що використання нечітких логічних відношень другого порядку окремо та відношень другого і третього порядку забезпечує найвищу точність прогнозування. Для тих рядів, де нечіткі відношення першого порядку давали більш якісні результати, було проведено налаштування параметрів функцій приналежності нечітких відношень другого порядку і отримані результати забезпечували нижчі значення показників помилок. У більшості випадків підхід із застосуванням нечітких свічкових патернів давав також якісні результати, однак він помітно програє

порівняно із більш простим у реалізації підходом із нечіткими логічними відношеннями вищих порядків, навіть без налаштування параметрів функцій приналежності. Також з отриманих результатів видно, що для даних таких об'ємів має сенс будувати прогноз базуючись лише на відношеннях другого та третього порядку. Застосування відношень більш високих порядків не дозволяє покращити результат.

Висновки. Розглянуто два ефективних методи прогнозування фінансових даних – на основі нечітких свічкових моделей та нечітких логічних відношень вищих порядків. Обидва методи забезпечили достатньо якісні результати на розглянутих наборах даних при незначній складності реалізації і тривалості виконання. Показано, що підхід із застосуванням логічних відношень дає кращі результати навіть без налаштування параметрів функцій приналежності нечітких множин та не вимагає від дослідника врахування для побудови прогнозу чотири ціни для кожного проміжку часу, що скорочує обсяг спостережень. З'ясовано, що для даних такого об'єму використання нечітких логічних відношень більш високих, ніж третій, порядків не є доцільним, оскільки такі дані не можуть забезпечити достатню інформацію про внутрішні залежності високих порядків. Для деяких рядів налаштування параметрів функцій приналежності може займати значний проміжок часу, хоча і не забезпечує значного зменшення помилок прогнозу. Тому у випадку, коли максимальна точність прогнозу не є надважливою, можна використовувати функції приналежності із заданими за замовчанням параметрами.

Бібліографічні посилання

1. **Song Q.** Fuzzy time series and its models / Q. Song, B.S. Chissom // Fuzzy Sets and Systems. – Vol. 54, – P. 269 – 277, – 1993.

2. **Song Q.** Forecasting enrollments with fuzzy time series – part I / Q. Song, B.S. Chissom // Fuzzy Sets and Systems. – Vol. 54, – P. 1–9, – 1993.

3. **Song Q.** Forecasting enrollments with fuzzy time series – part I / Q. Song, B.S. Chissom // Fuzzy Sets and Systems. – Vol. 62, – P. 1 – 8, – 1994.

4. **Chen S.M.** Handling forecasting problems based on high-order fuzzy time series and fuzzy-trend logical relationships / S.M. Chen and N.Y. Wang // Proceedings of the 2008 Workshop on Consumer Electronics, Taipei County, Taiwan, Republic of China, – P. 759 – 764, –2008.

5. **Yu H. K.** A bivariate fuzzy time series model to forecast the TAIEX / H.K. Yu and K.H. Huarng // Expert Systems with Applications. – Vol. 34, № 4, – P. 2945 – 2952, – 2008.

6. **Teoh H. J.** A hybrid multiorder fuzzy time series for forecasting stock markets / H. J. Teoh, T. L. Chen, C. H. Cheng, and H. H. Chu // Expert Systems with Applications. – Vol. 36, № 4, – P. 7888 – 7897, –2009.

7. **Chiung-Hon Leon Lee.** Pattern Discovery of Fuzzy Time Series for Financial Prediction / Leon Lee Chiung-Hon // IEEE Transactions on knowledge and data engineering. – Vol. 18, № 5, – 2006.

8. **Chao-Dian Chen.** A New Method for Forecasting the TAIEX Based on High-Order Fuzzy Logical Relationships / Chao-Dian Chen, Shyi-Ming Chen // Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, – 2009.

9. **Ємел'яненко Т.Г.** Використання нечітких моделей в задачах фінансового прогнозування / Т.Г. Ємел'яненко, Д.Л. Самарська // «Людина і космос»: XIV міжнародна молодіжна науково-практична конференція, 11–13 квітня 2012 р.: тези допов. – Д., 2012. – С. 271.

Надійшла до редколегії 29.06.2012

УДК 519.246.8

Я.В. Пирогова, Т.Г. Ємел'яненко

Дніпропетровський національний університет ім. Олеся Гончара

РОЗРОБКА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ КОРОТКОСТРОКОВОГО ПРОГНОЗУВАННЯ НЕЧІТКИХ ЧАСОВИХ РЯДІВ

Створено інформаційну технологію для побудови прогнозів нечітких часових рядів на базі нечіткої логіки з використанням генетичного алгоритму. Розглянутий підхід покладено до основи розробленого програмного продукту Fuzzy_Forecassting, який створено у середовищі Eclipse на мові програмування Java.

Ключові слова: *нечітка логіка, нечіткий часовий ряд, прогнозування, генетичний алгоритм, програмний продукт, гідрогеохімічний моніторинг.*

Создана информационная технология для прогнозирования нечетких временных рядов на базе нечеткой логики с использованием генетического алгоритма. Рассмотренный подход лежит в основе разработанного в среде Eclipse программного продукта Fuzzy_Forecassting на языке программирования Java.

Ключевые слова: *нечеткая логика, нечеткий временной ряд, прогнозирование, генетический алгоритм, программный продукт, гидрогеохимический мониторинг.*

It was created information technology to make forecast for fuzzy time series based on fuzzy logic and genetic algorithm. This approach was used in the developed software product Fuzzy_Forecassting which was created in Eclipse on the programming language Java.

Key word: *fuzzy logics, fuzzy time series, forecasting, genetic algorithm, software product, hydrogeochemical monitoring.*

Вступ. У галузі розробки інформаційних технологій актуальною є задача прогнозування процесу, що аналізується. Це обумовлено тим, що ефективне управління під час рішення як технічних, так і економічних задач, неможливе без достовірного й оперативного прогнозування. У теорії і практиці за минулі роки накопичено значний набір різноманітних методів розробки прогнозів. За оцінками вчених, нараховується понад 150 різних методів прогнозування; на практиці ж в якості основних використовується лише 15–20. Розвиток

інформатики і засобів обчислювальної техніки створює можливість розширення кола методів прогнозування та їх вдосконалення [1].

Мабуть, найбільш вражаючою властивістю людського інтелекту є здатність приймати правильні рішення в обстановці неповної і нечіткої інформації. Побудова моделей наближених міркувань людини і використання їх у комп'ютерних системах представляє сьогодні одне з дуже цікавих наукових питань. Значне просування в цьому напрямку зроблено професором Каліфорнійського університету (Берклі) Лотфі А. Заде (Lotfi A. Zadeh). Його роботи заклали основи моделювання інтелектуальної діяльності людини і стали початковим поштовхом до розвитку нової математичної теорії. Заде розширив класичне канторовське поняття множини, припустивши, що характеристична функція (функція приналежності елемента множині) може приймати будь-які значення в інтервалі [0; 1], а не тільки значення 0 або 1. Такі множини були названі їм нечіткими (fuzzy).

Актуальною є задача дослідження можливості застосування нечіткого підходу до прогнозування техногенного навантаження на територіях гірничо-збагачувальних комбінатів (ГЗК).

Аналіз останніх досліджень і публікацій. Прогнозування, яке базується на нечітких часових рядах (НЧР), притягує увагу дослідників протягом останніх 15 років. Більшість опублікованих до теперішнього часу праць використовували в якості тестової послідовності даних дані реєстрації студентів в Університеті шт. Алабама за період у майже 20 років. Дослідники Сонг та Чісс (Song-Chissom) [6 – 8] розглядали стаціонарні (time-invariant) та змінні у часі (time-variant) НЧР моделі прогнозу, які забезпечили помилки на рівні 3.18 % та 4.37 % (вікно прогнозування дорівнює 4), відповідно. Стаціонарна НЧР модель, яка була запропонована Ченом (Chen) [9] та протестована на тих самих даних, практично не змінила похибку прогнозу Сонга та Чісса, (3.23 %), а модифікований підхід Шаха и Дегтярьова (Şah, Degtiarev) [10] дозволив трохи покращити результати, знизивши середню помилку до 2.42 %. Одночасно, змінна у часі НЧР модель Хванга, Чена та Лі (Hwang-Chen-Lee) [11], залишаючись простою з точки зору матричних обчислень, забезпечила відносну похибку близько 3.12 %.

Найчастіше нечіткі часові ряди використовуються під час прогнозування даних фінансових ринків. Значно менша кількість робіт присвячена прогнозуванню екологічних даних з використанням нечіткого підходу [11 – 15]. Застосуванню нечітких часових рядів для розв'язання задачі прогнозування даних гідрогеохімічного моніторингу не приділено достатньої уваги.

Постановка задачі. Задано часовий ряд , що містить інформацію про концентрацію хімічної речовини в пробах ґрунтових вод на техногенно навантаженій території Північного гірничо-збагачувального комбінату (ГЗК). Необхідно створити інформаційну технологію для побудови прогнозу, тобто продовження часового ряду у вигляді

$$u_t, t = \overline{1, N+1}.$$

Основний матеріал. Останнім часом з'явилося багато різноманітних методів прогнозування, які основані на теорії нечітких множин. Вони відрізняються тим, що використовують різні алгоритми нечіткого виводу. Але ці методи мають певні обмеження. Найвідоміші алгоритми: Мамдані, Цукамото, Ларсена, Сугено. У цілому, зміст цих алгоритмів зводиться до наступних етапів:

1. Фаззифікація – це процедура знаходження значень функції належності вхідних лінгвістичних змінних на основі звичайних (не нечітких) даних. Іншими словами, треба вибрати вид функції належності та визначити проміжки, на яких вхідна лінгвістична змінна асоціюється із заданими на даному етапі лінгвістичними оцінками.

2. Побудова нечіткої бази знань – запис набора нечітких правил (експертних висловлювань), що мають вид «ЯКЩО – ТОДІ», що описують зміну рівнів часового ряду та факторів, які впливають на показники, що прогнозуються. База знань систем нечіткого виводу призначена для формального представлення емпіричних знань та знань експертів у той чи іншій проблемній галузі.

3. Агрегування – це процедура визначення степеня істинності умов за кожним із правил системи нечіткого вивода.

4. Активізація у системах нечіткого вивода – це процедура знаходження степені істинності кожного із підзаключень правил нечітких продукцій.

5. Аккумуляція – процедура або процес знаходження функції належності для кожної із вихідних лінгвістичних змінних.

6. Дефаззифікація – це процедура або процес знаходження звичайного (не нечіткого) значення для кожної вихідної лінгвістичної змінної множини. Для дефазифікації можуть бути використані такі методи: центра тяжіння, центра тяжіння для одноточкових множин, центра площини, лівого та правого модальних значень, центр максимумів [2].

Для збільшення якості прогнозів необхідно використовувати механізм налаштування моделі.

Схема реалізації цієї методики виглядає так.

Алгоритм 1

Крок 1. Визначення універсальної множини U. Нехай $D_{\min}$ та $D_{\max}$ найменше та найбільше значення вихідних даних. З метою отримання більш зручних меж інтервалу треба вибрати додатні числа D_1 і D_2. Тоді універсальну множину U можливо представити у вигляді U : $U = \left[D_{\min} - D_1; D_{\max} + D_2\right]$.

Крок 2. Розіб'ємо універсальну множину на k рівних інтервалів.

Крок 3. Визначимо середні точки інтервалів u_{icp}.

Крок 4. Визначимо нечіткі множини універсальної множини U. Визначимо A_i ($i = 1,...,k$) лінгвістичні значення лінгвістичної змінної за формулою

$$\mu_{A_i}(u_i) = \frac{1}{1 + (c \cdot (v - u_{icp}))^2}, \qquad (1)$$

де v – варіація, c – стала величина, яка вибирається так, щоб забезпечити належність $\mu_{A_i}(u_i)$ проміжку [0;1].

Крок 5. Фаззифікація варіацій, обчислених на кроці 1. Якщо для кварталу i варіація буде v_i, v_i належить u_j, то для u_j (u_j належить U – інтервали універсальної множини U) функція приналежності $\mu_{A_i}(u_i)$ обчислюється за формулою (1) з урахуванням $v = v_i$, тобто з універсальної множини розглядається той інтервал, до якого потрапляє варіація, що розглядається.

Крок 6. Треба вибрати базис w ($1 \le w \le l$, де l – кількість попередніх кварталів, які включені у прогноз). З урахуванням базису, обчислюється матриця нечітких відносин $R^w(t)$, на основі якої робиться прогноз. З цією метою після вибору w будується операційна матриця $i \times j$ $O^w(t)$ (i – число рядків, що відповідають кількості кварталів у послідовності $t-2, t-3, ..., t-w$; j – число стовпчиків, що відповідає кількості інтервалів варіацій) та матриця-критерій $1 \times j$ $K(t)$ для кварталу, що прогнозується, t (матриця-рядок, що відповідає нечіткій варіації за квартал $t-1$).

Крок 7. Для дефаззифікації результатів, отриманих на 6-му кроці, пропонується така формула

$$v(t) = \frac{\sum \mu_t(u_i) \cdot u_{icp}}{\sum \mu_t(u_i)},$$

де $\mu_t(u_i)$ – обчислені значення функції приналежності для кварталу, для якого проводиться прогноз.

Отже, в результаті отримано значення прогнозу варіації величини на наступний квартал і для того щоб отримати значення прогнозу самої величини, необхідно до поточного значення величини додати значення спрогнозованої варіації [3].

Нечіткі методи дуже успішно застосовуються, отримані результати мають високий ступень інтерпретації. Але точність автоматично отриманих правил дуже низька, часто виникають ситуації, коли необхідно використовувати допомогу експертів. Для виправлення цього недоліку, можливо використовувати еволюційні методи, щоб будувати та налаштовувати набір нечітких правил. Питаннями побудови та налаштуванням нечітких правил займається напрямок генетичних нечітких систем (genetic fuzzy system). Дослідження в цьому напрямку є дуже важливою актуальною науковою задачею, під час розв'язку якої можна отримати важливі результати як у теоретичному так і практичному плані [4].

Розглянемо один з сучасних методів прогнозування НЧР з використанням генетичного алгоритму. Цей алгоритм може бути представлений наступним чином.

Крок 1. Визначення універсальної множини U. Нехай $D_{\min}$ та $D_{\max}$ найменше та найбільше значення вихідних даних. З метою отримання більш зручних меж інтервалу треба вибрати додатні числа D_1 і D_2. Тоді універсальну множину U можна представити у вигляді U : $U = [D_{\min} - D_1; D_{\max} + D_2]$.

Крок 2. розіб'ємо універсальну множину на 7 інтервалів: $[D_{\min} - D_1; x_1], [x_1; x_2], [x_2; x_3], [x_3; x_4], [x_4; x_5], [x_5; x_6], [x_6; D_{\max} + D_2]$,

де x_1, x_2, x_3, x_4, x_5, x_6 – цілі числа та $x_1 < x_2 < x_3 < x_4 < x_5 < x_6$.

Крок 3. Визначимо хромосому. Нехай кожна хромосома складається з 6 генів, як показано нижче

$$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6,$$

де x_1, x_2, x_3, x_4, x_5, x_6 – цілі числа.

Під час ініціалізації випадковим чином генерується кожний ген. Випадковим чином згенеруємо популяцію N хромосом.

Крок 4. Значення прогнозу може бути знайдено використовуючи крок 3 – крок 6 з Алгоритму 1.

Крок 5. Кроссовер. Випадковим чином виберемо дві хромосоми з популяції, щоб провести процедуру кроссоверу. Під час процедури кросоверу система випадково вибирає точку кроссоверу (ціле число від 1 до 5).

Хромосома 1 $x_1 \; x_2 \; x_3 \; x_4 \; x_5 \; x_6$

Хромосома 2 $y_1 \; y_2 \; y_3 \; y_4 \; y_5 \; y_6$

Після кроссоверу:

Хромосома 1 $x_1 \; y_2 \; y_3 \; y_4 \; y_5 \; y_6$

Хромосома 2 $y_1 \; x_2 \; x_3 \; x_4 \; x_5 \; x_6$

Крок 6. Мутація. Випадковим чином вибираємо хромосому з популяції та випадковим чином вибираємо номер гена у хромосомі. Після проведення операцій мутації та кроссоверу, для кожної хромосоми проводимо процедуру прогнозування. Коли отримано всі значення прогнозів, можна обчислити помилку прогнозу. Значення помилки буде використовуватися як значення фітнесу для відповідної хромосоми. Зрозуміло, чим менше помилка, тим краще пристосована хромосома.

Крок 7. Вибираємо M ($M < N$) хромосом із популяції з найменшими значеннями помилки, і ще ($N - M$) хромосом генеруємо випадковим чином. Аналогічним чином для нової популяції повторюємо кроки 5 – 7. Таким чином, треба породити P популяцій. Хромосома, яка буде мати найменше значення помилки і буде найкращим розв'язком задачі прогнозування НЧР [5].

Розглянуті алгоритми покладені до основи розробленого програмного продукту Fuzzy_Forecasting. Середовище розробки програмного продукту – Eclipse. Мова програмування – Java. Мінімальні вимоги до системи: операційна система Windows XP/7; Linux. Програмний продукт спроектовано за шаблоном Модель-Вигляд-Контролер (Model – View – Controller). На рис. 1 наведено UML діаграму класів, а на рис. 2 – діаграму варіантів використання.

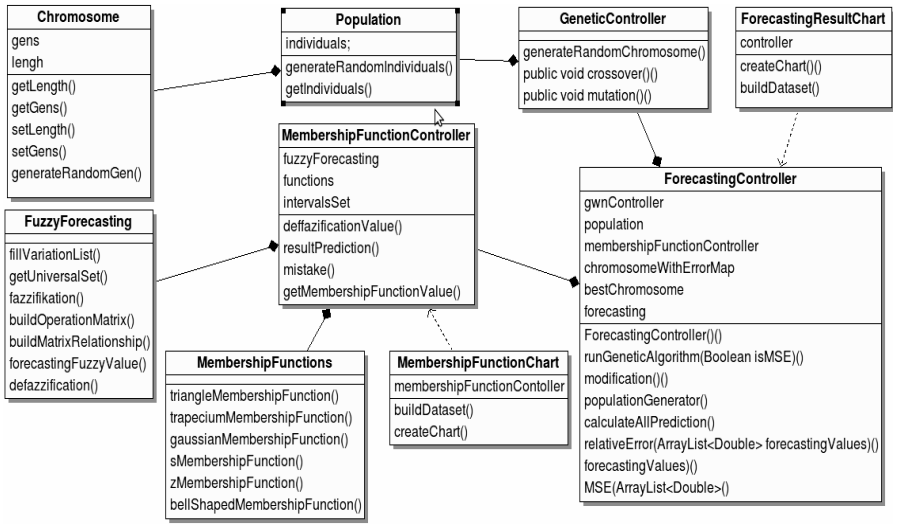


Рис. 1. UML діаграма класів

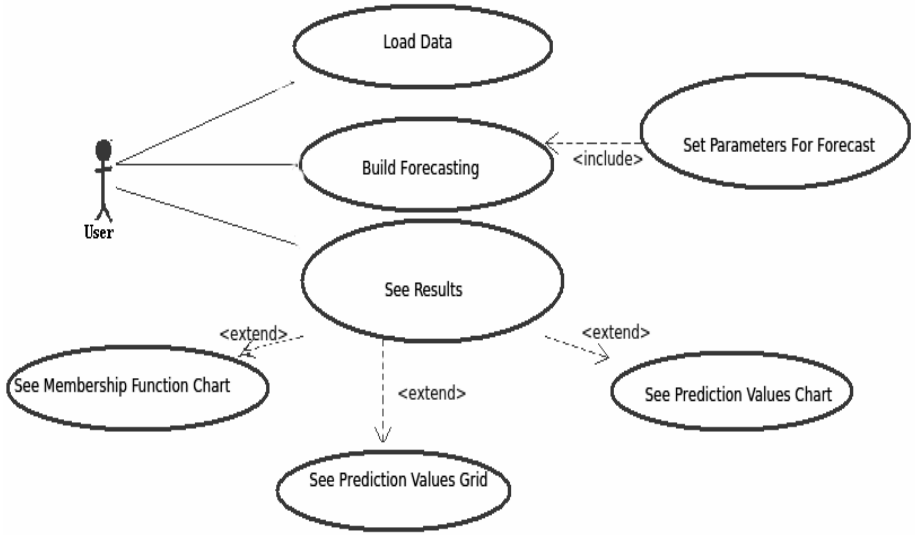


Рис. 2. UseCase діаграма

Розроблене застосування має дружній та зручний інтерфейс для користувачів, дані наочно представлені у вигляді таблиць та графіків.

Апробація розробленого програмного продукту виконана на даних гідрогеохімічного моніторингу території Північного гірничо-збагачувального комбінату (ГЗК) за період з лютого 1986 р. по травень 1992 р. На рис. 3 наведено графік функцій приналежності. На рис. 4 зображено графік прогнозу, побудованого за вище описаним алгоритмом. Історичні дані для прогнозу – це дані концентрації SO_2 у воді. У таблиці 1 наводяться обчислені значення середньоквадратичної та середньої відносної помилки для наборів даних SO_2, SO_4, SO, Na, S.

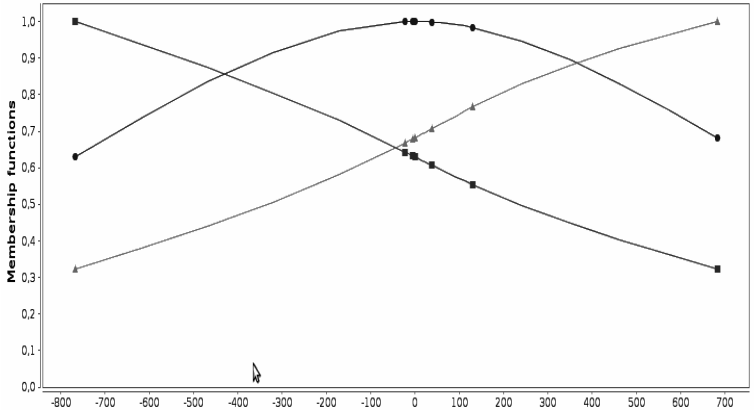


Рис. 3. Графіки функцій приналежності

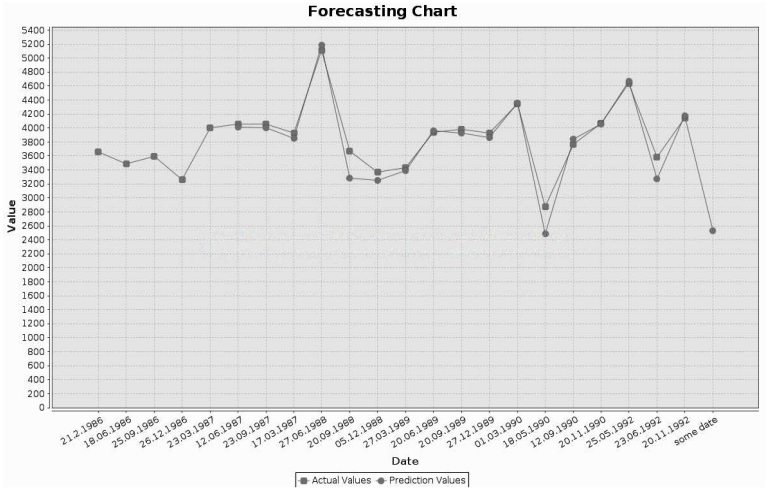


Рис. 4. Графік отриманих результатів

Таблиця 1

Обчислені похибки прогнозу

Дані	Середньоквадратична помилка	Середня відносна помилка (%)
SO_2	568231	2.77
SO_4	691091	2.82
SO	1924084	2.40
Na	176672	2.81
S	1682336	2.76

Висновки. У роботі розглянута інформаційна технологія для прогнозування нечітких часових рядів. Розглянута можливість застосування нечіткої логіки для аналізу та прогнозування часових рядів концентрації хімічних компонентів у пробах підземних вод на території гірничо-збагачувального комбінату. Подальші дослідження можуть бути спрямовані на вдосконалення розглянутого підходу для виявлення більш складних шаблонів у нечітких часових рядах.

Бібліографічні посилання

1. **Леоненков А.В.** Нечеткое моделирование в среде MATLAB и fuzzyTECH / А.В. Леоненков. – СПб., 2005. – 736 с.
2. **Мамедова М.Г.** Нечеткая логика в прогнозировании демографических аспектов рынка труда / М.Г. Мамедова, З.Г. Джабраилова // Сб. Трудов НИУЦ по труду и социальным проблемам. – 2005.
3. **Хмелевой С.В.** Инструментальные средства для создания базы знаний на основе нечетких продукций, настраиваемых с помощью генетических алгоритмов./ С.В. Хмелевой, Ю.А. Скобцов, А.М. Фонотов //Сучасні тенденції розвитку іформаційних технологій в науці, освіті та економіці: Матер. II Всеукр. науково-практ. конф. 8–10 квітня 2008 р. – Луганськ: Альма-матер. – 2008. – С.103–105.
4. **Shyi-Ming Chen.** Forecasting Enrollments of Students by Using Fuzzy Time Series and Genetic Algorithms / Shyi-Ming Chen, Nien-Yi Chung // Information and Management Sciences. – Vol. 17, № 3. 2006.
5. **Song Q.** Fuzzy time series and its models. / Q. Song, B.S. Chissom B.S. // Fuzzy Sets and Syst. – № 54, – 1993.
6. **Song Q.** Forecasting enrollments with fuzzy time series – part 1./ Q. Song, B.S. Chissom // Fuzzy Sets and Syst. – № 54, – 1993.
7. **Song Q.** Forecasting enrollments with fuzzy time series – part 2./ Q. Song, B.S. Chissom B.S. // Fuzzy Sets and Syst. – № 54, – 1993.
8. **Chen S.-M.** Forecasting Enrollments Based on Fuzzy Time Series // Fuzzy Sets and Syst. – 81.1996.
9. **Şah, M.** Forecasting Enrollment Model Based on First-Order Fuzzy Time Series. / Degtiarev, K.Y. // Proc. Int. Conf. Computational Intelligence. – 2004.

10. **Shyi-Ming** Chen. Forecasting Enrollments of Students by Using Fuzzy Time Series and Genetic Algorithms / Nien-Yi Chung // Information and Management Sciences. – Vol. 17. – Number 3. – 2006. – P. 1–7.
11. **Cheng, Ching-Hsue.** Predicting daily ozone concentration maxima using fuzzy time series based on a two-stage linguistic partition method / Ching-Hsue Cheng, Sue-Fen Huang, Hia-Jong Teoh // Computers & Mathematics with Applications. – Vol. 62, Issue 4. – 2011. – P. 2016–2028.
12. **Zounemat-Kermani.** Using adaptive neuro-fuzzy inference system for hydrological time series prediction / Mohammad Teshnehlab // Applied Soft Computing. – Vol. 8, Issue 2. – 2008. – P. 928–936.
13. **S. Marsili-Libelli.** Fuzzy pattern recognition of circadian cycles in ecosystems / S. Marsili-Libelli // Ecological Modelling. – Vol. 174, Issues 1–2. – 2004. – P. 67–84.
14. **Elisabetta Giusti.** Spatio-temporal dissolved oxygen dynamics in the Orbetello lagoon by fuzzy pattern recognition / Elisabetta Giusti, Stefano Marsili-Libelli // Ecological Modelling. – Vol. 220, Issue 19. – 2009. – P. 2415–2426.
15. **Alexandra P.** Development of rainfall–runoff models using Takagi–Sugeno fuzzy inference systems / Alexandra P. Jacquin, Asaad Y. Shamseldin // Journal of Hydrology. – Vol. 329, Issues 1–2. – 2006. – P. 154–173.

Надійшла до редколегії 09.07.2012

УДК 519.242

С.В. Земляна

Дніпропетровський національний університет ім. Олеся Гончара

ОБЧИСЛЮВАЛЬНІ СХЕМИ ПОСЛІДОВНОГО АНАЛІЗУ ПРИ КОНТРОЛІ НАДІЙНОСТІ З НЕПЕРЕРВНОЮ ВИБІРКОЮ

Розглядаються методи контролю на надійність високонадійних виробів, які працюють у кусково-стаціонарному режимі. Наведено обчислювальні процедури методу послідовного аналізу для вибіркової реалізації з неперервним часом на основі сплайн-експоненційного розподілу з одним вузлом залежно від взаємного розташування вузлів сплайна.

Ключові слова: *послідовний аналіз, вибіркова реалізація з неперервним часом.*

Рассматриваются методы контроля на надежность высоконадежных изделий, работающих в кусочно-стационарном режиме. Приведены вычислительные процедуры метода последовательного анализа для выборочной реализации с непрерывным временем на основе сплайн-экспоненциального распределения с одним узлом в зависимости от взаимного расположения узлов сплайна.

Ключевые слова: *последовательный анализ, выборочная реализация с непрерывным временем.*

In this paper we considered methods to control the reliability of highly reliable products that work in the piecewise stationary mode. We present computational method of sequential analysis of procedure for the sample realization the continuous-time based on spline-exponential distribution with a single node, depending on the relative position of nodes of the spline.

Key words: *sequential analysis, sample realization of continuous time.*

Постановка проблеми в загальному вигляді. Дана робота розглядає методи контролю на надійність високонадійних виробів, які працюють у кусково-стаціонарному режимі (це значить, що нестаціонарність викликана раптовими змінами в деякі моменти часу: які при цьому не завжди наперед відомі). Контролюються як самі моменти «розладу», так і параметри надійності на стаціонарних

дільницях при невідомих точних значеннях цих параметрів, а також меж інтервалів стаціонарності.

Під «високою надійністю» розуміють ситуацію, при якій під час проведення випробувань спостерігається мало відмов, або вони довгий час зовсім відсутні і тоді єдиною інформацією про надійність виробу є час неперервної безвідмовної роботи, що далі в тексті буде називатись «використанням неперервних реалізацій» протягом випробувань на надійність. Прикладами таких виробів можуть бути різні енергетичні установки, електроприводи систем автоматики, відмови яких призводять до катастрофічних наслідків, ракетна техніка, медична апаратура, вузли рухомого складу на залізничному транспорті та ін.

Аналіз останніх досягнень. Основний внесок у теорію планування випробувань вніс А. Вальд, який запропонував метод послідовного аналізу. В наступних роботах Дж. Кифера, Дж. Вольфовиця, Б. Марченко [1–3] ці методи набули подальшого розвитку. Запропоновано точні формули обчислень для вибіркової реалізації з неперервним часом на основі експоненційного закону розподілу часу напрацювання до відмови. Слід зазначити, що в даний час при вирішенні різних задач обробки статистичних даних знайшли широке застосування сплайн-розподіли [4], які найбільш адекватно і достовірно описують реальні процеси. Тому актуальним є використання сплайн-розподілів при розробці обчислювальних схем планування випробувань.

Мета роботи. Необхідно розробити обчислювальну технологію планування випробувань на основі сплайн-експоненційного розподілу з одним вузлом для вибіркової реалізації з неперервним часом залежно від взаємного розташування вузлів сплайна.

Основна частина. Припустимо, що щільність розподілу часу безвідмовної роботи об'єкта, який підлягає випробуванню на надійність, має сплайн-експоненційний закон з одним вузлом [4], тобто

$$p(t;\lambda_1,\lambda_2,T) = \begin{cases} \lambda_1 \exp(-\lambda_1) I_t(0,T] + \lambda_2 \exp(-\lambda_2 t - (\lambda_1 - \lambda_2)T) I_t(T,\infty), & t > 0; \\ 0, & t \le 0, \end{cases}$$

де $\lambda_1 > 0, \lambda_2 > 0, T > 0$ параметри, T - момент «розладу», а $I_t(a,b]$ та $I_t(a,b)$ індикатори:

$$I_t(a,b]=\begin{cases}1,\ t\in(a,b];\\ 0,\ t\notin(a,b).\end{cases}\quad I_t(a,b)=\begin{cases}1,\ t\in(a,b);\\ 0,t\notin(a,b).\end{cases}$$

У роботі [5] отримано загальний вигляд логарифму відношення правдоподібності. Процес відмов (кількість відмов, що виникли до моменту часу t) описується узагальненим процесом Пуассона з ведучою функцією

$$\mathrm{x(t)}=\Lambda(t)=-\left(\lambda_1 I_t(0,T]+\lambda_2 I_t(T,\infty)\right)t\ ,$$

гіпотези про стан об`єкта сформульовано у вигляді:

$\mathrm{H_0}$: $\lambda_1=\lambda_1',\lambda_2=\lambda_2',T=T'$; $p(t;\lambda_1',\lambda_2',T')$,

$\mathrm{H_1}$: $\lambda_1=\lambda_1'',\lambda_2=\lambda_2'',T=T''$; $p(t;\lambda_1'',\lambda_2'',T'')$.

Зараз розглянемо окремо часткові випадки різного взаємного розташування вузлів сплайна.

1) $T_0'<T_0''$

а) $t\in\left(0,T_0'\right]$

Введемо наступні позначення:

$$K_1=\frac{b}{\ln r_1},\ J_1=\frac{a_1}{\ln r_1},\ r_1=\frac{\lambda_1''}{\lambda_1'},\ c_1=\frac{\lambda_1''-\lambda_1'}{\ln r_1}.$$

Тоді

$$v_\lambda(s)=\begin{cases}0,\ s\le K_1;\\ 1,\ s\ge J_1,\end{cases}$$

а для $K_1<s<J_1$ при $s\ne J_1-1$, як показано в [2], $v_\lambda(s)$ задовольняє наступному рівнянню

$$cv_\lambda'(s)+\lambda v_\lambda(s)=\lambda v_\lambda(s+1). \qquad (1)$$

При виконанні умов

- $v_\lambda(s)$ неперервна для $\lambda<J_1$;
- $v_\lambda(K_1)=0$;
- $v_\lambda(s)=1$ для $s\ge J_1$,

рівняння (1) має наступний розв'язок

$$v_\lambda(s)=1-\exp\left(-\frac{\lambda}{c}s\right)\exp\left(\frac{\lambda}{c}K_1\right)\frac{\sum_{i=0}^{n(s)}\frac{(-1)^i}{i!}\left[\left(J_1-s-i\right)\frac{\lambda}{c}\exp\left(-\frac{\lambda}{c}\right)\right]^i}{\sum_{i=0}^{n(K_1)}\frac{(-1)^i}{i!}\left[\left(J_1-K_1-i\right)\frac{\lambda}{c}\exp\left(-\frac{\lambda}{c}\right)\right]^i},$$

де $n(s)=\left]J_1-s\right[$ – ціла частина від $\mathrm{J_1-s}$, тобто найбільше ціле число, що не перевищує $\mathrm{J_1-s}$.

Введемо наступне позначення

$$\Psi(x,y,d)=\sum_{m=0}^{]x[}\frac{(-1)^m}{m!}\left[(y-m)e^{-d}d\right]^m,$$

тоді

$$v_\lambda(s)=1-\exp\left(-\frac{\lambda}{c}s\right)\exp\left(\frac{\lambda}{c}K_1\right)\frac{\Psi(J_1-s,J_1-s,\lambda/c)}{\Psi(J_1-K_1,J_1-K_1,\lambda/c)}.$$

У даному випадку s = 0, тому

$$v_\lambda(0)=q\left(\frac{\lambda}{c}\right)=1-\exp\left(\frac{\lambda}{c}K_1\right)\frac{\Psi(J_1,J_1,\lambda/c)}{\Psi(J_1-K_1,J_1-K_1,\lambda/c)}. \qquad (2)$$

Тепер точні значення для a знаходяться з рівнянь:

$$q\left(\frac{\lambda_1'}{c_1}\right)=\alpha\ \text{ або }\ q\left(\frac{\lambda_1''}{c_1}\right)=1-\beta. \qquad (3)$$

Враховуючи, що

$$\frac{\lambda_1'}{c_1}=\frac{\lambda_1'\ln\frac{\lambda_1''}{\lambda_1'}}{\lambda_1''-\lambda_1'}=\frac{\ln\frac{\lambda_1''}{\lambda_1'}}{\frac{\lambda_1''}{\lambda_1'}-1}=\frac{\ln r_1}{r_1-1}, \qquad (4)$$

отримаємо з (2) з врахуванням одного з рівнянь (3) наступне співвідношення для визначення a_1

$$1-\alpha=\exp\left(\frac{b}{r_1-1}\right)\frac{\Psi\left(\frac{a_1}{\ln r_1},\frac{a_1}{\ln r_1},\frac{\ln r_1}{r_1-1}\right)}{\Psi\left(\frac{a_1-b}{\ln r_1},\frac{a_1-b}{\ln r_1},\frac{\ln r_1}{r_1-1}\right)}. \qquad (5')$$

б) $t\in\left(T',T''\right]$

Уведемо наступні позначення:

$$K_2 = \frac{b}{\ln r_2},\ J_2 = \frac{a_2}{\ln r_2},\ r_2 = \frac{\lambda_1''}{\lambda_2'},\ c_2 = \frac{\lambda_1'' - \lambda_2'}{\ln r_2},$$

$$n(s) =]J_2 - s[,\ R(0) = d_2,\ d_2 = -\frac{\lambda_1' - \lambda_2'}{\ln r_2} T_0';$$

$$v_\lambda(d_2) = q\left(\frac{\lambda}{c_2}\right) = 1 - \exp\left\{-\frac{\lambda}{c_2}(d_2 - K_2)\right\} \frac{\Psi(J_2 - d_2, J_2 - d_2, \lambda/c_2)}{\Psi(J_2 - K_2, J_2 - K_2, \lambda/c_2)}.$$

Тоді для визначення a_2 отримаємо наступне співвідношення

$$1 - \alpha = \exp\left\{\frac{1}{r_2 - 1}(b - d_2 \ln r_2)\right\} \frac{\Psi\left(\frac{a_2}{\ln r_2} - d_2, \frac{a_2}{\ln r_2} - d_2, \frac{\ln r_2}{r_2 - 1}\right)}{\Psi\left(\frac{a_2 - b}{\ln r_2}, \frac{a_2 - b}{\ln r_2}, \frac{\ln r_2}{r_2 - 1}\right)}. \quad (5'')$$

в) $t \in \left(T_0'', \infty\right)$

Введемо наступні позначення:

$$K_3 = \frac{b}{\ln r_3},\ J_3 = \frac{a_3}{\ln r_3},\ r_3 = \frac{\lambda_2''}{\lambda_2'},\ c_3 = \frac{\lambda_2'' - \lambda_2'}{\ln r_3},$$

$$n(s) = [J_3 - s],\ R(0) = d_3,\ d_3 = \frac{\lambda_1' - \lambda_2'}{\ln r_3} T_0' - \frac{\lambda_1'' - \lambda_2''}{\ln r_3} T_0'';$$

$$v_\lambda(d_3) = q\left(\frac{\lambda}{c_3}\right) = 1 - \exp\left\{-\frac{\lambda}{c_3}(d_3 - K_3)\right\} \frac{\Psi(J_3 - d_3, J_3 - d_3, \lambda/c_3)}{\Psi(J_3 - K_3, J_3 - K_3, \lambda/c_3)}.$$

Тоді для визначення a_3 отримаємо наступне співвідношення

$$1 - \alpha = \exp\left\{\frac{1}{r_3 - 1}(b - d_3 \ln r_3)\right\} \frac{\Psi\left(\frac{a_3}{\ln r_3} - d_3, \frac{a_3}{\ln r_3} - d_3, \frac{\ln r_3}{r_3 - 1}\right)}{\Psi\left(\frac{a_3 - b}{\ln r_3}, \frac{a_3 - b}{\ln r_3}, \frac{\ln r_3}{r_3 - 1}\right)}. \quad (5''')$$

Випадки 2), 3) розглядаються аналогічно.

З співвідношень (5′), (5″), (5‴) за допомогою відомих методів рішення рівнянь, наприклад, методом половинного ділення, по заданим α, β та r_1, r_2, r_3, d_2, d_3 можна знаходити відповідні значення a_1, a_2, a_3.

Область продовження випробувань у випадку 1) задається наступними нерівностями:

а) $t \in \left(0, T_0'\right]$

$$K_1 + c_1 t < x(t) < J_1 + c_1 t;$$

б) $t \in \left(T_0', T_0''\right]$

$$K_2 + c_2 t - d_2 < x(t) < J_2 + c_2 t - d_2;$$

в) $t \in \left(T_0'', \infty\right)$

$$K_3 + c_3 t - d_3 < x(t) < J_3 + c_3 t - d_3.$$

Приведемо формулу для визначення середньої кількості спостережень для прийняття остаточного рішення.

Позначимо середню кількість спостережень, коли істинне значення параметра розподілу дорівнює λ та $R(0) = s$, через $M_\lambda\left(t/\lambda', r, s\right)$. Введемо позначення

$$\Psi_1(x, y, d) = \sum_{m=0}^{]x[-1} e^{-d\,i} \sum_{k=0}^{m} \frac{(-1)^k}{k!} \left[(y - m - 1)d\right]^k.$$

Тоді, згідно з [4]

$$M_\lambda\left(t/\lambda', r, s\right) = \frac{n(s)+1}{\lambda} + c' \exp\left(-\frac{\lambda}{c} s\right) \Psi(J - s, J - s, \lambda/c) -$$

$$-\frac{1}{\lambda} \exp\left(\frac{\lambda}{c}(J - s - 1)\right) \Psi_1(J - s, J - s, \lambda/c)$$

при $K \le s \le J$, де постійна c′ визначається з умови $M_\lambda\left(t/\lambda', r, K\right) = 0$ та дорівнює

$$c' = \frac{1}{\lambda} \exp\left\{\frac{\lambda}{c}(J - 1)\right\} \frac{\Psi_1(J - K, J - K, \lambda/c) - n(K) - 1}{\Psi(J - K, J - K, \lambda/c)}.$$

Тоді середня тривалість спостережень до прийняття рішення у випадку сплайн-експоненційного розподілу з одним вузлом дорівнює

$$M_\lambda\left(t/\lambda', r\right) = M_\lambda\left(t/\lambda', r_1, 0\right) v_\lambda(0) + M_\lambda\left(t/\lambda', r_2, d_2\right) v_\lambda(d_2) +$$

$$+ M_\lambda\left(t/\lambda', r_3, d_3\right) v_\lambda(d_3).$$

Оперативна характеристика, згідно з [4], має вигляд

$$P_\lambda(H_0) = 1 - q\left(\frac{\lambda}{c}\right).$$

Проведемо порівняльний аналіз меж послідовного аналізу при реалізації моделей з дискретною вибіркою та з неперервною для випадку сплайн-експоненційного розподілу з одним вузлом часу безвідмовної роботи високо надійних виробів спеціального призначення. Для цього розглянемо наступний приклад.

Приклад. Треба побудувати (розрахувати) план проведення випробувань високонадійних виробів на надійність шляхом перевірки наступних гіпотез :

H_0: $\lambda_1' = 1.2$; $\lambda_2' = 0.6$; $T' = 1.2$ року;

H_1: $\lambda_1'' = 3.07$; $\lambda_2'' = 1.7$; $T'' = 1.2$ року.

Ймовірності похибок 1-го та 2-го роду обрано наступні: $\alpha = 0.1$; $\beta = 0.1$.

У результаті розрахунків за наведеними вище формулами отримуємо:

$a = \ln\frac{1-\beta}{\alpha} = 2.197$; $b = \ln\frac{\beta}{1-\alpha} = -2.197$; $r_1 = 2.558$; $r_2 = 2.833$;

$c_1 = 1.991$; $c_2 = 0.0585$; $d_2 = -0.887$; $K_1 = -2.339$; $K_2 = 2.11$;

$a_1 = 1.881$; $a_2 = 0.756$; $J_1 = 2.002$; $J_2 = 0.726$.

Дослідження проводились на протязі 2,5 років шляхом неперервних спостережень. Масив часу відмов (реалізація відмов, числа в роках): {0.059; 0.165; 0.2086; 0.336; 0.644; 1.95; 2.19}.

Результати послідовного аналізу представлено на рисунку.

[В, АД] – область продовження випробувань у випадку дискретної вибіркової реалізації; [В, АН] – область продовження випробувань у випадку неперервної вибіркової реалізації; x(t) – процес відмов.

З аналізу рис. витікає, що область продовження випробувань у випадку неперервної вибіркової реалізації менша, ніж та ж область у випадку дискретної вибіркової реалізації. Це призводить до зменшення кількості випробувань за планом з неперервною вибірковою реалізацією у порівнянні з планом випробувань з дискретною вибіркою. План з неперервною вибіркою після виникнення третьої відмови призводить до відхилення основної гіпотези. План з дискретною вибірковою реалізацією після четвертої відмови призводить до відхилення основної гіпотези.

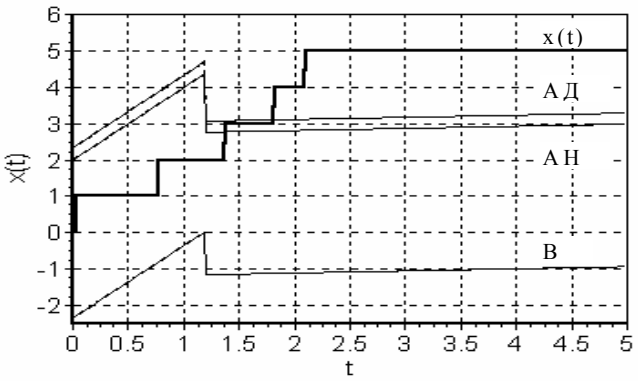


Рис. Результати послідовного аналізу

Висновки та перспективи подальшого розвитку. Рекомендується віддавати перевагу використанню методу послідовного аналізу з неперервною вибіркою, тому що він базується на точних порогових значеннях меж для аналізу процесу правдоподібності, крім того, цей метод потребує у середньому меншу кількість випробувань. Надалі представляється доцільним розробити більш прості для реалізації на ЕОМ обчислювальні схеми розрахунку границь випробувань.

Бібліографічні посилання

1. **Dvoretzky A., Kiefer J., Wolfowitz J.** Sequential decision problems for process with continuosis time parameter. Testing Hypotheses. // The Annals of Mathematical Statistics, June. – 1953. – 24, 2. – P. 254–264.

2. **Kiefer J., Wolfowitz J.** Sequential Tests of Hypothes about the mean occurense time of continuous parameter Poisson process // Naval research Logistics quartaly – 3, 3 (1956). – P. 205–219.

3. **Броди С.М.** Расчет и планирование испытаний систем на надежность. / С.М. Броди, О.Н. Власенко, Б.Г. Марченко. – К., – 1970. – 192 с.

4. **Приставка А.Ф.** Сплайн-распределения в статистическом анализе / А.Ф. Приставка – Д., – 1995. – 152 с.

5. **Земляна С.В.** Застосування послідовного аналізу при контролі надійності в нестаціонарних режимах: загальні положення / С.В. Земляна // Актуальні проблеми автоматизації та інформаційних технологій. – Д., 2011. – Т.15. – С.195–200.

Надійшла до редколегії 11.06.2012

УДК 534.4: 621.391

О.М. Карпов, К. Г. Чебров

Дніпропетровський національній університет ім. Олеся Гончара

РОЗРОБКА АЛГОРИТМІВ І ПРОГРАМ СИНТЕЗУ ЯКІСНОГО МОВЛЕННЯ ЗА ПРАВИЛАМИ

Наведено опис практичної реалізації алгоритму синтезу мовленевих сигналів.

Ключові слова: *алгоритм, синтез мовлення, формантний метод.*

Приведено описание практической реализации алгоритма синтеза речевых сигналов.

Ключевые слова: *алгоритм, синтез речи, формантный метод.*

In the article were the description of the practical realization of the algorithm of the synthesis of the speech signal considered.

Key words: *algorithm, Speech synthesis, formant synthesis method.*

Вступ. Усі способи синтезу мови можна поділити на групи [1– 4].

1. Параметричний синтез мови є кінцевою операцією у вокодерних системах, де мовний сигнал представляється набором невеликого числа параметрів, що безупинно змінюються. Параметричний синтез доцільно застосовувати в тих випадках, коли набір повідомлень обмежений і змінюється не надто часто. Перевагою такого способу є можливість записати мову для будь-якої мови і будь-якого диктора.

Якість параметричного синтезу може бути дуже високою (залежно від ступеня стиску інформації за параметричним поданням). Однак параметричний синтез не може застосовуватися для довільних, заздалегідь не заданих повідомлень.

2. Компіляційний синтез зводиться до складання повідомлення з попередньо записаного словника вихідних елементів синтезу. Очевидно, що зміст синтезованих повідомлень фіксується обсягом словника. Основна проблема в компілятивному синтезі – обсяги пам'яті для зберігання словника. У цьому зв'язку використовуються різноманітні методи стиску/кодування мовного сигналу.

3. Повний синтез мови за правилами (або синтез за друкованим текстом) забезпечує керування всіма параметрами мовного сигналу і, таким чином, може генерувати мову по заздалегідь невідомому тексту.

У цьому випадку параметри, отримані при аналізі мовного сигналу, зберігаються у пам’яті так само, як і правила з’єднання звуків у слова і фрази. Синтез реалізується шляхом моделювання акустичних процесів мовного тракту, застосуванням аналогової або цифрової техніки. Причому в процесі синтезування значення параметрів і правила з'єднання фонем вводять послідовно через певний часовий інтервал, наприклад 5–10 мс.

Виділяються два підходи. Перший – спрямований на побудову моделі мовної системи людини, він відомий під назвою артикуляторний синтез. Другий підхід – формантний синтез за правилами. Розбірливість і натуральність таких синтезаторів може бути доведена до величин, порівняних з характеристиками природної мови.

Загальна частина. Реальна побудова синтезатора утримує деяку сукупність правил усіх типів синтезаторів з обліком тривалості частоти, інтенсивності і варіативності (не стаціонарності) параметричних представлень мовних одиниць.

Спектр мови Y(ω) можна зобразити як добуток частотної функції джерела (джерела збудження) X(ω) та перехідної функції мовного тракту H(ω) (1):

$$\mathbf{Y(\omega) = X(\omega)\ H(\omega)} \qquad (1)$$

За часовою залежністю мовний сигнал буде представлятиметься відповідною згорткою (5.2) (для нерекурсивного фільтра):

$$y(t) = \int_0^{\infty} x(t)h(t-\tau)d\tau \qquad (2)$$

Для одержання необхідної F-картини фонеми може бути використаним банк фільтрів. Для одержання спектру фонеми /А/, яка була вимовлена автором даної роботи в останньому ненаголошеному складі слова «дама», який зображено на рис. 1, може використовуватися банк з п’яти фільтрів (відповідно до кількості формант).

Нехай зображений спектр фонеми має формуватись на базі спектра джерела збудження, котрий має вигляд як на рис. 2.

Для цього формують банк фільтрів, параметрами котрих задають вид формант, приведених на рис. 3.

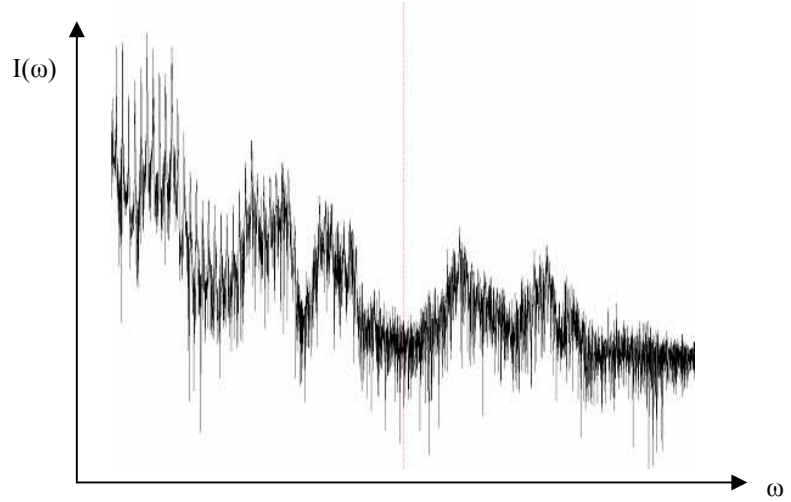


Рис. 1. Спектр фонеми /А/

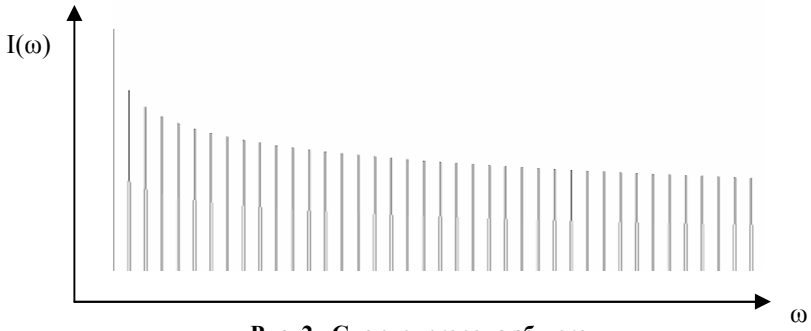


Рис. 2. Спектр джерела збудження

Рис. 3. Вид спектру, котрий породжується передатною функцією мовного тракту у вказаному випадку

Перейдемо до опису організації обчислювального процесу на базі наведених викладок. Програма має містити гнучкий інтерфейс. Вона реалізує синтез мовного сигналу за текстом на основі методу форматного синтезу з використанням можливостей компіляційного.
Вхідними даними є текстові файли, що містять інформацію про траекторії зміни F-картини фонем, траєкторії інтенсивностей, значення основних параметрів, які можуть варіюватись, wav-файли, базовий для синтезу текст. Вихідними даними є синтезований (відновлений за текстом) звук, котрий зберігається у wav-файл.

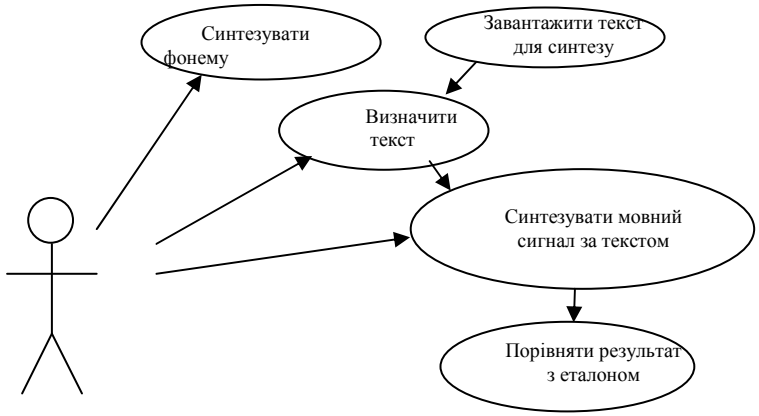


Рис. 4. Use-case діаграма програми

Программу було реалізовано засобами середовища MATLAB R2010a. До програми входять такі модулі:
amplMod.m – утримує службові функції для реалізації амплітудної модуляції сигналу;
applyFilter.m – слугує для таких задач як генерація сигналу голосового джерела збудження, побудова результуючого спектру вокалізованої фонеми за заданими параметрами;
fricative.m – слугує для таких задач як генерація сигналу шумового джерела збудження, побудова результуючого спектру фрікативних фонеми за заданими параметрами;
gui.m – головний модуль, котрий збирає до сукупності та організовує взаємодію функціональних модулів з користувачем з програмою;

nValsSlide.m – містить допоміжні функції для забезпечення неперервного (не скачковидного) способу зміни параметрів процесу синтезу;
Synth.m – головний модуль, котрий поєднує функціональні підмодулі синтезатора;
singleFilter.m – містить відносно низькорівневі функції для завдання однієї форманти.
Можливості програми та шлях взаємодії користувача з нею можна представити на рис. 4 за допомогою use-case діаграми.
Продемонструємо роботу програми. Згенеруємо слово «антивирусные». При цьому програма має вигляд як на рис. 5.
На рис. 6 и рис.7 наведено вигляд вікон з внутрішньою проміжною інформацією у графічному вигляді, що використовується при синтезі вище згаданого слова – траєкторії зміни F-картини та інтенсивності відповідно.

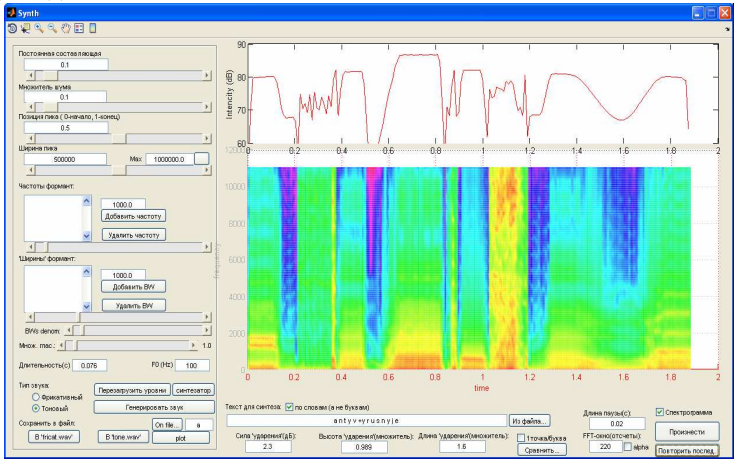


Рис. 5. Вигляд головного вікна програми після синтезу тексту

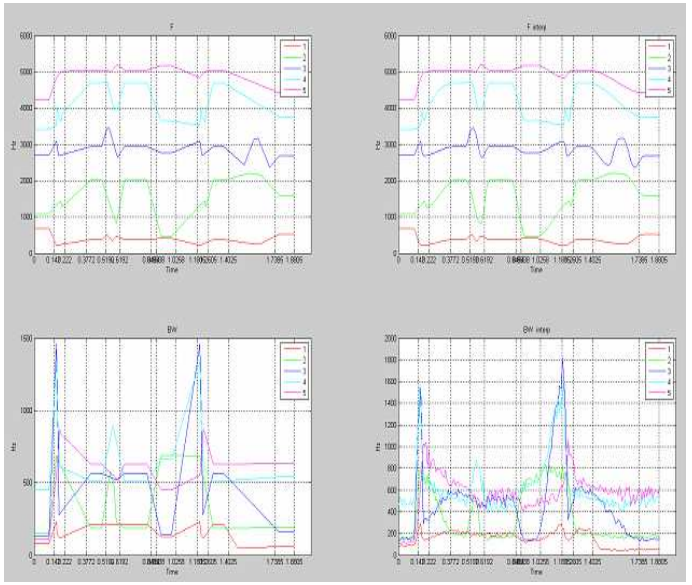


Рис. 6. Форматні траєкторії

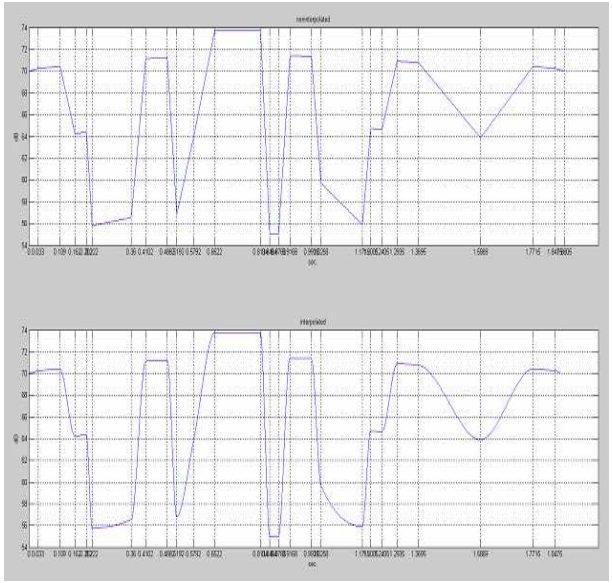


Рис. 7. Траєкторії інтенсивності

Під час проведення тестів програма показала коректну поведінку, відсутність помилок часу виконання.

Висновки. Зроблено програмний продукт для синтезу мовного сигналу з використанням методів формантного та компіляційного синтезу. Ця розробка відноситься до області надання автоматичним/автоматизованим системам мовного інтерфейсу за напрямком «машина–людина». Для реалізації поставленої задачі використовувалось середовище MATLAB R2010a.

Було зроблено огляд досягнень у галузі синтезу мовного сигналу, який показав, що провідні організації у всьому світі цікавляться та займаються дослідженнями в цьому напрямку. Але результати роботи існуючих систем не є повною мірою задовільними. Це обумовлено складністю процесу синтезу, зокрема складністю реалізації відображення перехідних процесів, нелінійностей та інших.

Проаналізувавши результати роботи програми, можна сказати, що, звичайно, якість синтезованої мови не дозволяє казати про передові показники, з чого слідує необхідність подальшого розвитку системи та використання додаткових методів, залучення механізмів артикуляційного синтезу. Під час проведення тестів програма показала коректну поведінку, відсутність помилок часу виконання.

Бібліографічні посилання

1. **Фант Г.** Акустическая теория речеробразования / Г. Фант. – М., 1964 – 284 с.

2. **Фланаган Д.** Анализ, синтез и восприятие речи / Д. Фланаган. – М., 1968 – 296 с.

3. **Сорокин В.Н.** Синтез речи / В.Н. Сорокин. – М., 1992. – 392 с.

4. **Лобанов Б.М.** Компьютерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник. – Минск, 2008. – 339 с.

Надійшла до редколегії 15.06.2012

УДК 519.688:336.76.066

О.П. Луценко, О.Г. Байбуз

Дніпропетровський національний університет ім. Олеся Гончара

ОГЛЯД МЕТОДІВ ПОШУКУ РОЗЛАДНАНЬ І ПЕРСПЕКТИВИ ЇХНЬОГО ЗАСТОСУВАННЯ У ТЕХНІЧНОМУ АНАЛІЗІ БІРЖОВИХ КОТИРУВАНЬ

Запропоновано використовувати методи визначення розладнань для визначення ділянок стаціонарності часових рядів котирувань при здійсненні технічного аналізу біржових ринків. Наведено огляд існуючих методів послідовного визначення розладнань.

Ключові слова*: статистика, визначення розладнань, біржова торгівля, котирування, часовий ряд, нестаціонарні процеси.*

Предложено использовать методы определения разладок для определения участков стационарности временных рядов котировок при проведении технического анализа биржевого рынка. Приведен обзор существующих методов последовательного определения разладок.

Ключевые слова: *статистика, определение разладок, биржевая торговля, котировки, временной ряд, нестационарные процессы.*

Proposed using change-point detection methods to determine areas of stationarity in time series of stock market quotations. The review of existing sequential methods of change-point detection provided.

Key words: *statistics, change-point detection, stock market, market quotations, time series, non-stationary process.*

Постановка проблеми. Торгівля на спекулятивному ринку фінансів і цінних паперів характеризується високою динамікою і великою потенційною прибутковістю, але поєднується з пропорційно великими ризиками, для мінімізації яких застосовується цілий комплекс засобів технічного аналізу. У зв'язку з ростом популярності даного виду торгівлі на протязі останніх років, у тому числі і на території України, попит на технології оцінки і мінімізації торгових ризиків надзвичайно зріс. Існуючі методи технічного аналізу перестають задовольняти ринкових гравців, і розробка нових напрямів аналізу ринку є актуальною задачею.

Аналіз досліджень у даній галузі. Тенденція до переходу від індикаторів і осциляторів технічного аналізу до більш точних і складних методів сформувалася відносно нещодавно. Це обумовлюється досягненнями персональними ЕОМ необхідного рівня обчислювальних можливостей. Популярності набули технології побудови апроксимаційних рівнянь кривих за допомогою генетичних алгоритмів, технології розпізнавання візуальних образів, технології нейромережевої побудови і оптимізації індикаторів.

Підхід, що пропонується авторами, відрізняється від існуючих тим, що часовий ряд валютних котирувань розглядається не як множина окремих точок, а як сукупність стаціонарних відрізків, на які він розбивається точками розладнань.

Знаходження точок розладки ряду дозволяє досягти двох цілей:

1) розділити ряд на ділянки з подібними статистичними властивостями;

2) отримати з великих вибірок даних стислу інформацію про статистичну динаміку ряду.

Отримані дані можуть бути використані у подальшому дослідженні закономірностей руху ринку за допомогою нейромережевого підходу.

Постановка цілей. Метою даної статті є викладення огляду методів пошуку розладнань, які слугують одним з основних компонентів експертної системи оцінки функцій ризику при здійсненні торгівлі на спекулятивному фінансовому ринку.

Виявлення зміни ринкової тенденції належить до задач послідовного виявлення розладнань. Апостеріорні методи можуть бути використані в тому випадку, якщо потрібні точні дані про час порушеня стаціонарності, і можуть бути застосовані лише за наявності певної кількості даних після можливого розладнання. Так як задачею аналізу ринку можливість роботи з найновішими даними є важливішою за максимізацію точності виявлення моменту розладнання у часі, в рамках даного дослідження розглядаються лише методи послідовного аналізу.

Основний матеріал. З математичної точки зору, загальна постановка задачі виявлення розладнання стаціонарного часового ряду $x_t = \{x_1, x_2, ..., x_n\}$ полягає у перевірці гіпотези H_0 про те, що випадкові величини x мають один і той же розподіл F_0 з деякої множини розподілів. Альтернативною є гіпотеза H_1 про кускову стаціонарність, тобто про існування такого моменту часу $\tau \geq 1$, що при $1 < \tau$ розподілом випадкових величин x_t є F_0, а при $t \geq \tau$ відмінний від F_0 деякий розподіл F_1.

Розглядаючи послідовні методи виявлення розладнання, можна говорити про кілька великих груп методів, заснованих на загальних підходах.

1 Алгоритм Гіршика-Рубіна-Ширяєва (ГРШ)

Уперше задача послідовного виявлення розладнання була поставлена в [1]. Розглядався випадок поточного контролю виробничого процесу, який може знаходитися в двох станах – налагодженому і розладженому – і має відомі величини щільності ймовірності до і після розладнання.

Нехай за першим станом щільність ймовірності значень ряду дорівнює $\omega(x_t, \theta_0)$, а в другому (розладнаному) – $\omega(x_t, \theta_0)$. Тоді

$$w = \frac{\omega(x_t, \theta_0)}{\omega(x_t, \theta_1)}.$$

На кожному кроці накопичується добуток

$$W_t = w_t(1 + W_{t-1})$$
$$W_0 = 0$$

Правило подачі сигналу про розладнання має вигляд:

$$\tau = \inf(t : W_t \geq b),$$

де b – поріг чутливості виявлення.

2 Алгоритми, засновані на накопиченні кумулятивних сум. Алгоритм кумулятивних сум (АКС, CUSUM) був розроблений Е. С. Пейджем [2]. Він являє собою послідовний аналіз А. Вальда, а точніше — послідовний критерій відношення ймовірності (ПКВЙ) для двох простих гіпотез H_1 (немає розладнання): $\theta = \theta_1$ і H_2 (є розладнання): $\theta = \theta_2$, де θ – деякий скалярний параметр щільності розподілу $\omega(x_t/\theta)$.

Ідея Е. С. Пейджа полягає в аналізі поведінки кумулятивної суми

$$S_t = S_{t-1} + \ln(\omega(x_t/\theta_2)\omega(x_t/\theta_1)).$$

На кожному кроці сума порівнюється із заданим порогом h. Якщо на кроці t $g_t > h$, то подається сигнал про розладнання, а накопичення суми починається заново з нуля. Таким чином:

$$S_t = \max(0, S_{t-1} + g_t),$$
$$g_t = \ln(\omega(x_t/\theta_2)/\omega(x_t/\theta_2)).$$

Сигнал про розладнання подається у момент часу

$$\tau = \inf(t \geq 1 : S_t > h). \quad (1)$$

Існує інше тлумачення АКС. На кожному кроці із заданим порогом порівнюється різниця

$$g_t = S_t - \min_{k<t} S_k . \quad (2)$$

Указані формули справедливі для випадку, коли середня величина θ збільшується. У випадку, якщо необхідно виявляти зміни θ у бік зменшення, різниця має вигляд

$$g_t = \max_{k<t} S_k - S_t . \quad (3)$$

Сигнал про розладнання подається у момент часу $\tau = \inf(t \geq 1 : g_t > h)$. Порівняння сум (2) і (3) з контрольною межею h може проводитися одночасно, щоб виявляти відхилення у будь-який бік.

Знаючи тип розподілу та визначаючи одну з імовірнісних характеристик розподілу як параметр θ, можна отримати рекурентні формули накопичення кумулятивної суми. Так, для випадку виявлення зміни середнього значення нормального розподілу формула для прирощення матиме вигляд

$$g_t = \frac{m_2 - m_1}{\sigma^2}\left(x - \frac{m_1 + m_2}{2} \right), \quad (4)$$

де m_1 – математичне очікування величини до розладнання; m_2 – передбачуване математичне очікування величини після розладнання; σ – середнє квадратичне відхилення; x – значення спостереження у момент часу t.

На підставі послідовного аналізу Вальда Г. Лорденом була розроблена процедура «максимальної правдоподібності» [3].

Нехай розподіл x відноситься до експоненціального сімейства розподілів з щільністю

$$\omega(x_t \mid \theta) = \exp(\theta T(x) - b(\theta)),$$

$b(\theta)$ – строго вигнута вгору функція, що диференціюється на всієї області визначення. Для прийнятого сімейства розподілів можна припустити, що при $\theta = 0$ $b(\theta) = 0$, за необхідності центруючи вибірку відносно середнього значення. Логарифм відношення правдоподібності гіпотези про присутність розладнання проти гіпотези про стаціонарність ряду рівний $\theta S_n - nb(\theta)$. Правило максимальної правдоподібності має вигляд:

$$\tau = \inf\left\{ t \geq 1 : \sup_{k \leq t} \sum_{i=k}^{t} T(x_i) - (t-k+1)b(\theta) \geq h \right\}.$$

Правило двобічного виявлення можна представити у формі U-маски: на кожному кроці накопичуючи суму $S_k^t = \sum_{i=k}^{t} T(x_i)$, можна порівнювати її з криволінійними порогами

$$\overline{c}_l = \inf_{\theta > \tilde{\theta}} \{ h/\theta + lb(\theta)/\theta \}$$

$$\underline{c}_l = \sup_{\theta < -\tilde{\theta}} \{ h/\theta + lb(\theta)/\theta \},$$

де $\tilde{\theta}$ – величина допускового інтервалу відхилення навколо θ.

На відміну від АКС Пейджа, ефективність якого падає при відхиленні від передбачуваного значення θ_2, алгоритм Лордена рівномірно ефективний для деякої множини значень θ_2. Недоліком методу є складність отримання рекурентного запису, що приводить до великої ресурсоємності рішення за допомогою ЕОМ.

3 Методи, засновані на лемі Неймана-Пірсона. Дана група методів заснована на перевірці гіпотези $\theta = \theta_1$ проти $\theta = \theta_2$, що виконується на кожному кроці для допоміжної вибірки обсягом $\tilde{N} : \{x_t^{t+\tilde{N}-1}\}$ згідно критерію максимальної правдоподібності. Для цієї вибірки обчислюється кумулятивна сума і порівнюється з порогом h. У разі нормального розподілу

$$S_{\tilde{N}}^t = ((\theta_2 - \theta_1)/\sigma^2)\left(\sum_{i=t}^{t+\tilde{N}-1} x_i - \tilde{N}(\theta_2 + \theta_1)/2 \right) . \quad (2)$$

Для даного випадку правило виявлення є еквівалентним використанню карт Шухарта.

Карта Шухарта [4] має дві контрольні межі щодо центральної лінії, які проводяться на відстані kσ від деякого еталонного значення (наприклад, середнє значення ряду), де σ – дисперсія випадкової величини, k – деяка контрольна межа (найчастіше використовується значення k=3, запропоноване самим Шухартом). Сигнал про розладнання подається, якщо:

$$x_{\tilde{N}} > m + k\sigma,$$

$$x_{\tilde{N}} < m - k\sigma,$$

$$x_{\tilde{N}} = \frac{1}{\tilde{N}} \sum_{i=t}^{t+\tilde{N}-1} x_i,$$

де $\tilde{N}$ – обсяг допоміжної виборки, m – математичне очікування.

4 *Алгоритми, засновані на експоненціальному згладжуванні.* Метод, заснований на експоненціальному згладжуванні, описаний в [5].

На кожному кроці спостереження накопичується сума

$$S_t = (1-k)S_{t-1} + k(x_t - m),$$

де m – математичне очікування.

Замість x може бути використане середнє значення $x_{\widetilde{N}}$, отримане з допоміжної вибірки. Також можливий варіант застосування алгоритму, коли величина k залежить від попередніх значень S_t, і таким чином здійснюється пристосування до зміни статистичних властивостей послідовності.

У разі збільшення середнього сигнал про розладнання подається згідно правила

$$\tau = \inf(t \geq 1 : S_t > h).$$

Правило для двобічної процедури:

$$\tau = \inf(t \geq 1 : | S_t | > h),$$

Існує модифікація даного методу, в якій накопичуються дві суми [6]:

$$S_t = (1-k)S_{t-1} + k(x_t - m),$$

$$R_t = (1-k)R_{t-1} + k | x_t - m |,$$

і на підставі їх обчислюється параметр

$$G(n) = \frac{S(n)}{R(n)}.$$

Сигнал про розладнання подається, якщо $G \geq h_2$ або $G \leq h_1$.

$$-1 < h_1 < h_2 < 1.$$

5 Байесівський підхід до послідовного виявлення розладнання. Задача послідовного виявлення розладнання може бути вирішена і з використанням *байесівського підходу*. Хоча байесівські формули ймовірності частіше використовуються в апостеріорному виявленні, у [7] був описаний метод послідовного визначення розладнання з прогнозуванням, заснований на рекурсивній байесівській оцінці очікування довжини ділянки без розладнання.

В основі алгоритму – твердження, що на кожному кроці спостереження можливі дві події: або розладнання не спостерігається, і отже збільшується довжина відрізку без розладнання, або відбувається розладнання, і відлік відрізку починається заново. Ймовірність цих подій:

$$P(r_t \mid r_{t-1}) = \begin{cases} H(r_{t-1}+1) & r_t = 0 \\ 1 - H(r_{t-1}+1) & r_t = r_{t-1}+1 \end{cases},$$

$$H = \frac{P_{gap}(g = \tau)}{\sum_{t=\tau}^{\infty} P_{gap}(g = t)},$$

де P_{gap} – апріорний дискретний розподіл ймовірностей на інтервалі без розладнання.

Досліджуваний ряд представляється у вигляді кусково-експоненціальної моделі. Функція правдоподібності для цієї моделі матиме вигляд:

$$P(x \mid \eta) = h(x)\exp(\eta^T U(x) - A(\eta)),$$

де

$$A(\eta) = \log \int d\eta\, h(x)\exp(\eta^T U(x)).$$

Вираз може бути переписаний у вигляді експоненціального розподілу навколо η

$$P(\eta \mid \chi, \nu) = \widetilde{h}(\eta)\exp(\eta^T \chi - \nu A(\eta) - \widetilde{A}(\chi, \nu)).$$

Таким чином, алгоритм визначення розладнання з прогнозуванням буде наступним:

1. Завдання початкових значень P, ν, χ.

Якщо відомо, що розладнання відбулося на один крок раніше початку спостереження, отже $P(r_0 = 0) = 1$. Інакше використовується нормалізована функція виживання:

$$P(r_0 = \tau) = \frac{1}{Z} S(\tau),$$

$$S(\tau) = \sum_{t=\tau+1}^{\infty} P_{gap}(g = t),$$

де $P_{gap}(g = t)$ – інтервал між розладнаннями.

2. Отримання наступного значення ряду x_t.
3. Обчислення прогнозованого значення ймовірності:

$$\pi_t^{(r)} = P(x_t \mid \nu_t^{(r)}, \chi_t^{(r)}).$$

4. Обчислення ймовірності продовження ділянки без розладнання:

$$P(r_t = r_{t-1} + 1, x_{1:t}) = P(r_{t-1}, x_{1:t-1})\pi_t^{(r)}(1 - H(r_{t-1})).$$

5. Обчислення ймовірності розладнання:

$$P(r_t = 0, x_{1:t}) = \sum_{r_{t-1}} P(r_{t-1}, x_{1:t-1})\pi_t^{(r)} H(r_{t-1}) .$$

6. Обчислення підтвердження:

$$P(x_{1:t}) = \sum_{r_t} P(r_t, x_{1:t}) .$$

7. Визначення розподілу на ділянці без розладнання:

$$P(r_t, x_{1:t}) = P(r_t, x_{1:t}) / P(x_{1:t}) .$$

8. Обчислення нових значень ν і χ:

$$\nu_{t+1}^{(r+1)} = \nu_t^{(r)} + 1,$$
$$\chi_{t+1}^{(r+1)} = \chi_t^{(r)} + u(x_t).$$

9. Прогнозування:

$$P(x_{t+1} \mid x_{1:t}) = \sum_{r_t} P(x_{t+1} \mid x_t^{(r)}, r_t) P(r_t \mid x_{1:t}) .$$

6 Непараметричні алгоритми послідовного виявлення розладнання. Алгоритми, що вимагають інформації про розподіл до і після розладнання, є точними, але, в той же час, можуть бути застосовані не завжди, адже часто інформацію про розподіл отримати неможливо. У такому разі застосовуються непараметричні методи.

У задачі аналізу стану ринку ймовірнісні характеристики процесу після розладнання є невідомою величиною, а отже використання параметричних методів виявлення розладнання ускладнене.

Наведена вище форма АКС потребує інформації про значення параметра θ після розладнання. Існує спосіб ослабити ці вимоги. Алгоритм налаштовується так, щоб реагувати на будь-які зміни заданої характеристики розподілу.

Нехай $\breve{\theta}_2$ – невідома величина, але відомий напрям її зміни, наприклад, $\theta_2 > \theta_1$. Тоді (4) можна переписати у формі, відомій також як тест Пейджа-Хінклі

$$\Delta g_t = x - m_1 - k \quad при\ m_2 > m_1 . \qquad (5)$$

Для випадку зменшення *m*

$$\Delta g_t = -x + m_1 + k \quad при\ m_2 < m_1 , \qquad (6)$$

де $k \geq 0$ – поріг чутливості методу для відхилення від θ_1.

У [8] була доведена еквівалентність процедур (5) і (6), вживаних одночасно, до наступної процедури:

$$R_t = \max \sum_{i=1}^{m} (x_i - \theta_i) - \min_{m \leq t} \sum_{i=1}^{m} (x_i - \theta_i) .$$

У [9] пропонується використовувати метод виявлення зміни медіани випадкової послідовності:

$$g_t = \max(0, g_{t-1} + \Delta g(x_t - k))$$

$$\Delta g(x) = \begin{cases} 1 & при\ x \geq 0 \\ -1 & при\ x < 0 \end{cases}$$

де k – значення медіани.

Бродським і Дарховським в [10] був запропонований непараметричний метод послідовного визначення, заснований на обчисленні наступної статистики:

$$Z(n) = \max_{\lfloor \alpha M \rfloor \leq k \leq \lfloor (1-\alpha) M \rfloor} \left| Y_M(k, n) \right|,$$

$$Y_M(k, n) = \frac{1}{k} \sum_{n-M+1}^{n-M+k} x(i) - \frac{1}{M-k} \sum_{n-M+k+1}^{n} x(i) ,$$

де n – довжина часового ряду, $0 < \alpha < \frac{1}{2}$.

Сигнал про розладнання подається, якщо значення Z(n) перевищує поріг визначення c.

Алгоритм має три параметри налаштування: α, *M*, *c*. Із збільшенням об'єму пам'яті M поліпшується якість виявлення, тому його можна вибирати виходячи з обчислювальних можливостей. Значення $\alpha < 2M$ вибирається з урахуванням того, що величина затримки визначення має порядок α*M*. Значення c обирається експериментальним шляхом для конкретного числового ряду.

Тими ж авторами були запропоновані непараметричні модифікації раніше згаданих методів послідовного виявлення.

1. Алгоритм кумулятивних сум. Авторами запропоновано використовувати значення ряду *x* як прирощення суми, а накопичену суму порівнювати з деяким «великим» числом c

$$y_t = \max(0, (y_{t-1} + x_t))$$
$$\tau = \inf(t : y_t \geq c) .$$

Процедура виявлення відхилення в обидві сторони:

$$\tau = \inf(t : | y_t | \geq c) .$$

Якщо ряд не є центрованим, необхідне додаткове центрування значень відносно маточікування ряду, підрахованого на діапазоні заданої довжини l.

2. Алгоритм ГРШ. У даній модифікації методу замість відношення щільності ймовірності використовується експоненціальна функція

$$W_t = (1 + W_{t-1})\, e^{x(t)}.$$

Як і в базовому методі ГРШ, вирішальне правило має вигляд: $\tau = \inf(t : W_t \geq b)$.

3. Метод заснований на експоненціальному згладжуванні

$$S_t = (1-k)S_{t-1} + kx_t.$$

В [11] запропоновано наступний непараметричний метод визначення зміни середнього

$$S_t = \max((p+q), (S_{t-1} + q\ sign(x_t - x_{t-m}) - p)),$$

де q > p – натуральні нескорочувані числа, p + q – порог чутливості алгоритму.

Метод виявляє зміну середнього у сторону збільшення. Для визначення зміни у сторону зменшення перепишемо рівняння у вигляді:

$$S_t = \max((p+q), (S_{t-1} - q\ sign(x_t - x_{t-m}) + p)).$$

Вирішальна функція алгоритму, заснованого на принципі нев'язок, формується як нев'язка (розбіжність) між моделлю випадкового процесу, що спостерігається, і прийнятою раніше моделлю [12]

$$G(n) = \frac{1}{\sqrt{2n}} \sum_{i=n-M}^{n} \left(\left(\frac{x_i - m_i}{\sigma_i^2} \right)^2 - 1 \right), \qquad (7)$$

де M – глибина пам'яті, m_i – математичне очікування процесу до розладнання, σ_i^2 – дисперсія до розладнання.

Формула (7) може бути перетворена до рекуррентного вигляду:

$$G(n) = \sqrt{\frac{n-1}{n}} \cdot G(n-1) + \frac{1}{\sqrt{2n}} \left(\left(\frac{x_n - m_n}{\sigma_n^2} \right)^2 - 1 \right).$$

Початкові умови:

$$G(1) = \frac{1}{\sqrt{2}} \left(\left(\frac{x_1 - m_1}{\sigma_1^2} \right)^2 - 1 \right).$$

У [13] розладнання пропонується розглядати зміну в оновлюючих послідовностях фільтра Калмана. Запис алгоритму в матричному вигляді:

$$G(n) = \begin{cases} 0, n < M \\ \det(S_n),\ n \geq M \end{cases},$$

$$S_n = \frac{1}{M-1} \sum_{i=n-M+1}^{n} (x_i - m_n)(x_i - m_n)^T, n \geq M,$$

де $\det(S_n)$ – визначник матриці S_n, M – глибина пам'яті, m – математичне очікування послідовності до розладнання:

$$m_n = \frac{1}{M} \sum_{i=n-M+1}^{n} x_i, n \geq M.$$

Незважаючи на те, що алгоритм призначений для роботи з матричними параметрами x, він може бути застосований і для аналізу одновимірних послідовностей. Прийнявши розмірність x за одиничну, приходимо до рекурентних формул:

$$S_n = S_{n-1} + \frac{x_n - x_{n-M}}{M-1} \left(x_n + x_{n-M} - 2m_{n-1} - \frac{1}{M} \right),$$

$$m_n = m_{n-1} + \frac{x_n - x_{n-M}}{M}, n > M.$$

Початкові умови:

$$S_M = \frac{1}{M-1} \sum_{i=1}^{M} (x_i - m_M)^2.$$

Метод визначення мінімуму інформаційної неузгодженості [14] також не потребує для роботи даних про характеристики розподілу після розладнання. Алгоритм заснований на знаходженні інформаційної неузгодженості автоковаріаційних матриць, побудованих на основі двох вибірок: $X_1 = (x_{1,1}, x_{1,1}, ..., x_{1,M_1})$ і $X_1 = (x_{2,1}, x_{2,1}, ..., x_{2,M_2})$ об'ємом M_1 і M_2 відповідно.

$$S_1 = \frac{1}{M_1} \sum_{i=1}^{M_1} x_{1,i}\ x_{1,i}^T,$$

$$S_2 = \frac{1}{M_2} \sum_{i=1}^{M_2} x_{2,i}\ x_{2,i}^T,$$

$$S_0 = (M_1 / M_0) S_1 + (M_2 / M_0) S_2,$$

де $M_0 = M_1 + M_2$.

Сигнал про розладнання подається за умови:

$$\lambda(X_0) = M_1\ \gamma_{1,0} + M_2\ \gamma_{2,0} - 0{,}5(M_0 n) \geq \ln(\lambda_0),$$

де $\gamma_{k,0} = 0{,}5(\mathrm{tr}(S_k\ S_0^{-1}) - \ln|\ S_k\ S_0^{-1}\ |)$ – величина інформаційної неузгодженості, $|\cdot|$ – визначник матриці, tr – слід квадратної матриці, n – константа.

Метод передбачає, що процес X є центрованим з нульовим математичним очікуванням.

Висновки. Методи пошуку розладнання дозволяють «стискати» великі об'єми даних, виділяючи з них тільки найважливішу для подальшого аналізу інформацію – дані про зміну динаміки процесу.

Скорочення об'ємів даних дозволяє працювати у процесі прогнозування з більшими часовими проміжками історії цін, не збільшуючи об'єм вхідних даних.

Указані методи використані в автоматизованій системі аналізу валютного ринку «Форекс-радник» в якості основи експертної системи оцінки ризиків торгівельних операцій.

Бібліографічні посилання

1. **Girshich M.A**. Bayes Approach to a Quality Control Model. / M.A. Girshich, H. A. Rubin // Ann. Math. Statist. – 1952. – Vol. 23. – № 1. – P. 114–125.
2. **Page E.S.** Continous Inspection Schemes / E.S. Page // Biometrika – 1954. – Vol. 41. – № 1. – P. 110–115.
3. **Lorden G.** Procedures for Reacting to a Change of Distribution / G. Lorden // Ann. Math. Statist. – 1971. – Vol. 42, № 1. – P.1897–1908.
4. **Shewart, W.A.** (1931). Economic Control of Quality of Manufactured Product / W.A. Shewhart – Seattle. Quality Press, 1980. – 501p.
5. **Roberts S.W.** Control chart tests based on geometric moving average / S.W. Roberts // Technometrics, 1959. – Vol. 1, № 3 – P.239–250.
6. **Калишев О.Н.** Метод диагностирования измерительных каналов с учетом предыстории // Автоматика и телемеханика. – 1988. – №6. – С. 135–143.
7. **Adams R.P.** Bayesian Online Changepoint Detection / R.P. Adams, D.J.C. McKay. Режим доступу: http://arxiv.org/pdf/0710.3742v1.pdf
8. **Nadler J.** Some characteristics of Page's twosided procedure for detecting a change in a location parameter / J. Naddler, N.B. Robbins // Ann. Math. Statist. – 1971. – Vol. 42, № 2. – P. 538–551.
9. **McGilchrist C.A.** Note on distribution-free CUSUM technique / C.A. McGilchrist, K.D. Woodyer // Technometrics, 1975 – Vol. 17, № 3. – P. 321–325.
10. **Brodsky B.E.** Nonparametric Methods in Change-Point Problems. / B.E. Brodsky, B.S. Darkhovsky – Dordrecht: Kluver Academic Publishings, 1993. – 210 p.
11. **Воробейчиков С.Э.** Об обнаружении изменения среднего в последовательности случайных величин / С.Э. Воробейчиков // Автоматика и телемеханика. – 1998. – №3. – С. 50–56.
12. **Бородкин Л.И**. Алгоритм обнаружения моментов изменения параметров уравнения случайного процесса / Л.И. Бородкин, В.В. Моттль // Автоматика и телемеханика. – 1976. – №6. – С. 23–32.

13. **Гаджиев Ч.М.** Проверка обобщенной дисперсии обновляющей последовательности фильтра Калмана в задачах динамического диагностирования / Ч.М. Гаджиев // Автоматика и телемеханика. – 1994. – №8. – С. 98–104.
14. **Акатьев Д.Ю.** Обнаружение разладки случайного процесса на основе минимума информационного рассогласования. / Д.Ю. Акатьев, В.В. Савченко // Автометрия, 2005. – Т.41, №2. – С. 68–74.

Надійшла до редколегії 20.10.2012

УДК 534.4: 621.391

О.І. Лучинкіна, О.Н. Карпов, А.Г. Матвеєнко

Дніпропетровський національний університет ім. Олеся Гончара

РОЗРОБКА АЛГОРИТМІВ І ПРОГРАМ РОЗУМІННЯ ТА СИНТЕЗУ ШЕПІТНОЇ МОВИ

Розглядається шепітна мова, як одна з різновидів мови, для якої з'являється проблема складного і неоднозначного сприйняття. Пропонується алгоритм її обробки з метою підвищення розбірливості і подальшого синтезу.

Ключові слова: *апроксимація, дзвіноподібні функції, лінкер, локон Аньєзі, неоднозначність, розпізнавання мови, основний тон, спектр, тональні характеристики сигналу, функція Гауса, частота, шепітна мова, mathlab*

Рассматривается шепотная речь, как одна из разновидностей речи, для которой появляется проблема сложного и неоднозначного восприятия. Предлагается алгоритм ее обработки с целью повышения разборчивости и дальнейшего синтеза.

Ключевые слова: *аппроксимация, колоколообразные функции, линкер, локон Аньези, неоднозначность, распознавание речи, основной тон, спектр, тональные характеристики сигнала, функция Гауса, частота, шепотная речь, mathlab*

The whispered speech is considered as a form of speech, for which a problem of complex and ambiguous perception take place. An algorithm for processing it in order to improve legibility and further synthesis is proposed.

Key words: *approximation, bell-shaped functions, linker, witch of Agnesi Agnese, ambiguity, speech recognition, the fundamental tone, spectrum, tonal characteristics of the signal, the the function Gauss, frequency, whisper, mathlab*

Огляд проблеми. Шепітна мова – один з дефектів функціонування мовотворчої системи.

Причини розладів голосу [1; 2] вельми різноманітні: захворювання і травматичні ушкодження гортані та голосових зв'язок; порушення резонаторних систем; хвороби органів дихання; захворювання серця і серцево-судинної системи; ендокринні розлади, зокрема, захворювання щитовидної залози; порушення слуху, що утрудняють загальне «настроювання» голосотворного апарату зважаючи на відсутність або недостатність слухового контролю; тривале куріння; систематичне вживання алкоголю; вплив отрутохімікатів; часте

перебування у запорошених приміщеннях; систематичне перенапруження голосу, особливо при неправильному користуванні ним; різкі температурні коливання, зокрема, пиття холодної води і особливо холодного молока і соків; психічні травми.

Зазначені етіологічні фактори призводять до органічних і функціональних порушень голосу.

Дефект може бути викликаний різними захворюваннями і виражається в тому, що голосової м'язи не напружується, внаслідок чого в мові не з'являється тональна компонента.

З точки зору утворення мовного сигналу на вхід мовотворчого тракту надходять сигнали від двох генераторів:

а) тонального – коливання голосової м'язи;

б) шумового – турбулентний потік повітря.

Такий дефект може з'являтись через різні хвороби та виражається у тому, що голосові зв'язки не напружуються, в результаті чого з мови зникає тональна компонента.

При відсутності коливань голосових м'яз мова формується як зазвичай, але не звучить. Огинають спектрів тональних звуків такі ж, як і при наявності коливань голосових м'язів. Істотна відмінність – у внутрішній структурі спектрів. Спектри тональних звуків при наявності голосового джерела є лінійчатими з частотами ліній кратними частоті основного тону (ОТ), де $\omega_k = k\omega_0$, де ω_0 - частота ОТ. У спектрах шепітної мови немає ліній на частотах. Завдання озвучування тональних звуків є завданням розпізнавання слів, визначення меж тональних звуків і їх озвучення відповідно до їх спектрів.

Задача розпізнавання за методом та алгоритмом добре відома [3], хоча надійність розпізнавання – це величина що залежить від багатьох причин. Для ізольованих слів можна отримати гарну надійність розпізнавання – понад 90 % при налаштуванні системи на конкретного диктора. Для невеликого словника – це близько до 99 %. Реальна задача озвучування шепітної мови пов'язана з безперервною мовою, а надійність розпізнавання безперервної мови в загальному випадку низька – нижче 80 %.

Завдання озвучування – це завдання синтезу за відомою спектрально-часовою функцією тональної частини шепітної мови. Істотно важливим є завдання сегментації, тому що треба максимально точно визначати межі озвучування. Оскільки чергування гучних і тональних звуків носить випадковий характер, то при вирішенні задачі синтезу необхідно вести підрахунок ділянок мови для того, щоб

визначати які з них відповідно до транскрипції необхідно озвучувати. При вирішенні задачі розпізнавання на фонемному рівні завдання визначення цих ділянок вирішується автоматично.

Тому постає проблема боротьби з таким дефектом. Одна з можливостей вирішення такої проблеми полягає у спробі озвучення сегментів тональних звуків шепітної мови. Отримання звичайної вимови за шепітним сигналом покращить можливості спілкування з людьми, з яких пошкоджені голосові зв'язки.

Сутність проблеми, полягає в тому, щоб розробити таку систему, яка б по вхідному шепітному сигналу могла б автоматично розпізнати та синтезувати сигнал з тональною компонентою. Також проблема ускладнюється тим, що більшість існуючих методів розпізнавання і сегментації розроблені для сигналів зі звичайною вимовою. Тому система повинна використовувати алгоритми, які не спираються на тональні характеристики сигналу. Завдяки використанню таких алгоритмів якість розпізнавання та сегментації шепітної мови має бути не гіршою, ніж для звичайної мови.

Проблема розпізнавання шепітної мови – це не лише проблема налаштування системи. Це проблема, яка, насамперед, пов'язана з розумінням мови. Навіть людині іноді важко розпізнати шепітну мову.

У статті будуть розглянуті методи та алгоритми, які допоможуть вирішити не лише задачу розпізнавання. Вони направлені на те, щоб, насамперед, вирішити задачу розуміння мови та перетворення її таким чином, щоб вона була зрозуміла як людині, так і машині.

Постановка задачі. Вхідними даними для подальшого аналізу є звуковий сигнал шепітної мови, $\{S_i, i = \overline{1,k}\}$.

Уведення сигналу відбувається через завантаження з wave-файлу.

При розпізнаванні шепітної мови особлива увага приділяється якості сигналу, тому що додавання шумів або зниження якості сигналу відповідно знизить надійність розпізнавання. Таким чином повинна забезпечуватися належна якість аудіосигналу.

Необхідно розробити алгоритм, за допомогою якого можливо:

1) отримати мовленевий сигнал шляхом завантаження звукового файлу;

2) обробити мовленевий сигнал таким чином, щоб шепітна мова стала більш розбірливою та придатною для розпізнавання;

3) озвучити отриманий за допомогою алгоритму сигнал.

Методи та алгоритми розв'язання задачі розпізнавання шепітної мови. Розглянемо алгоритм розпізнавання шепітної мови,

який засновано на апроксимації мовленевого сигналу за допомогою дзвіноподібних функцій [4].

1. Первинна обробка сигналу. Над сигналом проводиться базова фільтрація, яка здійснюється за принципом відокремлення шумів на початку та в кінці звукового потоку залежно від маскувального рівня сигналу.

Нормалізація даних в задачах моделювання процесів аналізу мовних сигналів необхідна через те, що перешкоджає різним носіям мови впливати на кінцевий результат аналізу, зводить параметри мовного сигналу до канонічного вигляду.

У процесі подальшої обробки ми отримаємо спектрально-часове представлення вхідного звукового сигналу.

Спектрально-часове представлення може бути отримане у різних ортогональних базисах функцій:

- Фур'е;
- Мат'е;
- Лежандра;
- Чебишева;
- Уолша;
- Лаггера;
- Ерміта та інших.

На рис.1 зображене спектрально-часове представлення слова «ноль».

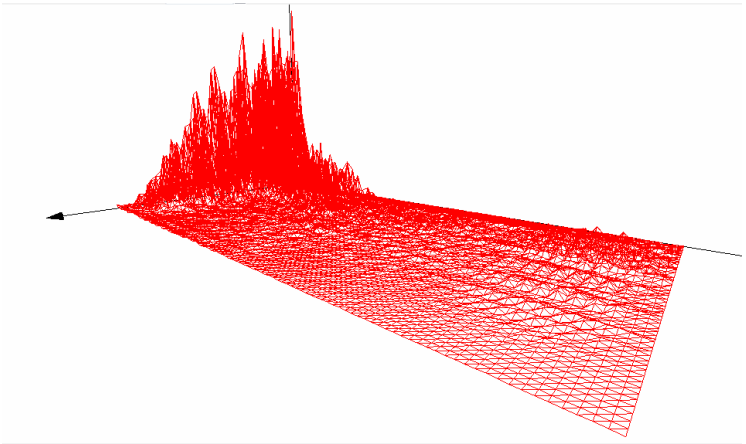


Рис. 1. Спектрально-часове представлення слова «ноль»

2. Апроксимація спектру класу дзвіноподібних функцій. Для подальшого аналізу сигналу спектрально-часове представлення, що отримано на попередньому кроці, необхідно очистити від шумів.

Можливі варіанти апроксимуючих дзвоноподібних функцій:

1. У класі комплексних функцій:

$$W_i(w_k)^2 = \frac{c_i^{\,2}(a_i^{\,2} + w_k^{\,2})}{(b_i^{\,2} + a_i^{\,2} - w_k^{\,2})^2 + (2w_k a_i)^2},\ h_i(t_l) = d_i e^{-(t_l - T_i)^2 f_i}, \qquad (1)$$

де $t_l = \ln \lambda_1$, λ_1 – час чи номер інтервалу.

2. Функції другого порядку у вигляді полінома:

$$W_i(w_k) = a_{i2} w_k^{\,2} + a_{i1} w_k + a_{i0},\ h_i(t_l) = b_{i2} t_l^{\,2} + b_{i1} t_l + b_{i0}. \qquad (2)$$

На основі парабол можна побудувати дзвоноподібну функцію з властивостями асимптотичності як сплайн з двома вузлами.

3. Модифікований локон Аньєзі:

$$W_i(w_k) = \frac{a_i^{\,3}}{c_i^{\,2} + (w_k - \beta_i)^2},\ h_i(t_l) = \frac{b_i^{\,2}}{d_i^{\,2} + (t_l - T_i)^2}. \qquad (3)$$

4. Функція Гауса:

$$W_i(w_k) = a_i e^{-(w_k - \beta_i)^2 v_i},\ h_i(t_l) = b_i e^{-(t_l - T_i)^2 \tau_i}. \qquad (4)$$

5. Експоненціальна функція:

$$W_i(w_k) = a_i e^{b_i w_k + v_i w_k^{\,2}},\ h_i(t_l) = r_i e^{d_i t_l + f_i t_l^{\,2}}. \qquad (5)$$

3. Вибір і формування типів мовних одиниць. В якості мовних одиниць можуть виступати:

а) сегменти;

б) склади;

в) фонеми;

г) слова.

4. Сегментація. Сегментація може проводитися одним з наступних методів:

а) фільтрація огинають спектра;

б) застосування порогових методів за груповими ознаками;

в) диференціювання функцій параметрів;

г) верифікація за сукупністю параметрів.

5. Навчання системи. Алгоритм розпізнавання вхідного сигналу, оснований на пошуку за базовим словником. Тому рішення задачі розпізнавання мови складається з двох частин: навчання системи (накопичення еталонів) і власне розпізнавання.

Етап навчання полягає в тому, що системі послідовно подаються сигнали навчальної вибірки, які будуть проаналізовані та додані до

словника. Наприклад, це можуть бути цифри від одного до дев'яти. Для кожного еталона виконуються всі етапи обробки сигналу:

- первинна обробка;
- частотно-часове перетворення;
- згладжування (апроксимація колоколоподібними функціями).

6. Розпізнавання мовленевого сигналу. Етап розпізнавання полягає у зіставленні поточного сигналу для розпізнавання з еталонами із словника. Поточний сигнал також повинен бути оброблений, щоб його можна були порівнювати з іншими еталонами. Для обробки, сигнал проходить всі описані вище етапи.

Цифрове представлення сигналів являє собою вектори частотних коефіцієнтів для кожного часового інтервалу (розбиття на часові інтервали описане вище). Для порівняння двох сигналів, попарно порівнюються вектори коефіцієнтів часових інтервалів цих сигналів. Для порівняння частотних коефіцієнтів використовується евклідова відстань між векторами

$$dist = \| V_1 - V_2 \|, \qquad (6)$$

де V_1, V_2 – вектори коефіцієнтів, $\|A\|$ – евклідова норма вектору.

Таким чином визначається відстань між двома сигналами d_i – ця відстань знаходиться за допомогою методу динамічного програмування. Метод динамічного програмування дозволяє вирішити завдання ослаблення часової нестаціонарності шляхом нелінійного розтягування короткої реалізації сигналу щодо довгої. Розпізнаним еталоном є еталон з найменшим значенням відстані

$$index = \arg\min_i (d_i). \qquad (7)$$

Потім за номером еталона index знаходиться відповідна транскрипція для шепітного сигналу.

Апроксимація мовленевого сигналу. Сформулюємо задачу апроксимації мовленевого сигналу [5,6]. Нехай задана прямокутна область $R = [w_a, w_b] \times [T_c, T_d]$, а в області R задана дискретна спектрально-часова функція $S(w_k, t_l)$, де w_k – дискретно задана частота, а t_l – дискретно заданий час. Область задана граничними значеннями w_a, w_b – частоти, T_c, T_d – часу.

Також нехай визначені класи функцій $\{W_i(w_k)\}$ та $\{h_i(t_l)\}$, які мають наступні властивості:

- функція повинна бути дзвонообразною (мати максимум на заданій частоті) для представлення резонансу;
- форма функції повинна залежати від деяких параметрів;

- функція повинна асимптотично наближатись до площини області R в будь-якому напрямку від максимуму;
- гілки функції повинні описувати інерційні та диференційні властивості при моделюванні частин системи чи монотонно спадати (зростати), якщо максимум знаходиться поза границями області.

Функція спектру сигналу $S(w_k,t_l)$ має будь-яку кількість сплесків спектральної енергії, розташованих у даній області. Необхідно описати функцію $S(w_k,t_l)$ в класі функцій $\{W_i(w_k)\}$ та $\{h_i(t_l)\}$, і визначити параметри сплесків функції спектру як параметри функцій $\{W_i(w_k)\}$ та $\{h_i(t_l)\}$.

Згідно з теоремою Колмогорова, будь-яка неперервна функція n змінних може бути отримана за допомогою композиції неперервних функцій однієї змінної та єдиної функції двох змінних $g(x,y) = x + y$. Він довів, що будь-яка функція f, неперервна на n-вимірному кубі, може бути представлена у вигляді

$$f(x_1,...,x_n) = \sum_{i=1}^{2n+1} h_i\left(\sum_{j=1}^{n} \varphi_{ij}(x_j)\right), \quad (8)$$

де функції h_i, φ_{ij} – неперервні, а функції φ_{ij} також стандартні, тобто не залежать від вибору функції f.

Визначимо модель опису функції $S(w_k,t_l)$ у вигляді

$$Z(w_k,t_l) = \sum_{i=1}^{n} | S_{gi}(w_k) || W_i(w_k) | h_i(t_l) \ , \quad (9)$$

де $S_{gi}(w_k)$ – спектральна функція генератора сигналу, n – кількість інформаційних сплесків.

Для визначення параметрів усіх функцій застосуємо схему послідовних вилучень інформаційних складових мовного сигналу (СЕТ). Параметри всіх функцій визначають послідовно для кожного доданка суми у функції $Z(w_k,t_l)$. Спектрально-часова функція може бути представлена у вигляді

$$S(w_k,t_l) = Z_{(1)}(\langle param_1 \rangle, w_k, t_l) + S_{(1)}(w_k,t_l). \quad (10)$$

Тоді після визначення параметрів params1 для функції $Z_{(1)}$, можна визначити залишок спектрально-часової функції у області

$$S(w_k,t_l) = Z_{(1)}(\langle param_1 \rangle, w_k, t_l) - S_{(1)}(w_k,t_l). \quad (11)$$

Дана процедура виділення залишку і знаходження параметрів повторюється далі для всіх доданків:

$$\begin{aligned} S(w_k,t_l) &= Z_{(i)}(\langle param_1 \rangle, w_k, t_l) + S_{(i)}(w_k,t_l) \\ S(w_k,t_l) &= Z_{(i)}(\langle param_1 \rangle, w_k, t_l) - S_{(i)}(w_k,t_l) \end{aligned}. \quad (12)$$

Функція S буде апроксимована повністю, коли будуть знайдені всі параметри кожної з n функцій доданків у Z.

Розглянемо алгоритм апроксимації.

1. Нехай маємо дискретну спектрально-часову функцію $S(w_k,t_l)$.

Для її апроксимації за схемою послідовних вилучень цю функцію розбивають на прямокутні регіони за часом та частоті. Розбиття за часом генерує інтервали однакової довжини на всьому часі сигналу – наприклад по 40 мс чи 4 шаги. Розбиття по частоті може бути рівномірним та нерівномірним. Низькі частоти несуть більше інформації, тому там потрібне детальніше розбиття. Вигляд розбиття зображено на рис. 2, регіони відділяються обмежуючими паралелепіпедами.

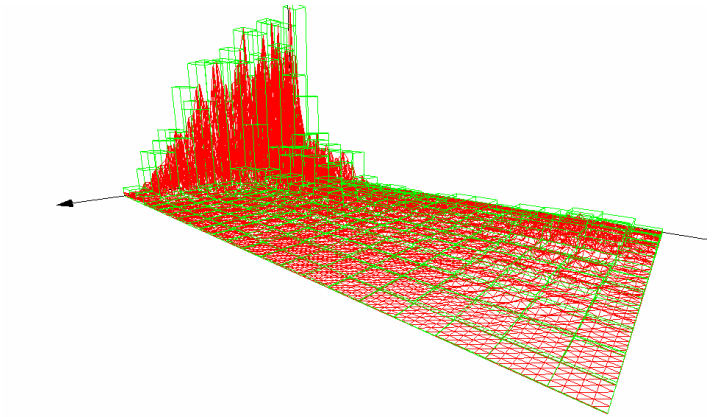


Рис. 2. Рівномірне розбиття на регіони для апроксимації для слова «ноль»

2. За допомогою методу покоординатного спуску проводимо апроксимацію частини спектру в кожному регіоні.

Суть методу полягає у послідовній мінімізації функції для кожного параметру окремо до того моменту поки не буде виконуватись умова зупинки алгоритму.

$$Z(w_k,t_l) = Z(p_1,...,p_m,w_k,t_l), \quad (13)$$

У даному випадку для функції вибирається наступний параметр для мінімізації p_j, а інші параметри фіксуються. Далі відбувається мінімізація одновимірної функції від обраного параметра

$$Z(w_k, t_l) = Z(p_j, w_k, t_l), \quad (13)$$

Так як градієнт даної функції невідомий, застосовується адаптивний алгоритм для знаходження значення параметра, який мінімізує функцію.

1) обирається початковий шаг step параметра;

2) функція тестується, щоб зробити крок у напрямках –step та +step;

3) якщо функція зменшується у деякому напрямку, то точка мінімуму зміщується і параметр змінюється відповідно;

4) якщо в обох напрямках функція більша за поточне значення, шаг параметра зменшується вдвічі;

5) умова зупинки алгоритму – шаг став менший деякого критичного значення.

Процес мінімізації багатовимірної функції повторюється для кожного параметра послідовно. Після останнього параметра знову обирається перший, доки не виконається умова зупинки алгоритму по координатного спуску. Ця умова – зміна значення функції на поточному та попередньому кроці менше деякого значення.

Інтеграція MATLAB в .NET. Для промови було використано програмне забезпечення синтезу реалізоване у середовищі MATLAB, тому постало питання інтеграції MATLAB у .NET [6].

Дуже часто перед програмістом постає завдання обчислення складної математики. MATLAB у свою чергу є відмінним засобом для його вирішення, але слабким у створенні повноцінного користувацького інтерфейсу.

Інструменти:

- Microsoft Visual Studio 2010 SP1
- MATLAB 2010a
- MATLAB Component Runtime

Крок 1. Налаштування лінкера.

Щоб зібрати dll-бібліотеку MATLAB'a для інтеграції у C #. NET, потрібно налаштувати лінкер, тобто яким середовищем ми будемо збирати проект. Для початку потрібно встановити середовище виконання MCR (MATLAB Component Runtime). Це набір dll-бібліотек для повної підтримки мови MATLAB. Установчий файл можна знайти: ... \ MATLAB \ R2011b \ toolbox \ compiler \ deploy \ win32 \ MCRInstaller. Для налаштування лінкера в командному вікні MATLAB'a набираємо mbuild -setup. З усім погоджуємося і вибираємо потрібну нам середу, в нашому випадку це MVS 2010 SP1. Результат представлено на рис.3.

Крок 2. Пишемо m-функцію.

Компілятор MATLAB'a розуміє тільки функції, тобто, кожен сценарій повинен починатися з function (бажано закінчуватися end) і бути окремим m-файлом.

```
Please choose your compiler for building standalone MATLAB applications:

Would you like mbuild to locate installed compilers [y]/n? y

Select a compiler:
[1] Lcc-win32 C 2.4.1 in C:\PROGRA~1\MATLAB\R2010a\sys\lcc
[2] Microsoft Visual C++ 2008 SP1 in C:\Program Files\Microsoft Visual Studio 9.0

[0] None

Compiler: 2

Please verify your choices:

Compiler: Microsoft Visual C++ 2008 SP1
Location: C:\Program Files\Microsoft Visual Studio 9.0

Are these correct [y]/n? y
```

Рис. 3. Налаштування лінкера

Крок 3. Отримуємо динамічну бібліотеку.

Набираємо в командному вікні MATLAB'a `deploytool` . Створюємо новий. NET Assembly проект MATLABplane, вказуємо розміщення.

Далі створюємо клас, додаємо в нього *.m файл і натискаємо кнопку build.

Після успішної компіляції створюється бібліотека *.dll, її шлях: ... \ distrib \ *.dll.

Крок 4. Створюємо додаток C #. NET.

У MVS 2010 SP1 створюємо додаток Windows Forms на C #.

Крок 5. Додаємо посилання на бібліотеки.

Перед використанням методів проекту необхідно додати посилання на скомпільовану бібліотеку *.dll і на бібліотеку MWArray.dll, знайти її можна за адресою ... \ MATLAB \ R2010a \ toolbox \ dotnetbuilder \ bin \ win32 \ v2.0.

Для використання бібліотек у проекті необхідно додати опис простору імен:

```
using MathWorks.MATLAB.NET.Utility;
using MathWorks.MATLAB.NET.Arrays;
```

using MATLABplane; // Простір імен з отриманої бібліотеки

Для отримання будь-яких значень з MWArray потрібно використовувати приведення типів. Тепер можна вільно використовувати написані на MATLAB'і функції.

Результати. На основі запропонованого алгоритму було створено програмний продукт розпізнавання шепітної мови.

Розглянемо результат роботи програми.

На рис. 4 представлена шепітна промова слова «ноль».

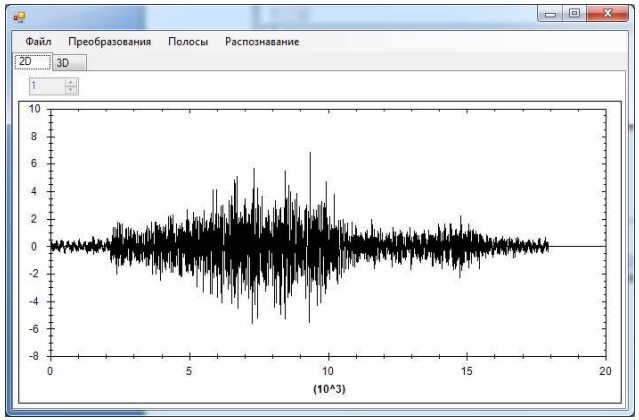


Рис. 4. Вигляд сигналу для шепітного слова «ноль»

Спектрально-часове (рис. 5, 6) та спектрально-смугове (рис.7) представлення звукового сигналу свідчить про його сильну зачумленність.

Апроксимація сигналу за допомогою дзвіноподібних функцій поліпшує картину. Як ми бачимо з рис. 8 за допомогою апроксимації спектрально-часове представлення (рис. 5, 6) було фідфільтровано та позбавлено зайвих високочастотних відхилень.

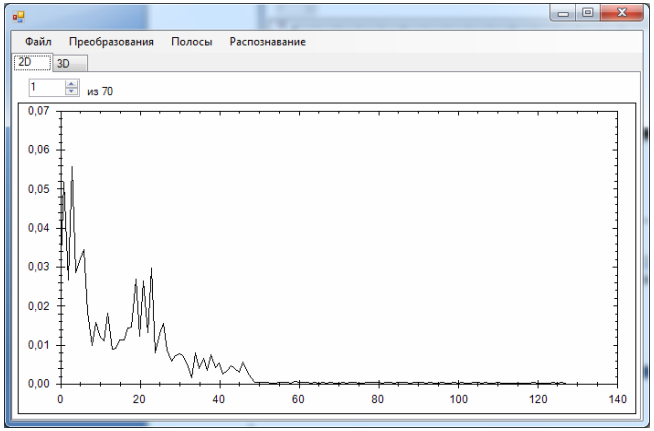


Рис. 5. Дискретне перетворення Фур'є для шепітного слова «ноль»

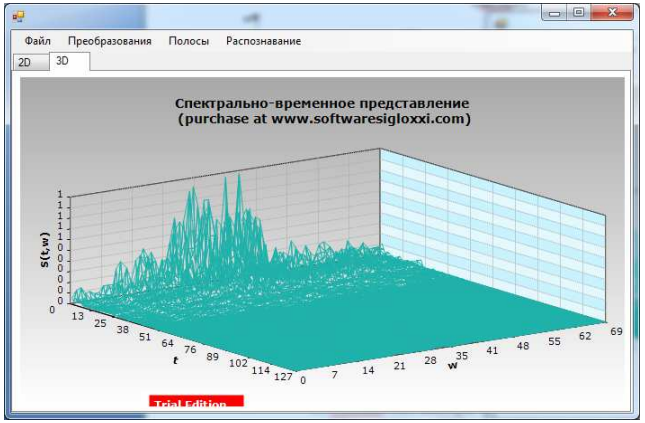


Рис. 6. Спектрально-смугове представлення шепітного слова «ноль»

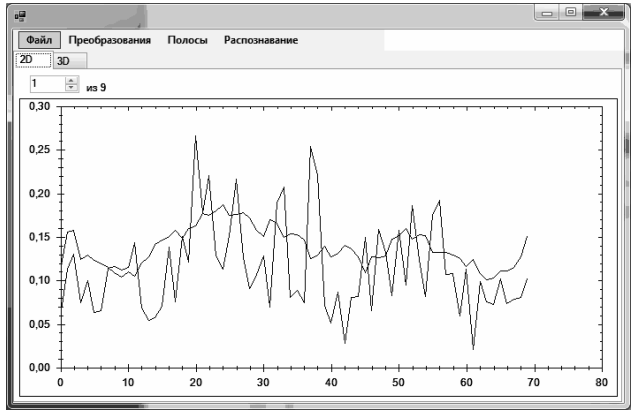


Рис. 7. Спектрально-смугове представлення та фільтрація слова «ноль»

Порівняння якості розпізнавання стандартного підходу з використанням спектрально-смужкового представлення та запропонований підхід з використанням розпізнавання апроксимованого за допомогою дзвоноподібних функцій спектрально-часового представлення показало, що запропонований підхід дає кращий результат для розпізнавання шепітної мови.

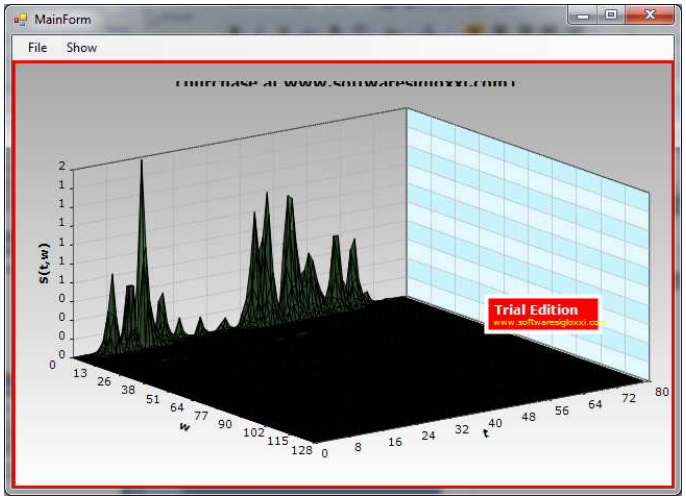


Рис. 8. Апроксимоване спектрально-часове представлення шепітного слова «ноль»

Недоліком даного підходу є складність програмної реалізації та потреба у більших системних ресурсах.

Отримані за допомогою побудованої системи розпізнавання мови результати було передано до системи MathLab для подальшого синтезу.

Висновки. Задача розпізнавання шепітної мови – це задача, по-перше, її розуміння та адаптації для сприйняття, а вже потім самого процесу розпізнавання.

Мовний сигнал формується під дією двох генераторів: тонального та шумового. При шепітній вимові звук утворюється тільки шумовим генератором, а голосові зв'язки не коливаються. Через зашумленість та специфіку вхідного сигналу задача розпізнавання шепітної мови не є тривіальною і не може бути вирішена за допомогою простих систем розпізнавання мови. У ході дослідження шепітної мови по запропонованому вище алгоритму було розроблено програмний продукт, за допомогою якого було поліпшено розбірливість мови, що сприяло підвищеню якості розпізнаванння. Крім того, оброблений мовленевий сигнал синтезовано за допомогою системи Mathlb.

Отриманий сигнал є більш сприятним для розпізнавання та зрозумілим.

Бібліографічні посилання

1. **Алмазова Е.С**. Логопедическая работа по восстановлению голоса у детей / Е.С. Алмазова – М., 2005. – 192 с.
2. **Парамонова Л.Г.** Логопедия для всех / Л.Г. Парамонова. – СПб, 1997. – 464 с.
3. **Карпов О.Н.** Компьютерные технологии распознавания речевых сигналов: моногр./ О.Н. Карпов, А.Г. Габович, Б.Г. Марченко. – К., 2005. – 138 с.
4. **Карпов О.Н.** Технология построения устройств распознавания речи: Моногр. / О.Н. Карпов – Д., 2001. – 182 с.
5. **Карпов О.Н.** Вычислительные схемы представления функций многих переменных в классе функций меньшего числа переменных. Методы анализа речевых сигналов: Моногр. / О.Н. Карпов. – Д., 2003. – 180 с.
6. **Карманов В.Г**. Математическое программирование / В.Г. Карманов. – М., 1980 – 256 с.
7. http://habrahabr.ru/post/132487/

Надійшла до редколегії 15.06.2012

УДК 519.233.2:519.254

О.М. Мацуга, Г.С. Шубіна

Дніпропетровський національний університет ім. Олеся Гончара

ОБЧИСЛЮВАЛЬНІ СХЕМИ ІДЕНТИФІКАЦІЇ СПЛАЙН-РОЗПОДІЛІВ ЗА ЙМОВІРНІСНИМ ПАПЕРОМ

Удосконалено запропоновані обчислювальні схеми автоматизованої ідентифікації сплайн-розподілів за ймовірнісним папером. Здійснено їх тестування на даних імітаційного моделювання з нормального, логарифмічно нормального, експоненціального та Вейбулла сплайн-розподілів.

Ключові слова: *ідентифікація, ймовірнісний папір, обчислювальна схема, сплайн-розподіл, вузол склеювання.*

Усовершенствованы предложенные вычислительные схемы автоматизированной идентификации сплайн-распределений по вероятностной бумаге. Проведено их тестирование на данных имитационного моделирования из нормального, логарифмически нормального, экспоненциального и Вейбулла сплайн-распределений.

Ключевые слова: *идентификация, вероятностная бумага, вычислительная схема, сплайн-распределение, узел склеивания.*

The computing schemes of spline-distributions automated identification using probability paper are improved. They are tested on modeling data from normal, logarithmic normal, exponential and Weibull spline-distributions.

Key words: *identification, probability paper, computing scheme, spline-distribution, node.*

Постановка проблеми. В інформаційному забезпеченні обробки статистичних даних актуальна проблема ідентифікації моделі розподілу за вибірковими даними. Один з найбільш потужних засобів ідентифікації є аналіз імовірнісного паперу, який робиться візуально. У роботі [1] авторами запропоновані обчислювальні схеми автоматизованої ідентифікації сплайн-розподілів за ймовірнісним папером. Їх ідея полягає в автоматизованому пошуку вузлів склеювання сплайн-розподілу на ймовірнісному папері. Проте результати тестування виявили, що запропоновані обчислювальні схеми інколи призводять до ідентифікації сплайн-розподілу із

завеликою кількістю вузлів. Дана робота націлена на подолання цього недоліку. Крім того, запропоновані в [1] обчислювальні схеми відпрацьовані лише на даних з нормальних сплайн-розподілів. Актуальною також представляється задача їх апробації на даних зі сплайн-розподілів і з інших родин.

Аналіз останніх досліджень і публікацій. Модель сплайн-розподілу введена та обґрунтована професором О.П. Приставкою [2]. Вона може бути візуально ідентифікована за ймовірнісним папером, на якому «чистому розподілу» відповідає пряма, а сплайн-розподілу – ламана лінія [2]. Вузли, в яких ламана змінює кут нахилу, відповідають вузлам склеювання сплайн-розподілу.

Нехай задано вибірку $\left\{x_i; i=\overline{1,N}\right\}$, за якою побудовано варіаційний ряд $\left\{x_i, n_i, p_i; i=\overline{1,r}\right\}$, де $x_1 < x_2 < \ldots < x_r$; r – кількість варіант; x_i – значення i-ї варіанти; n_i – частота i-ї варіанти; $\sum_{i=1}^{r} n_i = N$; $p_i = \frac{n_i}{N}$ – відносна частота i-ї варіанти; $\sum_{i=1}^{r} p_i = 1$. У кожній варіанті розраховано значення емпіричної функції розподілу $F_N\left(x_i\right) = \sum_{j=1}^{i} p_j$, $i=\overline{1,r}$.

Для ідентифікації сплайн-розподілу з родини розподілів $F\left(x\right)$ за вибіркою $\left\{x_i; i=\overline{1,N}\right\}$ будується ймовірнісний папір. З цією метою здійснюється перетворення функції $F\left(x\right)$ для надання їй лінійного вигляду. Справедливі такі перетворення для родин розподілів:

– нормального $\varphi\left(F\left(x\right)\right) = \Phi^{-1}\left(F\left(x\right)\right)$, $\varphi\left(x\right) = x$, де $\Phi^{-1}(\square)$ – обернена до функції Лапласа;

– логарифмічно нормального: $\varphi\left(F\left(x\right)\right) = \Phi^{-1}\left(F\left(x\right)\right)$, $\varphi\left(x\right) = \ln x$;

– експоненціального $\varphi\left(F\left(x\right)\right) = \ln\frac{1}{1-F\left(x\right)}$, $\varphi\left(x\right) = x$;

– Вейбулла $\varphi\left(F\left(x\right)\right) = \ln\ln\frac{1}{1-F\left(x\right)}$, $\varphi\left(x\right) = \ln x$.

На ймовірнісному папері відображається лінеарізована емпірична функція розподілу $F_N(x)$ у вигляді масиву $\left\{y_i = \varphi(x_i), z_i = \phi\left(F_N(x_i)\right); i = \overline{1,r}\right\}$.

У роботі [1] авторами запропоновано три обчислювальні схеми автоматизованого пошуку вузлів склеювання сплайн-розподілу на ймовірнісному папері.

Обчислювальна схема 1. *Пошук вузлів склеювання сплайн-розподілу на основі вимірювання кутів нахилу регресій.*

Фіксується варіанта x_t варіаційного ряду і формуються масиви

$$\left\{y_i, z_i; i = \overline{t-l,t}\right\}, \quad (1)$$

$$\left\{y_i, z_i; i = \overline{t,t+l}\right\}, \quad (2)$$

за якими відтворюються відповідні лінійні регресії:

$$z = k_1 y + b_1, \quad (3)$$

$$z = k_2 y + b_2. \quad (4)$$

Параметри k_1 та k_2 цих моделей являють собою тангенси кутів нахилу регресій до вісі абсцис, тому в якості різниці між кутами нахилу регресій можна ввести показник

$$K = \left|arctg\left|k_2\right| - arctg\left|k_1\right|\right|. \quad (5)$$

Ті варіанти, в яких значення показника K максимальні, доцільно вважати вузлами склеювання сплайн-розподілу.

1. Задати довжину «плеча» l, яка визначає довжини масивів (1) і (2).

2. Для кожної варіанти x_t, $t = \overline{l,r-l}$:

   2.1. Знайти оцінки параметрів $\hat{k}_1$ та $\hat{k}_2$ лінійних регресії (3) та (4) на основі масивів (1) та (2) відповідно, скориставшись методом найменших квадратів [3].

   2.2. Обчислити чергове значення показника K за формулою (5), і тим самим, сформувати масив $\left\{(x_t, K_t); t = \overline{l,r-l}\right\}$.

3. Здійснити згладжування (наприклад, медіанне) даних масиву $\left\{(x_t, K_t); t = \overline{l,r-l}\right\}$ з метою вилучення шумів (опціонально).

4. Знайти вузли склеювання сплайн-розподілу шляхом визначення варіант x_t, $t = \overline{l,r-l}$, яким відповідають максимальні елементи масиву $\left\{K_t; t = \overline{l,r-l}\right\}$. Максимальні елементи шукаються лише серед тих, що

перевищують заданий поріг. В якості порогу може бути використаний певний відсоток від максимального значення показника K, наприклад, значення $0{,}9K_{\max}$, де $K_{\max}$ – максимальне значення у масиві $\left\{K_t; t = \overline{l,r-l}\right\}$.

Обчислювальна схема 2. *Пошук вузлів склеювання сплайн-розподілу на основі порівняння двох регресійних прямих.*

Відмінність від попереднього випадку полягає в тому, що визначення наявності відмінності кутів нахилу регресій (3) та (4) зводиться до перевірки статистичної гіпотези $H_0: k_1 = k_2$ за альтернативи $H_1: k_1 \neq k_2$ на основі статистики [3]

$$K = \left| \frac{\hat{k}_1 - \hat{k}_2}{\sqrt{\frac{S^2_{\text{зал},1}}{(l+1)S_1^2} + \frac{S^2_{\text{зал},2}}{(l+1)S_2^2}}} \right|, \quad (6)$$

де $S^2_{\text{зал},1}$ та S_1^2 обчислюються на основі масиву (1) як:

$$S^2_{\text{зал},1} = \frac{1}{l+1}\sum_{i=t-l}^{t}\left(z_i - \hat{k}_1 y_i - \hat{b}_1\right)^2;$$

$$S_1^2 = \frac{1}{l+1}\sum_{i=t-l}^{t}\left(y_i - \overline{y}_1\right)^2; \qquad \overline{y}_1 = \frac{1}{l+1}\sum_{i=t-l}^{t} y_i;$$

$S^2_{\text{зал},2}$ та S_2^2 розраховуються за масивом (2) як:

$$S^2_{\text{зал},2} = \frac{1}{l+1}\sum_{i=t}^{t+l}\left(z_i - \hat{k}_2 y_i - \hat{b}_2\right)^2;$$

$$S_2^2 = \frac{1}{l+1}\sum_{i=t}^{t+l}\left(y_i - \overline{y}_2\right)^2; \qquad \overline{y}_2 = \frac{1}{l+1}\sum_{i=t}^{t+l} y_i.$$

Отже, обчислювальна схема 2 відрізняється від обчислювальної схеми 1 лише тим, що у пункті 2.2 чергове значення показника K розраховується за формулою (6).

Обчислювальна схема 3. *Пошук вузлів склеювання сплайн-розподілу на основі порівняння двох регресійних прямих, оцінених шляхом відтворення лінійної сплайн-регресії.*

У попередніх випадках регресії (3) та (4) на ймовірнісному папері будуються шляхом середньоквадратичного наближення до точок масивів (1) та (2), тому вони не обов'язково перетинаються. У даному випадку регресії будуються перетинними. Для реалізації цієї ідеї фіксується варіанта x_t варіаційного ряду і розглядається масив

$$\left\{y_i, z_i; i = \overline{t-l, t+l}\right\}, \tag{7}$$

за яким відтворюється лінійна сплайн-регресія вигляду

$$z = \begin{cases} k_1 y + b_1, & y \le y_t, \\ k_2\left(y - y_t\right) + k_1 y_t + b_1, & y \ge y_t. \end{cases} \tag{8}$$

Параметри k_1 та k_2, як і у першій схемі, являють собою тангенси кутів нахилу регресій до вісі абсцис.

1. Задати довжину «плеча» l, яка визначає довжину масиву (7).

2. Для кожної варіанти x_t, $t = \overline{l, r-l}$:

2.1. Обчислити оцінки параметрів $\hat{k}_1$ та $\hat{k}_2$ сплайн-регресії за масивом вигляду (7) [4].

2.2. Обчислити чергове значення показника *K* за формулою (5), і тим самим, сформувати масив $\left\{\left(x_t, K_t\right); t = \overline{l, r-l}\right\}$.

3. Провести згладжування (наприклад, медіанне) даних масиву $\left\{\left(x_t, K_t\right); t = \overline{l, r-l}\right\}$ з метою вилучення шумів (опціонально).

4. Знайти вузли склеювання сплайн-розподілу шляхом визначення варіант x_t, $t = \overline{l, r-l}$, яким відповідають максимальні елементи масиву $\left\{K_t; t = \overline{l, r-l}\right\}$. Максимальні елементи шукаються лише серед тих, що перевищують заданий поріг. Поріг може бути обраний як в обчислювальній схемі 1.

Як зазначалось вище, усі три обчислювальні схеми можуть знаходити надмірну кількість вузлів склеювання. Це пов'язано із тим, що на графіку залежності показника *K* від вузла склеювання може мати місце декілька близько розташованих локальних максимумів (рис. 1).

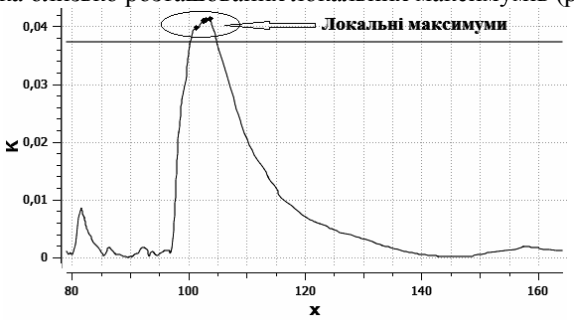


Рис. 1. Графік залежності показника *K* від вузла склеювання

Слід зазначити, що наявність пункту 3 (проведення згладжування) у кожній обчислювальній схемі націлена на подолання даної проблеми, проте не вирішує її повністю. Тому була поставлена наступна задача.

Постановка задачі. Нехай результати спостережень задано у вигляді вибірки $\left\{x_i; i = \overline{1, N}\right\}$, де N – кількість спостережень; x_i – спостережуване значення в *i*-му експерименті.

Ставиться задача ідентифікувати за цією вибіркою сплайн-розподіл шляхом автоматизованого визначення кількості його вузлів.

Для вирішення цієї задачі необхідно вдосконалити обчислювальні схеми 1–3 пошуку вузлів склеювання сплайн-розподілу на ймовірнісному папері, подолавши проблему знаходження декількох близько розташованих вузлів, та провести їх тестування на даних з нормального, логарифмічно нормального, експоненціального та Вейбулла сплайн-розподілів.

Основний матеріал. Пропонується модифікувати обчислювальні схеми 1–3, додавши до них ще один пункт, суть якого в наступному.

Нехай на основі пунктів 1–4 обчислювальних схем 1–3 знайдено оцінки вузлів склеювання $\left\{\hat{x}_{\text{в},i}; i = \overline{1, s}\right\}$. Для вилучення «зайвих» оцінок вузлів склеювання пропонуються два методи.

1. Метод інтервального оцінювання квантилів

Знайдені оцінки вузлів склеювання являють собою оцінки квантилів розподілу. Для кожного вузла на основі його оцінки будується довірчий інтервал як довірчий інтервал на квантиль. Якщо довірчі інтервали двох вузлів перетинаються, то залишається той вузол, для якого відповідне значення показника *K* більше, інший – видаляється.

Інтервальне оцінювання здійснюється згідно з нерівністю [3]

$$\hat{x}_{\text{в},i} - t_{1-\alpha/2,\nu}\sqrt{D\left(\hat{x}_{\text{в},i}\right)} \le x_{\text{в},i} \le \hat{x}_{\text{в},i} + t_{1-\alpha/2,\nu}\sqrt{D\left(\hat{x}_{\text{в},i}\right)},$$

де $D\left(\hat{x}_{\text{в},i}\right)$ – дисперсія оцінки вузла

$$D\left(\hat{x}_{\text{в},i}\right) = \frac{\hat{F}\left(\hat{x}_{\text{в},i}\right)\left(1 - \hat{F}\left(\hat{x}_{\text{в},i}\right)\right)}{N\hat{f}^2\left(\hat{x}_{\text{в},i}\right)};$$

$\hat{F}\left(\hat{x}_{\text{в},i}\right)$ та $\hat{f}\left(\hat{x}_{\text{в},i}\right)$ – значення оцінок функцій розподілу та щільності розподілу ймовірностей відповідно; α – імовірність «промаху» значення оцінки вузла повз довірчий інтервал (як правило, $\alpha = 0{,}05$).

Оцінювання функцій розподілу $\hat{F}(x)$ та щільності розподілу ймовірностей $\hat{f}(x)$ пропонується проводити за допомогою локальних поліноміальних сплайнів на основі *B*-сплайнів [5]. Їх перевага у високій швидкості роботи під час реалізації на ЕОМ та малій похибці апроксимації [5].

2. Метод сусідів

Суть методу полягає в тому, що попарно перевіряються знайдені оцінки вузли склеювання. Якщо для оцінок вузлів $\hat{x}_{в,k}$ та $\hat{x}_{в,l}$ різниця індексів $|k-l|$ менша за деякий заданий параметр *dist*, то такі оцінки вважаються близькими. Залишається той вузол, для якого відповідне значення показника *K* більше, інший – видаляється.

Запропоновані в роботі [1] обчислювальні схеми пошуку вузлів склеювання сплайн-розподілу та вищеописані модифікації цих схем було реалізовано у вигляді програмного забезпечення «SplineDistribution». Адекватність обчислювальних схем підтверджено результатами обчислювальних експериментів на даних імітаційного моделювання з нормального, логарифмічно нормального, експоненціального та Вейбулла сплайн-розподілів.

Нижче наводяться результати експерименту, під час якого моделювалась вибірка обсягом $n = 500$ зі сплайн-логнормального розподілу з одним вузлом з параметрами $x_0 = 0{,}8$, $m = 0{,}1$, $\sigma_1 = 0{,}5$, $\sigma_2 = 0{,}09$. На ймовірнісному папері логарифмічно нормального розподілу (рис. 2) чітко видно дві прямі з різними кутами нахилу, що відповідає сплайн-розподілу з одним вузлом.

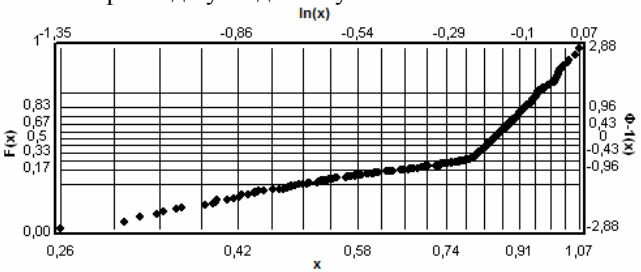


Рис. 2. Ймовірнісний папір логарифмічно нормального розподілу в експерименті 1

У таблиці 1 наведено результати ідентифікації за обчислювальними схемами на основі вимірювання кутів (ВК), порівняння двох регресійних прямих (ПРП), відтворення лінійної

сплайн-регресії (ВСР) та модифікованими схемами (м. – застосоване медіанне згладжування, ІОК – інтервальне оцінювання квантилів, МС – метод сусідів).

Таблиця 1

Ідентифіковані вузли сплайн-розподілу в експерименті 1

Обчислювальна схема	Оцінки вузлів	Параметри обчислювальної схеми
ВК	0,8001	Довжина «плеча» – 40, поріг – 90 %
ВК м.	0,8013	–
ВК ІОК	0,8001	–
ВК МС	0,8001	–
ВК м. ІОК	0,8013	–
ВК м. МС	0,8013	dist – 7
ПРП	0,8001	Довжина «плеча» – 40, поріг – 90 %
ПРП м.	0,8027	–
ПРП ІОК	0,8001	–
ПРП МС	0,8001	–
ПРП м. ІОК	0,8027	–
ПРП м. МС	0,8027	dist – 7
ВСР	0,79; 0,8021; 0,8027	Довжина «плеча» – 40, поріг – 90 %
ВСР м.	0,8007	–
ВСР ІОК	0,79; 0,8021; 0,8027	–
ВСР МС	0,79	dist – 7
ВСР м. ІОК	0,8007	–
ВСР м. МС	0,8007	–

Обчислювальні схеми на основі вимірювання кутів та порівняння двох регресійних прямих дозволили ідентифікувати один вузол склеювання, близький до заданого під час моделювання. При цьому не було потреби у застосуванні медіанного згладжування та інших модифікацій схем.

За обчислювальною схемою на основі відтворення лінійної сплайн-регресії ідентифіковано три близько розташовані вузли склеювання. Вирішити проблему надмірної кількості вузлів дозволило як застосування медіанного згладжування, так і методу сусідів. У кожному з випадків ідентифікувався один вузол склеювання.

При цьому всі ідентифіковані вузли склеювання за значенням дуже близькі до параметра моделювання. Найближчі значення оцінки вузла склеювання до параметра моделювання дала обчислювальна схема 2, що реалізує метод на основі порівняння двох регресійних прямих.

У цілому, результати чисельних експериментів засвідчили, що для якісної ідентифікації сплайн-розподілів з одним вузлом у більшості випадків достатньо базових обчислювальних схем 1 чи 2 або застосування медіанного згладжування. Для сплайн-розподілів з двома вузлами обов'язкове застосування методів інтервального оцінювання квантилів або сусідів.

Розглянемо результати експерименту, під час якого моделювалась вибірка обсягом $n = 300$ зі сплайн-нормального розподілу з двома вузлами з параметрами $x_0 = 95$, $x_1 = 120$, $m = 100$, $\sigma_1 = 10$, $\sigma_2 = 30$, $\sigma_3 = 10$. На ймовірнісному папері нормального розподілу (рис. 3) чітко виділяються три прямі з різними кутами нахилу, що відповідає випадку сплайн-розподілу з двома вузлами.

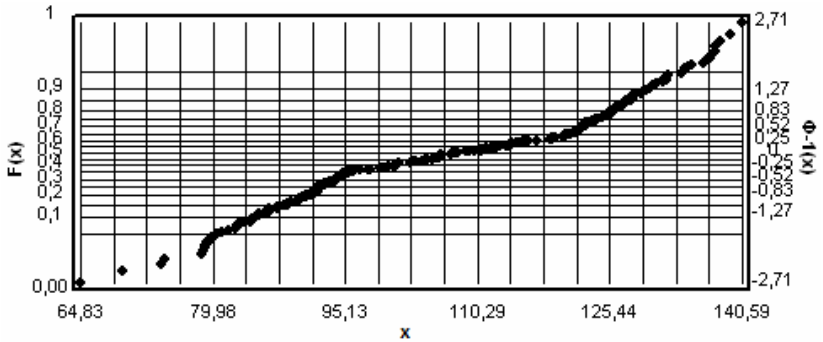


Рис. 3. Ймовірнісний папір нормального розподілу в експерименті 2

Результати ідентифікації вузлів сплайн-розподілу за різними методами для цього експерименту наведено у таблиці 2.

Усі три базові обчислювальні схеми ідентифікували надмірну кількість вузлів склеювання.

Застосування медіанного згладжування не виявилось ефективним для обчислювальних схем на основі вимірювання кутів та порівняння двох регресійних прямих. За його допомогою кількість ідентифікованих вузлів зменшилась, але не до потрібної кількості. Застосування методів інтервального оцінювання квантилів та сусідів як у сукупності з медіанним згладжуванням так і без нього дозволило одержати очікуваний результат – ідентифікувати два вузли склеювання, які мають місце в дійсності.

Таблиця 2

Ідентифіковані вузли сплайн-розподілу в експерименті 2

Обчислювальна схема	Оцінки вузлів	Параметри обчислювальної схеми
ВК	94,9; 95,0; 121,3; 122,0; 122,6; 123,4	Довжина «плеча» – 50, поріг – 90 %
ВК м.	94,9; 121,5; 122,5	–
ВК ІОК	95,0; 123,4	–
ВК МС	95,0; 123,4	dist – 19
ВК м. ІОК	94,9; 122,5	–
ВК м. МС	94,9; 122,5	dist – 13
ПРП	94,6; 95,0; 120,3; 121,3	Довжина «плеча» – 50, поріг – 90 %
ПРП м.	94,9; 120,9	–
ПРП ІОК	94,9; 121,3	–
ПРП МС	94,9; 121,3	dist – 6
ПРП м. ІОК	94,9; 120,9	–
ПРП м. МС	94,9; 120,9	–
ВСР	94,3; 94,8; 121,5; 121,9; 122,0; 122,6; 123,4	Довжина «плеча» – 50, поріг – 90 %
ВСР м.	94,7; 122,3	–
ВСР ІОК	94,8; 123,4	–
ВСР МС	94,8; 123,4	dist – 9
ВСР м. ІОК	94,7; 122,3	–
ВСР м. МС	94,7; 122,3	–

Для обчислювальної схеми 3 на основі відтворення лінійної сплайн-регресії застосування лише медіанного згладжування дозволило ідентифікувати два вузли склеювання.

Найближчі значення оцінок вузлів склеювання до параметрів моделювання дала обчислювальна схема 2.

З метою перевірки якості ідентифікації на основі запропонованих обчислювальних схем та їх модифікацій був проведений такий експеримент. Моделювалось по 100 вибірок обсягом 300 елементів з кожного сплайн-розподілу і за кожною вибіркою проводилась ідентифікація вузлів склеювання. Успіхом вважалась подія, коли ідентифікована така кількість вузлів, яка була задана під час моделювання. Результати експерименту (відсоток успіхів) наведено у таблиці 3. Під час експерименту використано такі параметри обчислювальних схем: довжина «плеча» – 40, поріг – 8 5 %, dist – 10.

Таблиця 3

Відсоток випадків, коли ідентифікована правильна кількість вузлів сплайн-розподілів

Обчислювальна схема	Сплайн-розподіл з одним вузлом	Сплайн-розподіл з двома вузлами
ВК	77,5	21,3
ВК м.	100,0	72,5
ВК ІОК	96,3	70,0
ВК МС	97,8	71,8
ВК м. ІОК	100,0	90,0
ВК м. МС	100,0	91,2
ПРП	91,3	25,0
ПРП м.	100,0	68,8
ПРП ІОК	98,8	75,0
ПРП МС	100,0	75,0
ПРП м. ІОК	100,0	86,3
ПРП м. МС	100,0	87,1
ВСР	33,7	6,3
ВСР м.	100,0	71,3
ВСР ІОК	65,0	61,3
ВСР МС	71,2	65,5
ВСР м. ІОК	100,0	87,5
ВСР м. МС	100,0	90,0

Дані таблиці 3 свідчать, що для сплайн-розподілів з одним вузлом базова обчислювальна схема 2 без жодних модифікацій дозволяє у більш ніж 90 % випадках ідентифікувати правильну кількість вузлів склеювання. Усі три обчислювальні схеми у сукупності з медіанним згладжуванням завжди забезпечують ідентифікацію правильної кількості вузлів склеювання. Отже, для сплайн-розподілів з одним вузлом немає потреби у застосуванні методів інтервального оцінювання квантилів та сусідів.

Для сплайн-розподілів з двома вузлами базові обчислювальні схеми є неефективні. Застосування медіанного згладжування частково покращує ситуацію. Усі обчислювальні схеми з медіанним згладжуванням дозволяють ідентифікувати правильну кількість вузлів склеювання приблизно у 70 % випадках, але це не можна вважати прийнятним результатом. Застосування обчислювальних схем без медіанного згладжування, але з методом інтервального оцінювання квантилів або сусідів також не забезпечує достатньо якісних результатів. Ефективним є застосування обчислювальних схем

одночасно з медіанним згладжуванням та методом інтервального оцінювання квантилів або сусідів. У такому разі правильна кількість вузлів склеювання ідентифікується приблизно у 90 %, а у випадку застосування обчислювальної схеми 1 з медіанним згладжуванням та методом сусідів – у більш ніж 90 % випадках.

Висновки. Було вдосконалено обчислювальні схеми ідентифікації сплайн-розподілів на основі ймовірнісного паперу. Модифікація полягала у застосуванні до ідентифікованих вузлів склеювання методу інтервального оцінювання квантилів та методу сусідів для вилучення надмірних вузлів склеювання. Останній метод виявився дуже ефективним та зручним у реалізації та використанні.

Модифіковані обчислювальні схеми було додано до ядра програмного забезпечення «SplineDistribution».

Здійснено тестування модифікованих обчислювальних схем на даних імітаційного моделювання сплайн-розподілів з класів нормального, логарифмічно нормального, експоненціального та Вейбулла. Результати тестування засвідчили, що модифіковані обчислювальні схеми дозволяють у більш ніж у 90 % випадків ідентифікувати сплайн-розподіл із правильною кількістю вузлів.

У подальшому доцільним представляється дослідження впливу параметрів обчислювальних схем на результати ідентифікації.

Бібліографічні посилання

1. **Мацуга О.М.** Обчислювальна схема ідентифікації розподілів з класу нормального / О.М. Мацуга, Г.С. Шубіна // Актуальні проблеми автоматизації та інформаційних технологій. – 2011. – Т.15. – С. 161– 172.

2. **Приставка О.П.** Сплайн-розподіли у статистичному аналізі / О.П. Приставка. – Д., 1995. – 152 с.

3. **Бабак В.П.** Статистична обробка даних / В.П. Бабак, А.Я. Білецький, О.П. Приставка, П.О. Приставка. – К., 2001. – 388 с.

4. **Приставка А.Ф.** Вычислительные методы и программная среда корреляционного и регрессионного анализа / А.Ф. Приставка, А.И. Передерий, О.В. Райко, В.М. Остропицкий. – Д., 1996. – 192 с.

5. **Приставка П.О.** Поліноміальні сплайни при обробці даних / П.О. Приставка. – Д., 2004. – 236 с.

Надійшла до редколегії 07.05.2012

УДК 519.234.2:004.67

О.М. Мацуга, Г.О. Басова

Дніпропетровський національний університет ім. Олеся Гончара

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ВІДНОВЛЕННЯ РОЗПОДІЛІВ ЗА МАЛИМИ ВИБІРКАМИ

Розроблено програмне забезпечення «SmallSample» відновлення розподілів за малими вибірками. Проведено порівняння роботи непараметричних методів відновлення розподілів за малими вибірками.

Ключові слова: *мала вибірка, програмне забезпечення, відновлення, функція розподілу, функція щільності.*

Разработано программное обеспечение «SmallSample» восстановления распределений по малым выборкам. Проведено сравнение работы непараметрических методов восстановления распределений по малым выборкам.

Ключевые слова: *малая выборка, программное обеспечение, восстановление, функция распределения, функция плотности*.

The program «SmallSample» for distributions restoration using small samples is developed. The distributions nonparametric restoration methods for small samples are compared.

Key words: *small sample, program, restoration, distribution function, density function*.

Постановка проблеми. В умовах обмеженості інформації часто доводиться мати справу з малими вибірками, під якими розуміються вибірки з малою кількістю спостережень над випадковою величиною. Більшість статистичних методів (як параметричних, так і непараметричних) малоефективні у таких випадках, оскільки вони асимптотичні. Тому існує необхідність у використанні методів, спеціально розроблених для вибірок малого обсягу. Такі методи запропоновано [1], проте їх програмна реалізація недоступна. У зв'язку з цим актуальна задача розробки програмного забезпечення, орієнтованого на статистичну обробку малих вибірок. У роботі увага приділена одному з етапів статистичної обробки – відновленню розподілів за малими вибірками.

Аналіз останніх досліджень і публікацій. Історично одними з перших робіт в області статистичного аналізу малої вибірки можна

вважати роботи Стюдента і Фішера, в яких досліджувались задачі оцінювання характеристик випадкової величини. Потім у 60–70-х роках XX століття над цією проблемою працювало багато інших математиків, таких як А.Н. Колмогоров, А.А. Петров, Л.Н. Большев, І.Н. Володін та інші [1].

Пік основних публікації стосовно задачі оцінювання функції розподілу ймовірностей випадкової величини за вибіркою малого обсягу припадає на 70–80-ті роки XX ст. Уперше дана задача була розглянута у роботі В.В. Чавчанідзе, В.А. Кумсішвілі. Результати цього та подальших досліджень узагальнено наприкінці 80-х років XX ст. у монографії Д.В. Гаскарова та В.І. Шаповалова [1], яку можна вважати основною роботою з питань статистичної обробки малих вибірок.

Серед запропонованих методів оцінювання функції розподілу випадкової величини за малою вибіркою можна виділити методи прямокутних вкладів, зменшення невизначеності, апріорно-емпіричної функції, які можуть бути застосовані навіть у випадку вибірки з 3–5 елементів [1]. Для них характерний індивідуальний підхід до кожної окремої реалізації вибірки, що відрізняє їх від методів побудови оцінок за великими вибірками. В якості додаткової апріорної інформації передбачається знання інтервалу $[a;b]$, на якому змінюється випадкова величина.

Оцінювання функції розподілу по зазначеним методам проводиться наступним чином.

Нехай спостереження над дійсною неперервною випадковою величиною X задано у вигляді вибірки $\left\{x_i; i=\overline{1,N}\right\}$. За вибіркою побудовано варіаційний ряд $\left\{x_i, k_i, p_i; i=\overline{1,n}\right\}$, де $x_1 < x_2 < \ldots < x_n$; n – кількість варіант; x_i – значення i-ї варіанти; k_i – частота i-ї варіанти; $\sum_{i=1}^{n} k_i = N$; $p_i = \frac{k_i}{N}$ – відносна частота i-ї варіанти; $\sum_{i=1}^{n} p_i = 1$.

Метод прямокутних вкладів (МПВ) дозволяє знайти оцінку функції щільності розподілу ймовірностей випадкової величини за вибіркою у вигляді [1]

$$f^*(x)=\frac{1}{N+1}\left(f_0(x)+\sum_{i=1}^{N}\psi_{x_i}(x)\right),$$

де $f_0(x)$ – апріорна компонента, відносно якої припускається що вона являє собою щільність рівномірного на $[a;b]$ розподілу

$$f_0(x)=\begin{cases}\dfrac{1}{b-a}, & x\in[a;b],\\ 0, & x\notin[a;b];\end{cases}$$

$\psi_{x_i}(x)$ – функція вкладу реалізації x_i довжиною d, яка забезпечує «розмивання» інформації про випадкову величину, яку було отримано від реалізації x_i

$$\psi_{x_i}(x)=\begin{cases}\dfrac{1}{d}, & x\in\left[x_i-\dfrac{d}{2};x_i+\dfrac{d}{2}\right],\\ 0, & x\notin\left[x_i-\dfrac{d}{2};x_i+\dfrac{d}{2}\right].\end{cases}$$

Оптимальне значення величини d залежить від обсягу вибірки та типу оцінюваного розподілу [1].

Оцінка функції розподілу $F^*(x)$ для МПВ одержується шляхом інтегрування функції щільності $f^*(x)$:

$$F^*(x)=\frac{1}{N+1}\left(F_0(x)+\sum_{i=1}^{N}\varphi_{x_i}(x)\right),$$

де

$$F_0(x)=\begin{cases}0, & x<a,\\ \dfrac{x-a}{b-a}, & x\in[a;b],\\ 1, & x>b;\end{cases}$$

$$\varphi_{x_i}(x)=\begin{cases}0, & x<x_i-\dfrac{d}{2},\\ \dfrac{1}{d}\left(x-x_i+\dfrac{d}{2}\right), & x\in\left[x_i-\dfrac{d}{2};x_i+\dfrac{d}{2}\right],\\ 1, & x>x_i-\dfrac{d}{2}.\end{cases}$$

У методі зменшення невизначеності (МЗН) оцінка функції розподілу знаходиться за варіаційним рядом за формулою [1]

$$F^*(x)=\frac{x-x_{i-1}}{x_i-x_{i-1}}\left(F^*(x_i)-F^*(x_{i-1})\right)+F^*(x_{i-1}),$$

коли $x\in[x_{i-1};x_i]$ та, якщо $x=x_i$, то

$$F^*(x_i)=\frac{1}{N+1}\left(\frac{x_i-a}{b-a}+(i-0,5)+(k_i-1)\right).$$

За методом апріорно-емпіричної функції (МАЕФ) оцінка функції розподілу визначається виразом [1]

$$F(x)=\omega F_a(x)+(1-\omega)F_N(x),$$

де $F_a(x)$ – апріорно задана функція розподілу; $F_N(x)$ – емпірична функція розподілу; ω – коефіцієнт достовірності інформації про апріорний розподіл.

Оцінка функції щільності $f^*(x)$ у МЗН та МАЕФ може бути одержана шляхом диференціювання $F^*(x)$ за змінною x.

Слід зазначити, що емпірична функція розподілу $F_N(x)$ є найбільш проста та традиційна непараметрична оцінка функції розподілу $F(x)$. Можливість її використання в якості оцінки функції розподілу було доведено В.І. Глівенко та Ф.П. Кантеллі. Емпірична функція розподілу визначається варіаційним рядом за формулою

$$F_N(x)=\begin{cases}0, & x\le x_1,\\ \sum\limits_{j=1}^{i}p_j, & x\in(x_i;x_{i+1}],\\ x, & x>x_n.\end{cases}$$

Розглянуті вище методи є непараметричні.

Спеціальних параметричних методів для оцінювання функції розподілу за малими вибірками не запропоновано. Застосування ж традиційних параметричних методів (максимальної правдоподібності, найменших квадратів, моментів тощо [2]) потребує знання мінімального обсягу вибірки, за якого вони можуть бути застосовані. Існуючі дослідження з цього приводу малочисельні та розрізнені. Деякі автори вважають обмеженим вибірки обсягом менше 200 елементів, інші називають малими вибірки в менш ніж 50 елементів [1]. Тому поставлена у роботі задача передбачає програмну реалізацію не лише

непараметричних методів оцінювання функцій розподілу за малими вибірками, а й параметричних з оцінкою мінімального обсягу вибірки.

Постановка задачі. Задано вибірку спостережень $\{x_i; i = \overline{1,N}\}$ над дійсною неперервною величиною X з невідомою функцією розподілу $F(x)$. Припускається, що вибірка малого обсягу. Ставиться задача за вибіркою знайти оцінку $F^*(x)$ функції розподілу $F(x)$, або, що тотожно, знайти оцінку $f^*(x)$ щільності розподілу $f(x)$.

Для вирішення цієї задачі необхідно розробити програмне забезпечення, яке б дозволяло здійснювати непараметричне оцінювання функцій розподілу і щільності розподілу за МПВ, МЗН, МАЕФ, побудувати емпіричну функцію розподілу, а також проводити параметричне відновлення розподілів за класичними методами та оцінювати мінімальний обсяг вибірки, за якого параметричні методи забезпечують задану точність відновлення.

Основний матеріал. Для вирішення поставленої задачі створено програмне забезпечення «SmallSample» у середовищі Microsoft Visual Studio 2010 на мові програмування C#.

Структурними елементами програмного забезпечення є:

1. Обчислювальне ядро.
2. Блок роботи з даними, який забезпечує завантаження, моделювання та збереження вибірок.
3. Блок візуалізації, який реалізує представлення результатів у вигляді таблиць і графіків, а також забезпечує зручний інтерфейс, надаючи можливість аналізувати результати.

Ядро програмного забезпечення дозволяє проводити:

1. Непараметричне оцінювання функції розподілу класичним методом, тобто шляхом побудови емпіричної функції розподілу.
2. Непараметричне оцінювання функцій розподілу та щільності розподілу ймовірностей випадкової величини на основі МПВ, МЗН, МАЕФ [1].
3. Параметричне оцінювання функцій рівномірного, нормального, експоненціального, Релея та Вейбулла розподілів на основі методів максимальної правдоподібності (ММП), найменших квадратів (МНК), моментів (ММ) [2]. З метою ідентифікації передбачена побудова ймовірнісного паперу для кожного з вказаних типів розподілів.
4. Визначення мінімального обсягу вибірки, за якого параметричні методи забезпечують задану точність оцінювання функції розподілу.

5. Моделювання вибірок з розподілів рівномірного, нормального, експоненціального, Релея та Вейбулла з метою тестування зазначених вище методів оцінювання на вибірках малого обсягу [3].
6. Обчислення максимальної різниці між оцінкою функції розподілу та теоретичною (змодельованою) функцією.

Загальна схема роботи програми представлена на рис. 1.

Рис. 1. Блок схема роботи програми «SmallSample»

Створений програмний продукт має головне вікно, на якому розміщений блокнот із сімома закладками:

1. «Моделювання даних». На цій закладці можна задати параметри розподілу, який моделюється, обсяг вибірки та переглянути у таблиці змодельовану вибірку (рис. 2).
2. «Вар. ряд; ряд, розбитий на класи». Тут міститься таблиця з варіаційним рядом та рядом розбитим на класи, а також графік гістограми. На закладці можна задати кількість класів.
3. «Непараметричне відновлення». Тут міститься ще один блокнот, кожна закладка якого відповідає певному методу непараметричного

відновлення розподілу (класичний, МПВ, МЗН, МЕАФ). Для кожного методу на графіки виводяться оцінки функцій розподілу та щільності розподілу, а також теоретичні функції у разі моделювання (рис. 3)

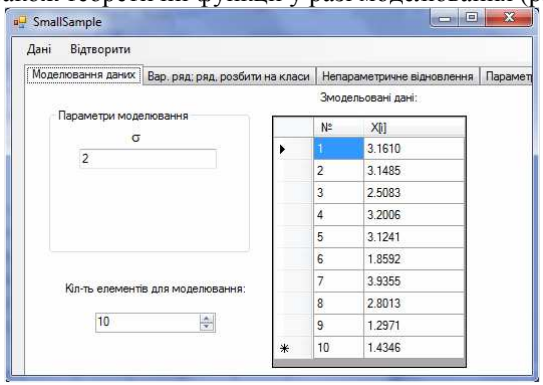


Рис. 2. Вигляд закладки «Моделювання даних»

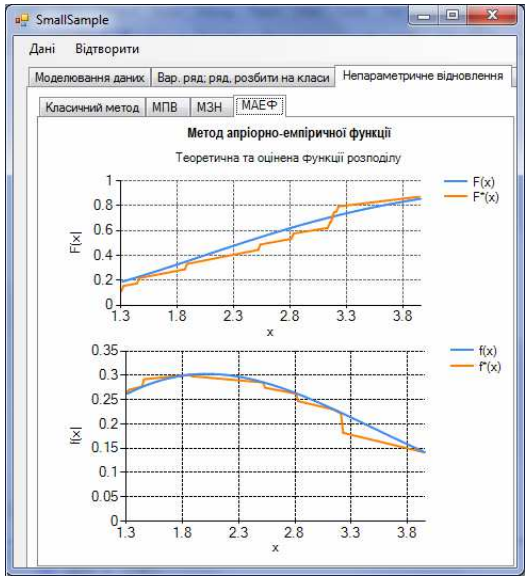


Рис. 3. Вигляд закладки «Непараметричне відновлення»

4. «Параметричне відновлення». На закладці розміщено таблицю, в яку виводяться знайдені оцінки параметрів та у разі моделювання даних значення параметра моделювання і його відхилення від оцінки (рис. 4). Також на закладці знаходиться графік з ймовірнісним

папером (будується папір для того типу розподілу, який обрано для відновлення) та графіки з функціями розподілу і щільності розподілу (оціненими та, у разі моделювання даних, теоретичними).

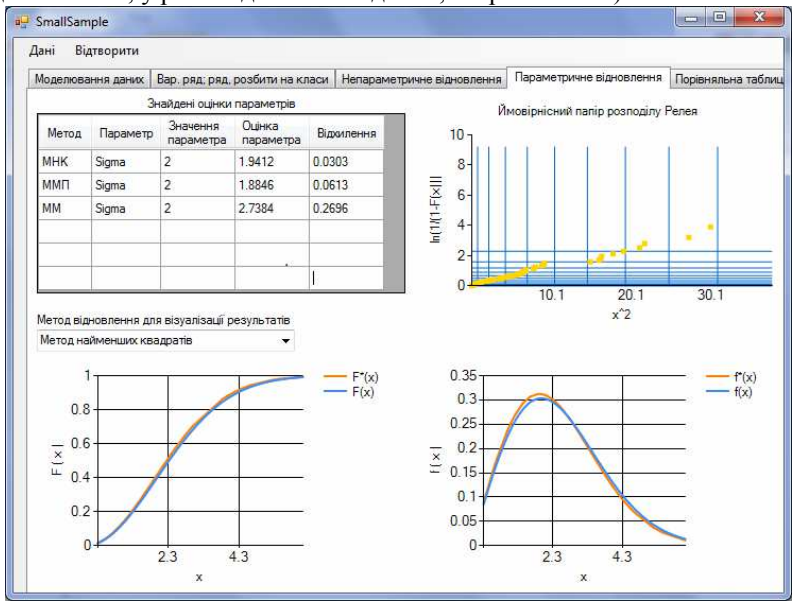


Рис. 4. Вигляд закладки «Параметричне відновлення»

5. «Порівняльна таблиця». Тут розміщена таблиця, до якої виводиться максимальна різниця між оціненими та теоретичними функціями розподілу і щільності розподілу. Таблиця заповнюється лише у випадку, коли вибірка була змодельована.

6. «Експеримент 1». На закладці при натисненні відповідної кнопки до таблиці виводяться результати наступного експерименту. Моделюється задана кількість вибірок вказаного обсягу з певного розподілу. За кожною вибіркою усіма непараметричними методами оцінюється функція розподілу, знаходиться максимальна різниця між оціненою функцією розподілу та теоретичною. За масивом максимальних різниць обчислюються середнє арифметичне та середньоквадратичне, які і виводяться у таблицю.

В якості прикладу нижче наведено результати одного з експериментів, під час яких моделювалось по 100 вибірок обсягом 5 з кожного розподілу (табл. 1). Через брак місця результати

представлено частково, лише для рівномірного, нормального та експоненціального розподілів.

Таблиця 1

Результати експерименту 1

Тип розподілу	Класичний метод	МПВ	МЗН	МАЕФ
Середнє арифметичне максимальної різниці				
Рівномірний	0,304	0,23	0,2	0,17
Нормальний	0,301	0,17	0,20	0,18
Експоненціальний	0,298	0,22	0,19	0,19
Середньоквадратичне максимальної різниці				
Рівномірний	0,120	0,095	0,077	0,060
Нормальний	0,109	0,087	0,073	0,063
Експоненціальний	0,118	0,110	0,072	0,058

Дані таблиці 1 свідчать, що найкращі результати дозволяє одержати метод апріорно-емпіричної функції. Висновок підтверджується і на інших типах розподілів та інших обсягах вибірок.

7. «Експеримент 2». На цій закладці при натисненні кнопки до таблиці виводяться результати такого експерименту. Для заданого типу розподілу моделюється задана кількість вибірок різного обсягу (мінімальний, максимальний обсяг та крок, з яким він має змінюватися, задаються). За кожною вибіркою здійснюється відновлення розподілу трьома параметричними методами і обчислюється максимальна різниця між оціненою функцією розподілу та теоретичною. За масивом максимальних різниць розраховується середнє, яке і виводяться до таблиці.

В якості прикладу наведено результати експериментів, під час яких моделювались по 100 вибірок з експоненціального та Релея розподілів (табл. 2). У таблиці через брак м'ясця представлено результати лише для п'яти обсягів. Одержані дані свідчать, що зі збільшенням обсягу вибірки максимальна різниця між оціненою та теоретичною функціями розподілу зменшується, що цілком природно. Задавшись точністю в 0,01, можна визначити мінімальний обсяг вибірки, за якого середнє значення максимальної різниці не перевищує цю точність. Для експоненціального розподілу ця умова починає виконуватися для вибірки обсягом 30 елементів. Для Релея у разі застосування методу максимальної правдоподібності мінімальний обсяг вибірки також 30 елементів, але для інших методів він вищий – 100 елементів. Результати експериментів на вибірках з різних розподілів засвідчили,

що під час застосування параметричних методів оцінювання розподілу, щоб максимальна різниця оціненої та теоретичної функцій розподілу була менше 0,01, обсяг вибірки має перевищувати 70–100 елементів.

Таблиця 2

Результати експерименту 2

Тип розподілу	Обсяг вибірки	ММП	МНК	ММ
Експоненціальний	10	0,0304	0,0518	0,0304
	15	0,0590	0,0445	0,0597
	25	0,0589	0,0267	0,0534
	30	0,0095	0,0076	0,0087
	50	0,0058	0,0048	0,0058
Релея	10	0,0340	0,0522	0,3616
	15	0,0323	0,0343	0,3305
	25	0,0167	0,0135	0,2567
	30	0,0093	0,0154	0,1120
	50	0,0052	0,0168	0,1871

Висновки. Створено програмне забезпечення «SmallSample» для відновлення розподілів за малими вибірками. Програмне забезпечення пройшло ретельне тестування на даних імітаційного моделювання, результати якого дозволяють рекомендувати його для використання у практичних задачах. Результати тестування методів непараметричного відновлення розподілів засвідчили, що найбільш точно оцінити функцію розподілу дозволяє метод апріорно-емпіричної функції.

Бібліографічні посилання

1. **Гаскаров Д.В.** Малая выборка / Д.В. Гаскаров, В.И. Шаповалов. – М., 1978. – 248 с.

2. **Бабак В.П.** Статистична обробка даних / В.П. Бабак, А.Я. Білецький, О.П. Приставка, П.О. Приставка. – К., 2001. – 388 с.

3. **Байбуз О.Г.** Системи масового обслуговування / О.Г. Байбуз, О.П. Приставка, П.О. Приставка. – Д., 2001. – 84 с.

Надійшла до редколегії 21.06.2012

УДК 519.24:519.652.3

П. О. Приставка, О. Г. Чолишкіна

Національний авіаційний університет

ОЦІНКА ПАРАМЕТРА ЕКСПОНЕНЦІАЛЬНОГО РОЗПОДІЛУ У ВИПАДКУ МАЛОЇ ВИБІРКИ

Запропоновано спосіб оцінки параметра експоненціального розподілу при обмеженій кількості даних експерименту. Показано переваги способу у порівняні з методом моментів на основі оцінки середнього арифметичного вибірки.

Ключові слова: *експоненційний розподіл, метод моментів, варіаційний ряд, мала вибірка.*

Предложен метод оценки параметра экспоненциального распределения при ограниченном количестве данных эксперимента. Показаны преимущества метода в сравнении с методом моментов на основе оценки среднего арифметического выборки.

Ключевые слова: *экспоненциальное распределение, метод моментов, вариационный ряд, малая выборка.*

The authors propose a method for estimating the parameters of the exponential distribution with a limited amount of experimental data. The advantage of the method in comparison with the method of moments based on the evaluation of arithmetic mean are shown.

Key words: *exponential distribution, the method of moments, variation series, small sample.*

Постановка проблеми. Питання про можливість аналізу після обробки обмежених обсягів статистичних даних є неоднозначним та суперечливим. Ефективність та спроможність оцінок на основі малих вибірок забезпечити практично неможливо, але, в той же час, потреба в результатах аналізу виникає в багатьох випадках практичної діяльності: дослідження рідкісних явищ, оцінка терміну активного існування високо надійних технічних систем, діагностика за певними видами захворювань, тощо. Тому винесення рекомендацій за технологіями опрацювання малих вибірок, зокрема, визначення нових підходів до оцінювання функції розподілу ймовіростей, було та залишається питанням актуальним.

Аналіз публікацій. У теорії перевірки статистичних гіпотез про однорідність вибірок питання винесення висновків на основі малих обсягів даних традиційно вирішується на основі t-тесту, f-тесту, критеріїв дисперсійного аналізу, непараметричних критеріїв (Манна-Уітні, H-критерій, тощо). Обґрунтуванню зазначених критеріїв приділяли увагу відомі вчені: У. Госсет (t-розподіл Стьюдента), Р. Фішер, Дж. Снедекор та ін. Для обробки малих обсягів дво- та багатовимірних даних, з метою встановлення наявності зв'язку між окремими ознаками, можна використовувати рангові коефіцієнти Спірмена, Кенделла, коефіцієнти таблиць сполучень.

Задача оцінки функції розподілу ймовірностей на разі малої кількості даних безпосередньо стосується винесення висновків про властивості генеральної сукупності, тож обсяг вибірки в першу чергу є визначальним з точки зору її репрезентативності.

У припущенні нормального закону теоретичної функції розподілу ймовірності випадкової величини, оцінка математичного сподівання приобсягах даних від 2 до 10 вирішується у середньому адекватно на підставі оцінки середнього арифметичного вибірки. Для оцінки середньоквадратичного σ досить використання формул, що наведено в таблиці [1, с.18], наприклад, при обсягах 3 та 4 даних мають місце співвідношення

$$\sigma = 0,5908\left(t_3 - t_1\right),$$

$$\sigma = 0,4539\left(t_4 - t_1\right) + 0,1102\left(t_3 - t_2\right),$$

де t_i, $i = \overline{1,4}$ – елементи варіаційного ряду.

Питанню непараметричного оцінювання функції розподілу ймовірності неперервної випадкової величини [2]. Показано, що оцінка емпіричної функції розподілу ймовірності, отримана на підставі методу вкладень та його модифікацій, є стійкою відносно статистичних характеристик вихідного ряду.

У [3; 4] оцінку функції щільності розподілу за малими вибірками запропоновано вирішувати на основі локальних поліноміальних сплайнів, близьких до інтерполяційних у середньому, на основі B-сплайнів [5; 6]. На основі моделювання з параметричних розподілів вибірок обмеженого обсягу (15 та більше даних) показано [4] переваги зазначених сплайнів при оцінюванні у середньому перших чотирьох моментів у порівнянні з методами прямокутних вкладень та апріорно-емпіричних функцій. Проте, варто відзначити, що не вдається на такому обсязі даних досягти прийнятної якості оцінювання при суттєвій лівій асиметрії теоретичних функцій щільності розподілу.

Поставимо за мету даної роботи провести дослідження питання оцінки експоненціального розподілу при обсягах даних до 10 елементів. Актуальними такі дослідження можуть бути при вирішенні задач теорії надійності та при розробці моделей систем масового обслуговування, зокрема, для оцінювання інтенсивностей потоків вимог та обслуговування.

Виклад основного матеріалу. Нехай за рівномірним розбиттям $\Delta_h : t_i = ih$ $\left(\tilde{\Delta}_h : t_i = (i+0,5)h\right)$, $i \in Z$, $h > 0$ осі реалізацій випадкової величини $\xi(\omega)$, на підставі вибірки $\Omega_{1,N} = \left\{t_l; l = \overline{1,N}\right\}$ проведено гістограмну оцінку, а отже одержано F_{1,N_i}, $t \in [t_i; t_{i+1})$, $i \in Z$ – масив емпіричної оцінки функції розподілу, тобто, будемо вважати, що на підставі $\Omega_{1,N}$ одержано масив

$$\Delta_h : \left\{t_i, F_{1,N_i}; i = \overline{0,m-1}\right\},$$

де m – кількість класів при гістограмній оцінці.

Тоді найпростішим, з точки зору реалізації у програмному забезпеченні обробки даних, при оцінюванні функції розподілу $F(t)$ для $\forall t \in [0; t_{\max}]$ (тут $t_{\max}$ – максимальне із зафіксованих спостережень) буде [6] застосування сплайну $S_{3,0}\left(F_{1,N}, t\right)$:

$$S_{3,0}\left(F_{1,N}, t\right) = \frac{1}{48}\left(-F_{1,N_{i-1}} + 3F_{1,N_i} - 3F_{1,N_{i+1}} + F_{1,N_{i+2}}\right)x^3 +$$
$$+\frac{1}{16}\left(F_{1,N_{i-1}} - F_{1,N_i} - F_{1,N_{i+1}} + F_{1,N_{i+2}}\right)x^2 +$$
$$+\frac{1}{16}\left(-F_{1,N_{i-1}} - 5F_{1,N_i} + 5F_{1,N_{i+1}} + F_{1,N_{i+2}}\right)x +$$
$$+\frac{1}{48}\left(F_{1,N_{i-1}} + 23F_{1,N_i} + 23F_{1,N_{i+1}} + F_{1,N_{i+2}}\right),$$

де $x = \frac{2}{h}(t - t_i) - 1$; $i = \left[\frac{t}{h}\right] + 1$; $[\square]$ – ціла частина.

Зважаючи на властивості функції розподілу, в якості невизначених значень $F_{1,N_{-2}}$, $F_{1,N_{-1}}$, F_{1,N_m}, достатньо узяти такі:

$$F_{1,N_{-2}} = 0, \quad F_{1,N_{-1}} = 0, \quad F_{1,N_m} = 1.$$

Для оцінювання функції експоненціального розподілу ймовірностей

$$F(t;\lambda) = \begin{cases} 0, & -\infty < t < 0, \\ 1 - \exp(-\lambda t), & 0 \le t < \infty, \end{cases} \quad (1)$$

згідно методу моментів, достатньо визначити оцінку параметр λ так:

$$\hat{\lambda} = \frac{1}{\overline{t}},$$

де

$$\overline{t} = \sum_{i=1}^{N} t_i \Big/ N \quad (2)$$

– оцінка математичного сподівання за вибіркою $\Omega_{1,N}$. Для сплайн-оцінювання математичного сподівання можна скористатись квадратурною формулою (наприклад, лівих прямокутників) для чисельного обрахунку за виразом для теоретичного моменту

$$\nu_1 = \int_0^{\infty} t\, dF(t),$$

де, в якості оцінки функції $F(t)$ виступає сплайн $S_{3,0}\left(F_{1,N}, t\right)$:

$$\hat{\nu}_1 = \sum_{j=1}^{n} t_j \left(S_{3,0}\left(F_{1,N}, t_j + \Delta t\right) - S_{3,0}\left(F_{1,N}, t_j\right)\right), \quad (3)$$

де t_j, $j = \overline{1,n}$ – деякі значення з інтервалу $[0; t_{\max}]$, узяті з рівномірним кроком Δt, досить малим, щоб забезпечити прийнятну точність вразу (2).

Іншим чином оцінку математичного сподівання можна отримати із (1). Зокрема, в точці оцінки теоретичного моменту $\tilde{\nu}_1^* \in [0, t_{\max})$, згідно оцінювання параметра λ в основі методу моментів, має виконуватись

$$F\left(\hat{\nu}_1^*, \frac{1}{\hat{\nu}_1^*}\right) = 1 - \exp\left(-\frac{1}{\hat{\nu}_1^*} \cdot \hat{\nu}_1^*\right),$$

або

$$F\left(\hat{\nu}_1^*, \frac{1}{\hat{\nu}_1^*}\right) = 1 - \exp(-1),$$

звідки оцінка $\hat{\nu}_1^*$ визначається як квантиль оцінки функції розподілу при рівні ймовірності $\alpha = 1 - \exp(-1)$, тобто

$$\hat{\nu}_1^* = \hat{F}^{-1}\left(\hat{\nu}_1^*,\ 1\Big/\hat{\nu}_1^*\right), \qquad (4)$$

де в якості оцінки $\hat{F}\left(\hat{\nu}_1^*,\ 1\Big/\hat{\nu}_1^*\right)$ пропонується використати непараметричну оцінку на основі сплайну $S_{3,0}\left(F_{1,N}, t\right)$. Визначення величини $\hat{\nu}_1^*$ на основі сплайн-оцінки неважко зробити, наприклад, на основі методу ділення відрізку навпіл.

Проведемо експериментальні дослідження оцінок (2)–(4) з метою визначення їх властивостей при малих обсягах даних вибірки $\Omega_{1,N}$. Опис схеми експерименту такий.

Крок 1. Випадковим чином визначаємо кількість даних вибірки від 4-х до 9-ти: $N \in [4; 9]$.

Крок 2. Моделюємо дані вибірки $\Omega_{1,N}$, розподілені за експоненціальним законом розподілу ймовірностей згідно виразу

$$t_l = \frac{1}{\lambda}\ln\frac{1}{1-random}, \qquad l = \overline{1, N},$$

де λ – параметр розподілу (для визначеності $\lambda = 5$); *random* – рівномірно розподілене псевдовипадкове число з інтервалу $[0;1)$.

Крок 3. Для змодельованої вибірки $\Omega_{1,N}$ проводиться формування варіаційного ряду та розбиття його на класи, причому кількість класів m гістограмної оцінки визначимо так:

$$m = \begin{cases} 3, & N \in [4;6], \\ 4 & N \in [7;9]. \end{cases}$$

Крок 4. Обраховуємо та зберігаємо для обробки оцінки (2) – (4).

Крок 5. Кроки 1 – 4 повторюємо 2000 разів для отримання представницького обсягу вбіркових статистик для подальшого аналізу.

Результати проведеного експерименту зведено до таблиці (табл.1) та наведено на графіках (рис.1 – 4).

Таблиця 1

Значення оцінок математичного сподівання за результатами експерименту

Оцінка	Середнє оцінки	Медіана оцінки	Теоретичне значення
$\overline{t}$	0,1241	0,1198	0,2
$\hat{\nu}_1$	0,12689	0,12246	
$\hat{\nu}_1^*$	0,17092	0,16304	

Зауважимо, що потреба в залучені до аналізу медіани оцінки математичного сподівання за результатами експерименту викликана асиметричністю функції щільності вибіркового розподілу статистик (2) – (4) що продемонстровано на графіках (рис.1–3).

Як видно з таблиці (табл.1), прийнятно оцінити математичне сподівання на основі виразів (2) та (3) у середньому неможливо. Поясненням цього є те, що для моделі експоненціального розподілу поява реалізації випадкової величини $\xi(\omega)$ на «хвості» розподілу – малоймовірна подія і для малих обсягів, що є предметом дослідження, зустрічається не часто, тобто, маємо заниження оцінки у порівнянні з теоретичним моментом. Проте, навіть, коли така подія має місце, це автоматично впливає на оцінку, шляхом її суттєвого завищення (правий хвіст функції щільності на графіках).

Для оцінки за виразом (4) ситуація порівняно краща. Приблизно чверть значень оцінок, отриманих у ході експерименту, розташовано в межах від 0,17 до 0,23 і майже половина значень у межах від 0,14 до 0,26.

Якщо проаналізувати залежність оцінювання за виразом (4) від обсягу даних вибірки (рис. 4), то видно, що регресія оцінки $\hat{\nu}_1^*$, залежно від обсягу даних вибірки має лінійну модель і для кількості від 7-ми до 9-ти практично відповідає теоретичному значенню математичного сподівання 0,2.

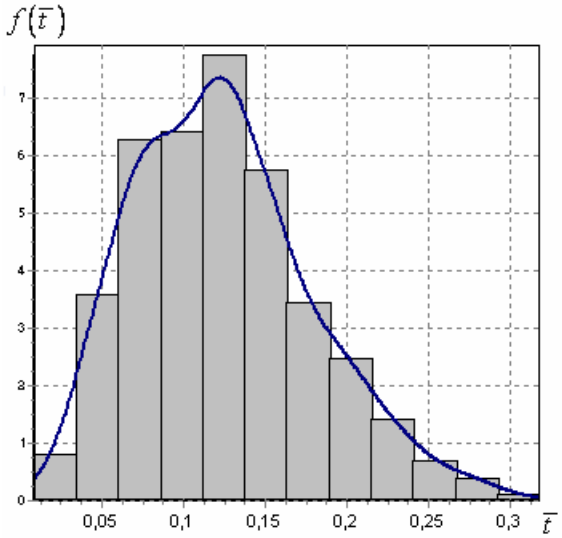


Рис. 1. Функція щільності та нормована гістограма оцінки (2) за результатами експерименту

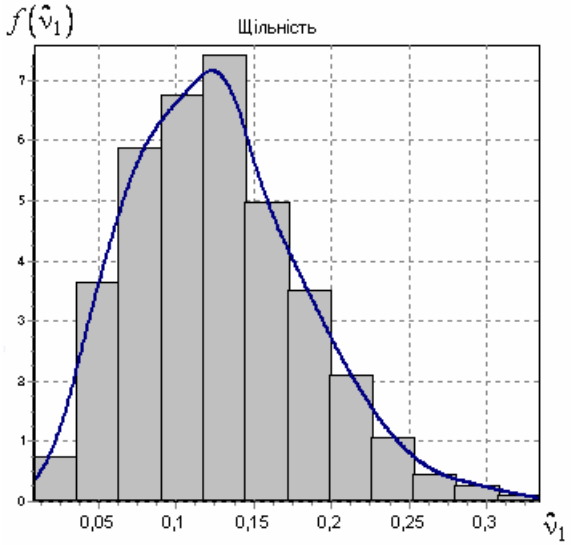


Рис. 2. Функція щільності та нормована гістограма оцінки (3) за результатами експерименту

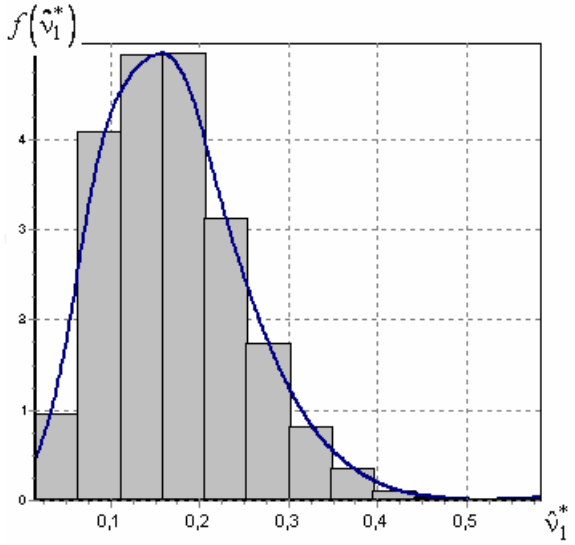


Рис. 3. Функція щільності та нормована гістограма оцінки (4) за результатами експерименту

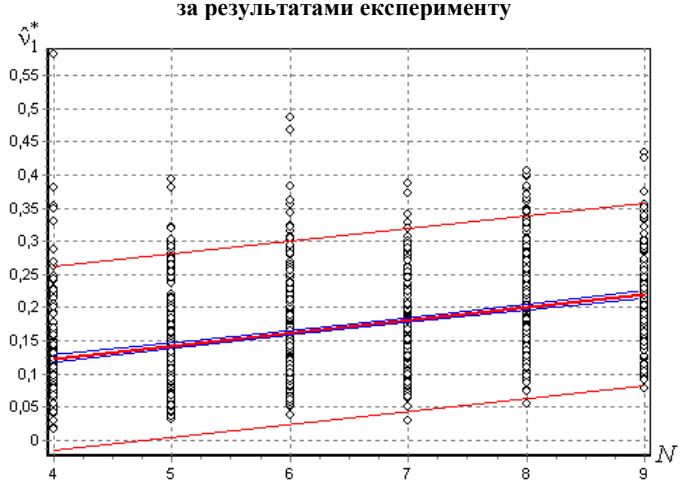


Рис. 4. Регресія з довірчим інтервалом та толерантними межами оцінки $\hat{\nu}_1^*$, залежно від обсягу даних вибірки

Висновки. На основі результатів експериментів з імітаційного моделювання (табл.1) показано, що при оцінюванні параметра λ в моделі експоненціального розподілу за методом моментів спостерігається завищення такої оцінки. Пов'язана така ситуація з тим, що в малих вибірках значення з «хвостів» експоненціального розподілу зустрічаються нечасто, отже, оцінка математичного сподівання, що є оберненою величиною оцінки $\hat{\lambda}$, у більшості випадків експериментів є заниженою (рис.1).

Таким чином, якщо в якості оцінки математичного сподівання обирати оцінку середнього вибірки, результат буде не адекватним. Аналогічна ситуація при використанні оцінки теоретичного моменту на основі непараметричної сплайн-оцінки функції розподілу (рис.2). Більш адекватним може бути оцінка на основі виразу (4), яка побудована на оцінці квантилі, що отримана за використанням локального поліноміального сплайну а основі *B*-сплайнів третього порядку, близького до інтерполяційного у середньому (рис.3 – 4).

Подальше дослідження може полягати в дослідженні запропонованого підходу при обробці даних у задачах моделювання систем масового обслуговування та при вирішенні задач оцінки терміну активного існування високо надійних технічних систем.

Бібліографічні посилання

1. **Бабак В.П.** Статистична обробка даних / В.П. Бабак, А.Я. Білецький, О.П. Приставка, П.О. Приставка. – К., 2001. – 388 с.
2. **Гаскаров Д.В.** Малая выборка. / Д.В. Гаскаров, В.И. Шаповалов. – М., 1978. – 248 с.
3. **Приставка Ф.А.** Непараметрическая оценка плотности вероятностей по малым выборкам / Ф.А. Приставка // Придніпровський науковий вісник. – 1998. – № 101 (168). – С. 25–29.
4. **Приставка П.О.** Обробка вибірок обмеженого обсягу з використанням поліноміальних сплайнів / П.О. Приставка, О.О. Смойловська // Актуальні проблеми автоматизації та інформаційних технологій: зб. наук. праць. – 2001.– Т.4. – С. 86–95.
5. **Лигун А.А.** Асимптотические методы восстановления кривых. / А.А. Лигун, А.А. Шумейко. – К., 1996. – 358 с.
6. **Приставка П.О.** Поліноміальні сплайни при обробці даних / П.О. Приставка. – Д., 2004. – 236 с.

Надійшла до редколегії 10.06.2012

УДК 519.652:519.254

П.О.Приставка

Національний авіаційний університет

ПРОЦЕДУРИ НЕБІНАРНОГО *SUBDIVISION* НА ОСНОВІ ЛІНІЙНИХ КОМБІНАЦІЙ B-СПЛАЙНІВ, БЛИЗЬКИХ ДО ІНТЕРПОЛЯЦІЙНИХ У СЕРЕДНЬОМУ

Запропоновано загальну постановку задачі на проведення небінарного *subdivision* на основі локальних сплайнів, наведено відповідні лінійні оператори, визначено місцезнаходження відліків для проектування числових послідовностей.

Ключові слова: *небінарний subdivision, локальний сплайн, апроксимація, дискретна згортка.*

Предложена общая постановка задачи на проведение небинарного *subdivision* на основе локальных сплайнов, приведены соответствующие линейные операторы, определено метонахождение отсчетов для проектирования числовых последовательностей.

Ключевые слова: *небинарный subdivision, локальный сплайн, аппроксимация, дискретная свертка.*

A general statement of the problem of conducting non-binary *subdivision* from local splines is proposed. The relevant linear operators are given and the location of indications to design numerical sequences is defined.

Key words: *non-binary subdivision, local spline, approximation, discrete compression.*

Постановка проблеми. Стала тенденція до збільшення обсягів даних, що обробляються, на сьогодні є об'єктивною реальністю. Серед інформації, котра зберігається в електронному вигляді є певна категорія даних, які потребують при обробці математичних процедур, що побудовані на основі методів інтерполяції: цифрові зображення та відео, часові ряди та цифрові сигнали, комп'ютерні анімаційні та інші просторові моделі, тощо. Розробники відповідного програмного забезпечення, що працюють з такими даними, давно використовують швидкодіючі процедури, засновані на локальній інтерполяції (типово – *B*-сплайни), проте, є необхідність враховувати потребу у зменшенні кількості обчислювальних операцій при побудові апроксимацій.

Дані, про які мова, в електронному вигляді представлені у вигляді послідовностей відліків (сигналів, тощо), заданих у вузлових точках.

Зміна кількості відліків у послідовностях за довільну кількість раз (необов'язково у цілочисельну) можна забезпечити на основі неперервних інтерполяцій. Але, якщо зміна масштабу послідовності здійснюється на певний наперед відомий коефіцієнт, то в цьому разі для апроксимацій, що мають явний вигляд, можна отримати обчислювальні процедури з меншою обчислювальною складністю, як часткові випадки. Як приклад – *subdivision*-процедури, для котрих отримання нових та вдосконалення алгоритмізації існуючих є задачею актуальною.

Аналіз досліджень та постановка задачі. З точки зору постановки задачі, застосування апроксимацій гладких функцій є спрощенням при виборі методу неперервної апроксимації, придатного для зміни кількості членів цифрових послідовностей. Таке спрощення не є грубим на разі використання методів на основі фінітних функцій, зокрема *B*-сплайнів. Вагомий внесок до фундаментальної розробки апарату апроксимацій на основі *B*-сплайнів внесли І. Шоенберг, К. Де Бор, М. П. Корнійчук, А. О. Лигун [1 – 3] та ін. Теоретичні та практичні дослідження сплайнів на основі *B*-сплайнів, близьких до інтерполяційних у середньому, подано, наприклад, в [4; 5]. Стосовно останнього типу сплайнів можна відмітити, що вибір в якості апарату апроксимації операторів, що є близькими до інтерполяційних у середньому, обумовлений більш високою стійкістю оцінки наближення за даними, що є різного роду результатами вимірювань та можуть бути використані в якості моделі аналогового сигналу [6].

Нехай з кроком $h>0$ задано розбиття дійсної вісі $\Delta_h : t_i = ih$, $i \in \mathbf{Z}$, у кожній точці якого отримано значення деякої неперервної функції $p(t)$, визначеної на $\mathbf{R}_1(-\infty;\infty)$. Вважаємо, що інформацію про функцію $p(t)$ задано у вузлах розбиття Δ_h у вигляді інтеграла

$$\overline{p}_i = \frac{1}{h}\int\limits_{(i-0,5)h}^{(i+0,5)h} p(t)\,dt, \qquad (1)$$

при цьому, істинне значення функції $p(t)$ у вузлах таке:

$$p_i = \overline{p}_i + \varepsilon_i,\ i \in \mathbf{Z},$$

де ε_i – похибка.

Для апроксимації функції $p(t)$ за значеннями типу (1) у вузлах розбиття Δ_h, застосовують поліноміальні сплайни, що є близькими до інтерполяційних у середньому на основі *B*-сплайнів другого – п'ятого порядків [4; 5; 7]. Наприклад, сплайн

$$S_{2,1}(p,t) = \sum_{i\in Z}\left(p_i - \frac{1}{6}\Delta^2 p_i\right) B_{2,h}\left(t-(i+0,5)h\right), \qquad (2)$$

де

$$B_{2,h}(t) = \begin{cases} 0, & t \notin [-3h/2;3h/2], \\ (3+2t/h)^2/8, & t \in [-3h/2;-h/2], \\ 3/4-(2t/h)^2/4, & t \in [-h/2;h/2], \\ (3-2t/h)^2/8, & t \in [h/2;3h/2]; \end{cases}$$

$$\Delta^2 p_i = p_{i-1} - 2p_i + p_{i+1},$$

має високу якість апроксимації, зокрема, при $h \to 0$ для довільної $p(t) \in C^3$ буде виконуватись

$$\left\| p(t) - S_{2,1}(\overline{p},t)\right\| = \frac{h^3}{12\sqrt{3}}\left\| p^{(3)}(t)\right\| + \mathrm{o}\left(h^3\right).$$

З іншого боку, наприклад, для сплайна

$$S_{4,0}(p,t) = \sum_{i\in Z} p_i B_{4,h}\left(t-(i+0,5)h\right), \qquad (3)$$

де

$$B_{4,h}(t) = \begin{cases} 0, & t \notin [-5h/2;5h/2], \\ (t/h+5/2)^4/24, & t \in [-5h/2;-3h/2], \\ \Psi(t/h)-5(t/h+1/2)^4/12, & t \in [-3h/2;-h/2], \\ \Psi(t/h), & t \in [-h/2;h/2], \\ \Psi(t/h)-5(t/h-1/2)^4/12, & t \in [h/2;3h/2], \\ (t/h-5/2)^4/24, & t \in [3h/2;5h/2]; \end{cases}$$

$$\Psi(t) = \frac{115}{192} - \frac{5}{8}t^2 + \frac{1}{4}t^4,$$

притаманна властивість згладжування: при $h \to 0$ для довільної функції $p(t) \in C^2$ справедливо таке:

$$\left\| p(t) - S_{4,0}(\bar{p},t) \right\| = \frac{h^2}{4} \left\| p''(t) \right\| + \mathrm{o}\left(h^2\right).$$

Отже, при застосуванні того, чи іншого типу сплайна, залежно від потреби, можна отримувати або асимптотично точні оцінки, або ж забезпечувати згладжування локальних осциляцій в даних. При обчисленнях сплайни зручно подавати в явному вигляді, наприклад, вирази (2) та (3) можна записати так:

$$S_{2,1}(p,t) = \frac{1}{48}\Big(-(1-x)^2 p_{i-2} + \left(2-16x+10x^2\right) p_{i-1} +$$
$$+\left(46-18x^2\right) p_i + \left(2+16x+10x^2\right) p_{i+1} - (1+x)^2 p_{i+2}\Big), \qquad (4)$$

$$S_{4,0}(p,t) = \frac{1}{384}\Big((1-x)^4 p_{i-2} + \left(76-88x+24x^2 +\right.$$
$$\left.+8x^3-4x^4\right) p_{i-1} + \left(230-60x^2+6x^4\right) p_i +$$
$$+\left(76+88x+24x^2-8x^3-4x^4\right) p_{i+1} + (1+x)^4 p_{i+2}\Big), \qquad (5)$$

де $x = \frac{2}{h}\left(t-(i+0{,}5)h\right)$, $|x| \le 1$.

Перехід від неперервних наближень на основі лінійних комбінацій *B*-сплайнів до процедур *subdivision* викликає практичний інтерес майже протягом останніх тридцяти років. Наприклад у [8] є чимало подібних процедур, при цьому бібліографія складає 180 робіт. Автори [9] започаткували цілу низку публікацій, які пропонують підходи до побудови *subdivision* що базуються на локальних сплайнах на основі *B*-сплайнів, близьких до інтерполяційних у середньому, наприклад [10]. У [11; 12] отримані лінійні оператори небінарного масштабування одно- та двовимірних послідовностей відліків гладких функцій, які дозволяють здійснювати вказану операцію як з вимогою згладжування, так і асимптотично точно. Проте, загальної постановки задачі для процедур небінарного *subdivision* зроблено не було. Обумовлено це тим, що вибір конкретних часткових випадків локальних сплайнів, близьких до інтерполяційних у середньому, потребує автоматизованого визначення вузлів інтерполювання. Тож, це питання потребує узагальнення, що й визначає мету даної публікації.

Виклад основного матеріалу. Нехай у вузлах розбиття

$$\Delta_{h_M} : t_j = jh_M,\ h_M > 0$$

задано $P = \left\{p_{j,\kappa}\right\}_{j\in\mathbf{Z}}$, $\kappa = 1,2,\ldots$ – послідовність відліків гладкої функції $p(t)$. Під κ будемо розуміти ітераційний крок небінарного масштабування проектування послідовності P. Тоді, якщо $\left\{p_{i,\kappa-1}\right\}_{i\in\mathbf{Z}}$ – початкова послідовність, визначена у вузлах розбиття

$$\Delta_{h_N} : t_i = ih_N,\ h_N > 0,$$

а $\left\{p_{j,\kappa}\right\}_{j\in\mathbf{Z}}$ – утворена, шляхом «проектування» будь-яких N послідовних точок в M, то маємо визначення для $p_{j,\kappa}$, наведене нижче.

Для довільного індексу i зафіксуємо N вузлів $\left\{t_{i+u,\kappa-1}\right\}_{\substack{i\in\mathbf{Z}_l \\ u=\overline{0,N-1}}}$ в яких визначено значення $\left\{p_{i+u,\kappa-1}\right\}_{\substack{i\in\mathbf{Z}_l \\ u=\overline{0,N-1}}}$. Суть проектування N точок до M полягає в тому, що на кожній ітерації в межах інтервалу

$$\left[-h_N/2 + ih_N,\ ih_N + (N-1)h_N + h_N/2\right]$$

визначаємо нові вузли з рівномірним кроком $h_M > 0$.

На інтервалі

$$\left[-h_M/2 + jh_M,\ jh_M + (M-1)h_M + h_M/2\right],$$

де

$$-h_N/2 + ih_N = -h_M/2 + jh_M;$$
$$h_M = \frac{N}{M} h_N.$$

будемо визначати члени послідовності $P = \left\{p_{j,\kappa}\right\}_{j\in\mathbf{Z}}$ на основі лінійних операторів, що базуються на даних попереднього кроку рекурсії:

$$p_{j+l,\kappa} = A_l\left(p^{\kappa-1,i}\right),\ l = \overline{0,M-1},$$

де $A_l\left(p^{\kappa-1,i}\right)$ неважко отримати як часткові випадки сплайнів на зразок (2), (3).

У загальному випадку для конкретного l знаходимо величини

$$u^* = \frac{N(2l+1)}{2M}, \qquad u = \left[u^*\right], \text{ де } [\square] - \text{ ціла частина,}$$

$$x_l = 2\left(u^* - u\right) - 1$$

та підставляємо до явної формули для сплайна, визначеного на розбитті Δ_{h_N}. Наприклад, з виразу (4) отримуємо

$$p_{j+l,\kappa} = \frac{1}{48}\Big(-\left(1-x_l\right)^2 p_{i+u-2,\kappa-1} + \left(2-16x_l+10x_l^2\right) p_{i+u-1,\kappa-1} + \tag{6}$$

$$+\left(46-18x_l^2\right) p_{i+u,\kappa-1} + \left(2+16x_l+10x_l^2\right) p_{i+u+1,\kappa-1} - \left(1+x_l\right)^2 p_{i+u+2,\kappa-1}\Big).$$

Для зменшення обчислювальної складності пропонується замість формул на зразок (6) використовувати наперед визначені часткові випадки сплайнів при конкретних значеннях аргумента x. Наприклад, при $x_l = 1/3$ вираз (6) перепишеться так:

$$p_{j+l,\kappa} = \frac{1}{108}\left(-p_{i+u-2,\kappa-1} - 5p_{i+u-1,\kappa-1} + 99p_{i+u,\kappa-1} + 19p_{i+u+1,\kappa-1} - 4p_{i+u+2,\kappa-1}\right)$$

або

$$p_{j+l,\kappa} = \sum_{ii=i+u-2}^{i+u+2} \gamma^{(2,1)}_{\left(\frac{1}{3}\right)ii-i+u} \cdot p_{ii,\kappa-1},$$

де $\gamma^{(2,1)}_{\left(\frac{1}{3}\right)} = \left(-1 \quad -5 \quad 99 \quad 19 \quad -4\right)^T / 108$.

Загалом, для часткових випадків локальних поліноміальних сплайнів парних ступенів на основі *B*-сплайнів, близьких до інтерполяційних у середньому [4; 5], має місце загальна формула для визначення лінійних операторів небінарного *subdivision*

$$p_{j+l,\kappa} = \sum_{ii=i+u-w}^{i+u+w} \gamma^{(r,v)}_{(x)ii-i+u} \cdot p_{ii,\kappa-1}, \tag{7}$$

де $\gamma^{(r,v)}_{(x)}$ – маска оператора дискретної згортки; $2 \cdot w + 1$ – ширина маски; r – порядок сплайна; v – порядок уточнення сплайна; x – часткове значення аргументу в явному поданні сплайну.

Наприклад, для x, що може набувати значення із множини

$$\left\{-1, -\frac{4}{5}, -\frac{3}{4}, -\frac{2}{3}, -\frac{3}{5}, -\frac{1}{2}, -\frac{2}{5}, -\frac{1}{3}, -\frac{1}{4}, -\frac{1}{5}, 0, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, 1\right\}, \tag{8}$$

мають місце такі маски, отримані з часткових випадків виразу (4):

$$\gamma^{(2,1)}_{(-1)} = \left(-1 \quad 7 \quad 7 \quad -1 \quad 0\right)^T / 12;$$

$$\gamma^{(2,1)}_{\left(-\frac{4}{5}\right)} = \left(-81 \quad 530 \quad 862 \quad -110 \quad -1\right)^T / 1200;$$

$$\gamma^{(2,1)}_{\left(-\frac{3}{4}\right)} = \left(-49 \quad 314 \quad 574 \quad -70 \quad -1\right)^T / 768;$$

$$\gamma^{(2,1)}_{\left(-\frac{2}{3}\right)} = \left(-25 \quad 154 \quad 342 \quad -38 \quad -1\right)^T / 432;$$

$$\gamma^{(2,1)}_{\left(-\frac{3}{5}\right)} = \left(-16 \quad 95 \quad 247 \quad -25 \quad -1\right)^T / 300;$$

$$\gamma^{(2,1)}_{\left(-\frac{1}{2}\right)} = \left(-9 \quad 50 \quad 166 \quad -14 \quad -1\right)^T / 192;$$

$$\gamma^{(2,1)}_{\left(-\frac{2}{5}\right)} = \left(-49 \quad 250 \quad 1078 \quad -70 \quad -9\right)^T / 1200;$$

$$\gamma^{(2,1)}_{\left(-\frac{1}{3}\right)} = \left(-4 \quad 19 \quad 99 \quad -5 \quad -1\right)^T / 108;$$

$$\gamma^{(2,1)}_{\left(-\frac{1}{4}\right)} = \left(-25 \quad 106 \quad 718 \quad -22 \quad -9\right)^T / 768;$$

$$\gamma^{(2,1)}_{\left(-\frac{1}{5}\right)} \left(-9 \quad 35 \quad 283 \quad -5 \quad -4\right)^T / 300;$$

$$\gamma^{(2,1)}_{(0)} = \left(-1 \quad 2 \quad 46 \quad 2 \quad -1\right)^T / 48;$$

$$\gamma^{(2,1)}_{\left(\frac{1}{5}\right)} = \left(-4 \quad -5 \quad 283 \quad 35 \quad -9\right)^T / 300;$$

$$\gamma^{(2,1)}_{\left(\frac{1}{4}\right)} = \left(-9 \quad -22 \quad 718 \quad 106 \quad -25\right)^T / 768;$$

$$\gamma^{(2,1)}_{\left(\frac{1}{3}\right)} = \left(-1 \quad -5 \quad 99 \quad 19 \quad -4\right)^T / 108;$$

$$\gamma^{(2,1)}_{\left(\frac{2}{5}\right)} = \left(-9 \quad -70 \quad 1078 \quad 250 \quad -49\right)^T / 1200;$$

$$\gamma^{(2,1)}_{\left(\frac{1}{2}\right)} = \left(-1 \quad -14 \quad 166 \quad 50 \quad -9\right)^T / 192;$$

$$\gamma^{(2,1)}_{\left(\frac{3}{5}\right)} = \left(-1 \quad 25 \quad 247 \quad 95 \quad -16\right)^T / 300;$$

$$\gamma^{(2,1)}_{\left(\frac{2}{3}\right)} = \left(-1 \quad -38 \quad 342 \quad 154 \quad -25\right)^T / 432;$$

$$\gamma_{\left(\frac{3}{4}\right)}^{(2,1)} = \left(-1 \quad -70 \quad 574 \quad 314 \quad -49\right)^T / 768;$$

$$\gamma_{\left(\frac{4}{5}\right)}^{(2,1)} = \left(-1 \quad -110 \quad 862 \quad 530 \quad -81\right)^T / 1200;$$

$$\gamma_{(1)}^{(2,1)} = \left(0 \quad -1 \quad 7 \quad 7 \quad -1\right)^T / 12.$$

Для сплайна (5), коли x може набувати значення із множини (8), мають місце наступні маски:

$$\gamma_{(-1)}^{(4,0)} = \left(1 \quad 11 \quad 11 \quad 1 \quad 0\right)^T / 24;$$

$$\gamma_{\left(-\frac{4}{5}\right)}^{(4,0)} = \left(6561 \quad 97516 \quad 121286 \quad 14636 \quad 1\right)^T / \left(24 \cdot 10^4\right);$$

$$\gamma_{\left(-\frac{3}{4}\right)}^{(4,0)} = \left(2401 \quad 38620 \quad 50726 \quad 6556 \quad 1\right)^T / 98304;$$

$$\gamma_{\left(-\frac{2}{3}\right)}^{(4,0)} = \left(625 \quad 11516 \quad 16566 \quad 2396 \quad 1\right)^T / 31104;$$

$$\gamma_{\left(-\frac{3}{5}\right)}^{(4,0)} = \left(256 \quad 5281 \quad 8171 \quad 1291 \quad 1\right)^T / 15000;$$

$$\gamma_{\left(-\frac{1}{2}\right)}^{(4,0)} = \left(81 \quad 1996 \quad 3446 \quad 620 \quad 1\right)^T / 6144;$$

$$\gamma_{\left(-\frac{2}{5}\right)}^{(4,0)} = \left(2401 \quad 71516 \quad 137846 \quad 28156 \quad 81\right)^T / \left(24 \cdot 10^4\right);$$

$$\gamma_{\left(-\frac{1}{3}\right)}^{(4,0)} = \left(16 \quad 545 \quad 1131 \quad 251 \quad 1\right)^T / 1944;$$

$$\gamma_{\left(-\frac{1}{4}\right)}^{(4,0)} = \left(625 \quad 21212 \quad 57926 \quad 18460 \quad 81\right)^T / 98304;$$

$$\gamma_{\left(-\frac{1}{5}\right)}^{(4,0)} = \left(81 \quad 3691 \quad 8891 \quad 2321 \quad 16\right)^T / 15000;$$

$$\gamma_{(0)}^{(4,0)} = \left(1 \quad 76 \quad 230 \quad 76 \quad 1\right)^T / 384;$$

$$\gamma_{\left(\frac{1}{5}\right)}^{(4,0)} = \left(16 \quad 2321 \quad 8891 \quad 2321 \quad 81\right)^T / 15000;$$

$$\gamma_{\left(\frac{1}{4}\right)}^{(4,0)} = \left(81 \quad 18460 \quad 57926 \quad 21212 \quad 625\right)^T / 98304;$$

$$\gamma_{\left(\frac{1}{3}\right)}^{(4,0)} = \left(1 \quad 251 \quad 1131 \quad 545 \quad 16\right)^T / 1944;$$

$$\gamma_{\left(\frac{2}{5}\right)}^{(4,0)} = \left(81 \quad 28156 \quad 137846 \quad 71516 \quad 2401\right)^T / \left(24 \cdot 10^4\right);$$

$$\gamma_{\left(\frac{1}{2}\right)}^{(4,0)} = \left(1 \quad 620 \quad 3446 \quad 1996 \quad 81\right)^T / 6144;$$

$$\gamma_{\left(\frac{3}{5}\right)}^{(4,0)} = \left(1 \quad 1291 \quad 8171 \quad 5281 \quad 256\right)^T / 15000;$$

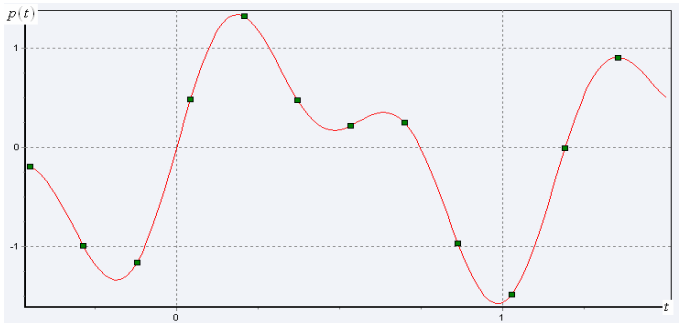
$$\gamma_{\left(\frac{2}{3}\right)}^{(4,0)} = \left(1 \quad 2396 \quad 16566 \quad 11516 \quad 625\right)^T / 31104;$$

$$\gamma_{\left(\frac{3}{4}\right)}^{(4,0)} = \left(1 \quad 6556 \quad 50726 \quad 38620 \quad 2401\right)^T / 98304;$$

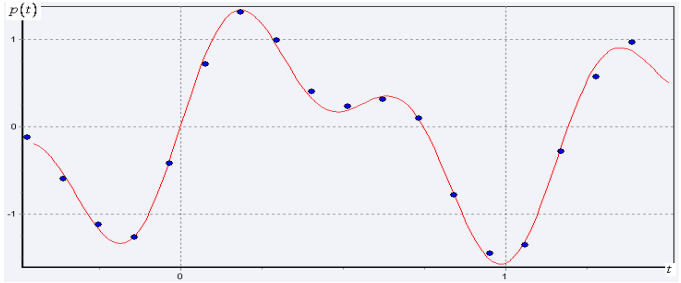
$$\gamma_{\left(\frac{4}{5}\right)}^{(4,0)} = \left(1 \quad 14636 \quad 121286 \quad 97516 \quad 6561\right)^T / \left(24 \cdot 10^4\right);$$

$$\gamma_{(1)}^{(4,0)} = \left(0 \quad 1 \quad 11 \quad 11 \quad 1\right)^T / 24.$$

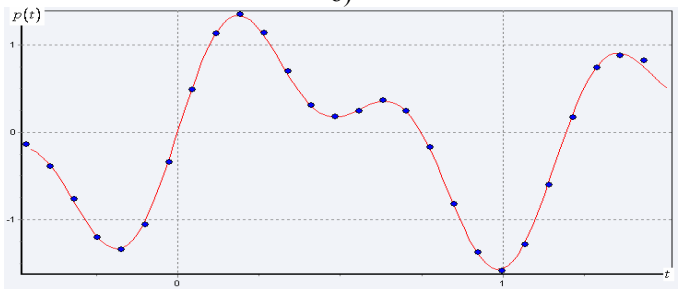
На графіках (рис.1) подано приклад небінарного *subdivision* при проектуванні кожних 2-х послідовних точок у 3-і. В якості лінійних операторів масштабування брались часткові випадки сплайну (4) відповідно в точках $x = -1/3$, $x = -1$, $x = 1/3$, котрим, згідно виразу (7) відповідають маски $\gamma_{\left(-\frac{1}{3}\right)}^{(2,1)}$, $\gamma_{(-1)}^{(2,1)}$, $\gamma_{\left(\frac{1}{3}\right)}^{(2,1)}$.



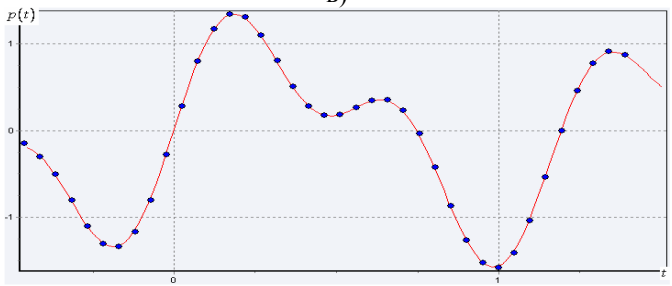
а)



б)



в)



г)

Рис.1. Небінарний *subdivision* «2 в 3» на основі сплайна (4): а) функція $p(t)=\sin(5t)+0,6\sin(11t)$, $t\in[-0,45;1,52]$ та її початкові відліки; б) *subdivision* при $\kappa=1$; в) *subdivision* при $\kappa=2$; г) *subdivision* при $\kappa=3$

Висновки. У роботі запропоновано загальну постановку задачі на проведення небінарного *subdivision* на основі лінійних комбінацій *B*-сплайнів, близьких до інтерполяційних у середньому. Викладено загальний спосіб визначення місцезнаходження відліків для проектування числових послідовностей. Наведено часткові випадки

сплайн-операторів у вигляді дискретної згортки послідовності відліків та відповідних масок.

Подальші дослідження можуть бути спрямовані на оцінку якості наближення функцій спостережень, при умові, коли данні задано з різними типами завад. Вартим уваги може бути питання узагальнення поданих процедур не бінарного *subdivision* на випадок багатовимірних послідовностей.

Бібліографічні посилання

1. **Де Бор К.** Практическое руководство по сплайнам / К. Де Бор. – М., 1985. – 303 с.

2. **Роджерс Д.** Математические основы машинной графики: пер. с англ. / Д. Роджерс, Дж. Адамс – М., 2001. – 604 с.

3. **Чуи Ч.** Введение в вэйвлеты: Пер. с агнгл. / Ч. Чуи – М., 2001. – 412 с.

4. **Лигун А.А.** Асимптотические методы восстановления кривых. / А.А. Лигун, А.А. Шумейко. – К., ІМ НАН України, 1996. – 358 с.

5. **Приставка П.О.** Поліномаільні сплайни при обробці даних / П.О. Приставка. – Д., 2004. – 236 с.

6. **Приставка П.О.** Лінійні комбінації *B*-сплайнів, близькі до інтерполяційних у середньому, в задачі моделювання аналогових сигналів / П.О. Приставка // Актуальні проблеми автоматизації та інформаційних технологій: зб. наук. праць. – 2011. – Т.15. – С. 4–17.

7. **Приставка П.О.** Дослідження *B*-сплайну п’ятого порядку та їх лінійної комбінації / П.О. Приставка, О.Г. Чолишкіна / Математичне моделювання. – 2007. – № 1 (16). – С. 14–17.

8. **Andersson L.-E.,** Stewart N. Introduction to the mathematics of subdivision surfaces. – Philadelphia: Society for Industrial and Applied Mathematics, 2010. – 356 p.

9. **Лигун А.А.** Исследования линейных операторов, порожденных методами пополнения данных / А.А. Лигун, А.А. Шумейко // Математичне моделювання. – 2000, 2 (5), С. 11–19.

10. **Приставка П.О.** Поліноміальні сплайни в задачах бінарного поповнення / П.О. Приставка // Актуальні проблеми автоматизації та інформаційних технологій. – 2003. – Т. 7. –С. 39–53.

11. **Приставка П.О.** Небінарне поповнення послідовностей відліків гладких функцій лінійними операторами на основі поліноміальних сплайнів / П.О. Приставка // Вісник НАУ. – 2008. – № 3. – С. 85–89.

12. **Приставка П.О.** Небінарне поповнення послідовностей відліків гладких функцій двох змінних лінійними операторами / П.О. Приставка // Вісник НАУ. – 2009. – № 2. – С. 173–177.

Надійшла до редколегії 11.07.2012

УДК 681.3

Т.А. Рузова

Днепропетровский национальный университет им. Олеся Гончара

МОДЕЛЬНЫЕ АГРЕГАТЫ ДИСПЕРСИЙ. ДЕКОМПОЗИЦИЯ ПО СТРУКТУРЕ

Запропоновано метод сегментації агрегованих елементів дисперсних утворень сферичної форми (крапель емульсій), заснований на інформації про структуру й контури агрегатів. Метод може бути використаний при розробці систем виміру й аналізу емульсій та інших мікрооб'єктів.

Ключові слова: *кістяк зображення, агрегат, дисперсні утворення, сегментація, точки приєднання частинок.*

Предложен метод сегментации агрегированных элементов дисперсных образований сферической формы (капель эмульсий), основанный на информации о структуре и контурах агрегатов. Метод может быть использован при разработке систем измерения и анализа эмульсий и других микрообъектов.

Ключевые слова: *скелет изображения, агрегат, дисперсные образования, сегментация, точки присоединения частиц.*

There is developed method for segmentation of aggregated structures of spherical particles (emulsion drops) in dispersive formations. Method is based on information of aggregates structure and contours. Method may be used to design systems for emulsions (and other micro objects) measuring and analyzing.

Key words: *image skeleton, aggregate, dispersive structures, segmentation, particles connecting points.*

Постановка проблемы. Развитие компьютерных технологий привело к появлению принципиально нового инструмента контроля качества диспергации – измерению микрообъектов по видеоизображениям. К основным проблемам указанного метода диагностики относится сложность сегментации изображений изучаемых объектов: низкая контрастность, неоднородность освещения и зашумленность фона, решению которых посвящены работы [1; 2].

Другой, не менее важной проблемой является наличие слипшихся

частиц – агрегированных образований, некорректная обработка которых приводит к существенному искажению результатов измерений.

Анализ последних достижений и публикаций. Известные методы решения этой задачи [3–7] основаны на анализе геометрии всей области, исследовании кривизны границы агрегата, распознавании перекрывающихся объектов на основе анализа сегментов, движущихся с разными скоростями, использовании морфологических подходов, пороговой сегментации с применением теории графов. Перечисленные подходы, как правило, приводят к излишней детализации объектов, чувствительны к шумам, задаваемым пороговым значениям, требуют значительных вычислительных затрат, вследствие чего не позволяют с достаточной точностью определить положение точек касания для агрегатов, состоящих из большого числа частиц, а также для образований, состоящих из объектов с высокой степенью перекрытия.

Постановка задачи. Целью данной работы является создание метода сегментации модельных агрегированных элементов дисперсных образований сферической формы путем определения точек их стыковки в агрегате по данным о его структуре и контуре, устойчивого к шумам изображения и позволяющего обрабатывать агрегаты сложной формы.

Основной материал. Общая схема метода для монохромного изображения выглядит следующим образом: после определения координат точек контура агрегата строится его скелет, на каждой ветке которого ищем точки присоединения капель – наиболее узкие места рассматриваемого агрегата – перешейки $Q_1^{(1)}Q_2^{(1)}$, $Q_1^{(2)}Q_2^{(2)}$, $Q_1^{(3)}Q_2^{(3)}$ на рис. 1. Для этого в каждой точке P рассматриваемой ветки скелета AB строим перпендикуляр, находим точки P_1, P_2 пересечения его с контуром агрегата. Назовем P_1P_2 – внутренним участком перпендикуляра. Так как перешейки являются наиболее узкими участками агрегата, им соответствуют перпендикуляры с наиболее короткими внутренними участками.

Но невыпуклость агрегата усложняет задачу, так как в общем случае прямая может пересекать контур невыпуклого объекта более чем в двух точках (рис. 2). В этом случае, поиск перешейков требует специальной схемы. Рассмотрим ее более подробно.

Первым шагом метода является фильтрация изображения, перевод его в монохромный режим [8]. Следующими этапами являются

определение координат точек контура агрегата [2] и построение по ним его скелета.

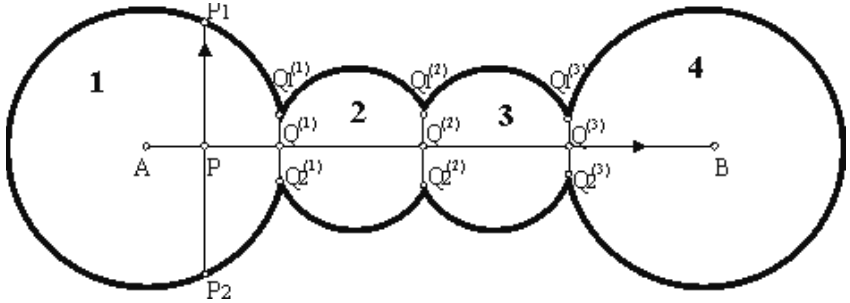


Рис. 1. Общая схема метода декомпозиции агрегатов дисперсных образований: АВ – ветка скелета; Р – произвольная точка ветки АВ; P_1, P_2 – точки пересечения контура агрегата с перпендикуляром к АВ, проведенным в точке Р; $Q^{(1)}, Q^{(2)}, Q^{(3)}$ – точки минимума функции ширины агрегата; $Q_1^{(1)}Q_2^{(1)}$, $Q_1^{(2)}Q_2^{(2)}$, $Q_1^{(3)}Q_2^{(3)}$ – точки присоединения капель

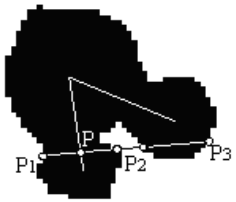


Рис. 2. Пересечение прямой контура невыпуклого объекта: Р – точка линии скелета; P_1, P_2, P_3 – точки пересечения контура агрегата с перпендикуляром к линии скелета, проведенным в точке Р

Предложенный в [9] метод позволяет построить скелет агрегата в виде набора отрезков прямых линий между узлами ветвления и включает следующие шаги: построение базового скелета на основе алгоритма Зонга-Суня, удаление избыточной связности, нахождение опорных точек и представление скелета в виде набора веток между ними, уточнение структуры скелета с целью представления ветвей отрезками прямых, упорядочивание ветвей соответственно продвижению от периферии агрегата к центру.

Последовательно перебирая ветви скелета, анализируем их положение относительно контура агрегата следующим образом. Ввиду того, что ветви скелета, согласно [9], хранятся в виде координат их начальной и конечной точек (А и В), используем какой-либо из алгоритмов построения отрезков [10], например, алгоритм Брезенхема, для разложения рассматриваемой ветви в растр и представления ее в виде последовательности точек. В каждой точке Р растеризованной ветки АВ скелета (рис. 1) строим перпендикуляр $y = kx + b$, точки $P_1, P_2, ..., P_N$ пересечения которого с контуром фигуры определяем, анализируя контур, из условия

$$\left| y - (kx + b) \right| \le \varepsilon ,$$

где x, y – координаты точек контура объекта.

Ввиду того, что в общем случае контур является невыпуклым $N \ge 2$. Разделим множество точек пересечения $\overline{P} = \left\{ P_{1,\ldots,} P_N \right\}$ на два подмножества $\overline{P}_+$ и $\overline{P}_-$ точек, расположенных, соответственно, справа и слева относительно прямой АВ (для наблюдателя, находящегося в точке P, лицом к точке В), используя свойства векторного произведения векторов

$$\overline{PB} \times \overline{PP_i}\ ,\ P_i \in \overline{P}\ . \qquad (1)$$

Для точек, расположенных по одну сторону АВ, выражение (1) сохраняет знак, а для точек расположенных по разные стороны – меняет на противоположный. В качестве границ внутреннего участка перпендикуляра выбираем ближайшие с обеих сторон из ($\overline{P}_+$ и $\overline{P}_-$) к P точки (P_1 и P_2 на рис. 1 и 2). В случае, когда все точки пересечения перпендикуляра с контуром располагаются по одну сторону от соответствующей ветки скелета, т. е одно из множеств точек пересечения оказывается пустым $\overline{P}_+ = \varnothing$ либо $\overline{P}_- = \varnothing$ (рис. 3) в качестве границ внутреннего участка перпендикуляра выбираем ближайшие точки из оставшегося множества. В дальнейшем, будем обозначать граничные точки внутреннего перпендикуляра в точке P как P_1 и P_2.

Результатом этого этапа является построение для каждой точки $P^{(i)}$ ветки АВ внутреннего перпендикуляра $P_1^{(i)}P_2^{(i)}$, длина которого $L^{(i)}$ определяет ширину отвечающего АВ участка агрегата в точке $P^{(i)}$

$$L_i = L\left(P^{(i)}\right) = \left|P_1^{(i)}P_2^{(i)}\right|.$$

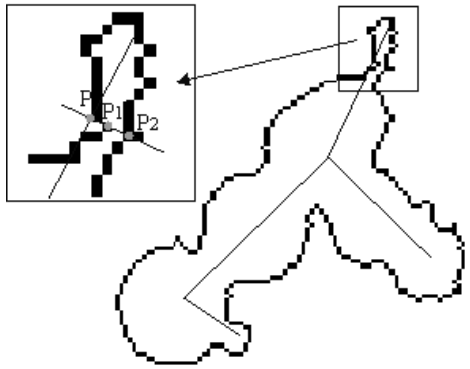


Рис. 3. Расположение граничных точек внутреннего перпендикуляра по одну сторону от ветки скелета: обозначения см. на рис.1

Места присоединения капель соответствуют локальным минимумам функции *L*. Назовем ее функцией ширины объекта. Постоянство функции *L* свидетельствует об отсутствии перешейков на рассматриваемом участке агрегата. Будем считать функцию *L*, определенную на отрезке АВ, постоянной, если выполняется условие

$$Max(L) - Min(L) \le 0{,}1\overline{L},$$

где $\overline{L} = \dfrac{\sum_{i=0}^{M-1} L_i}{M}$.

Для сглаживания функции L с целью уменьшения влияния шумов контура целесообразно использовать цифровые низкочастотные фильтры. Процедуру сглаживания не целесообразно проводить при рассмотрении коротких веток скелета (M<50), чтобы не допустить потери элементов рельефа контура на участках, соответствующих местам присоединения капель малого (относительно всего агрегата) размера.

Ввиду того, что значения координат точек контура в растровой графике представлены целыми числами, идущие последовательно значения функции L часто совпадают. Чтобы упростить алгоритм поиска локальных минимумов этой функции, проведем ее корректировку, линейно аппроксимируя функцию на участках, где ее значения совпадают. Так, если значения функции L совпадают на некотором участке *[k,t]*, $L(k-1) \ne L[t+1]$ ($k > 0$, $t < M-1$) аппроксимируем ее прямыми на отрезках $[i_0, i_c]$ и $[i_c, i_1]$, где $i_0 = k-1$, $i_1 = t+1$, $i_c = [(i_0 + l_1)/2]$ по схеме 1, приведенной на рис. 4, а. В случаях, когда $k = 0$ или $t = M-1$ аппроксимация проводится по схеме 2, представленной на рис 4, б.

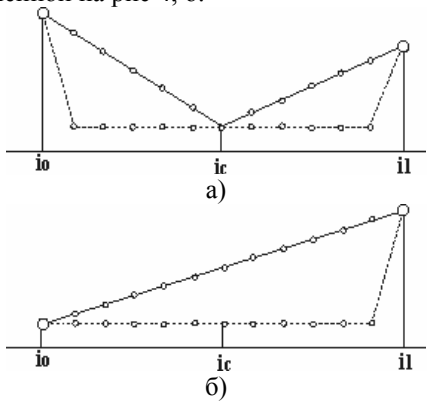


Рис. 4 Схема корректировки функции ширины объекта: а – схема 1; б – схема 2; ○----○----○ – исходное значение функции; □——□——□ – аппроксимированное значение функции

Точку j_0 функции L будем считать точкой перешейка, если выполняются следующие условия:

$$L_{j_0} < \overline{L}, \quad L_{j_0} < L_{j_0+i},$$

где $i = [-s/2; s/2]$; s – минимальный размер окрестности точки j_0 $\{i : L_i < L_{j_0}\}$. Значение s выбираем, исходя из величины ветки скелета. Так,

$M \le 10,$ $s = 3$;

$10 < M \le 50,$ $s = 10$;

$M > 50,$ $s = 0{,}2M$.

Таким образом, для рассматриваемой ветки скелета AB получаем множество точек $Q^{(1)}, Q^{(2)} \ldots Q^{(R)}$, соответствующих точкам присоединения капель (рис. 1). Внутренние перпендикуляры $Q_1^{(1)}Q_2^{(1)}$, $Q_1^{(2)}Q_2^{(2)}$, … , $Q_1^{(R)}Q_2^{(R)}$, построенные в этих точках, разбивают отвечающий AB участок агрегата на отдельные объекты (частицы). Граничные точки этих перпендикуляров являются точками присоединения частиц. По имеющимся точкам стыковки частиц распределяем точки границы агрегата между составляющими его объектами, включая внутренние перпендикуляры в контур обеих разделяемых ими частиц. Агрегат на рис. 1 разделим следующим образом:

объект 1: $\cup Q_1^{(1)}Q_2^{(1)} + Q_2^{(1)}Q_1^{(1)}$;

объект 2: $Q_1^{(1)}Q_2^{(1)} + \cup Q_2^{(1)}Q_2^{(2)} + Q_2^{(2)}Q_1^{(2)} + \cup Q_1^{(2)}Q_1^{(1)}$;

объект 3: $Q_1^{(2)}Q_2^{(2)} + \cup Q_2^{(2)}Q_2^{(3)} + Q_2^{(3)}Q_1^{(3)} + \cup Q_1^{(3)}Q_1^{(2)}$;

объект 4: $Q_1^{(3)}Q_2^{(3)} + \cup Q_2^{(3)}Q_1^{(3)}$.

По информации о координатах точек контурах частиц (x_n, y_n), $n = \overline{1, N}$, составляющих агрегат, вычисляем для каждой из восстановленных частиц площадь S и периметр P. Радиус каждой частицы определяем как

$$R = (R_1 + R_2)/2,$$

где $R_1 = \sqrt{S/\pi}$, $R_2 = P/2\pi$ – радиусы окружностей, соответственно, той же площади и периметра, что и анализируемая фигура. Центр –

$$x_c = \sum_{n=0}^{N-1} x_n \Big/ N \,;\; y_c = \sum_{n=0}^{N-1} y_n \Big/ N \,.$$

Окружности, соответствующие контурам капель, составляющих агрегат, могут быть восстановлены по известным параметрам. Таким образом, последовательно анализируем все ветки скелета агрегата.

Описанный метод разделения агрегатов проиллюстрирован на примере изображений аналитических фигур (рис. 5).

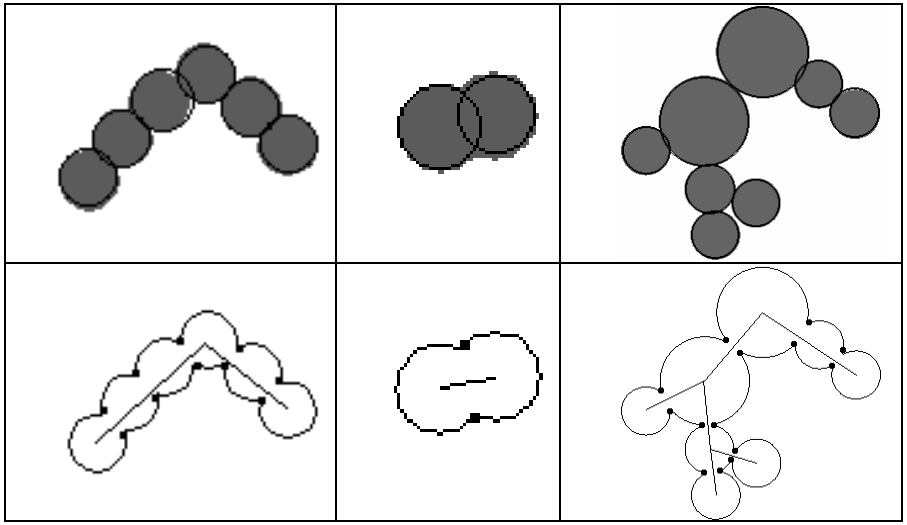


Рис. 5 Декомпозиция модельных объектов: линиями обозначены скелеты агрегатов, точками – места стыковки частиц

Выводы. В данной работе предложен метод сегментации агрегатов, состоящих из частиц сферической формы, каковыми являются, например, капли эмульсий и многих дисперсных образований. Метод основан на информации о структуре и контурах агрегатов.

Метод включает несколько этапов: фильтрация исходного изображения, перевод в монохромный режим, определение координат точек контура агрегата, построение его скелета, введение функции для каждой ветки скелета, характеризующей ширину отвечающего ей участка агрегата, определение точек присоединения частиц как точек локального минимума введенной функции.

Предложенный метод позволяет осуществлять декомпозицию агрегатов сложной конфигурации, состоящих из большого числа частиц сферической формы. Метод может быть использован при

разработке систем измерения и анализа эмульсий и других микрообъектов.

Библиографические ссылки

1. **Гонсалес Р.** Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – М., 2005. – 1072 с.
2. **Рузова Т. А.** Оперативный контроль параметров частиц дисперсных образований / Т. А. Рузова, О. Н. Карпов, Л. А. Флеер // Наук. вісник Нац. Гірнич. ун–ту. – 2004. – № 2. – С. 83–88.
3. **Honkanen M**. Analysis of the overlapping images of irregularly-shaped particles, bubbles and droplets / M. Honkanen // International Conference on Multiphase Flow. – Leipzig, 2007. – P. 370–382.
4. **Kutalik Z.** Occluding convex image segmentation for e.coli microscopy images / Z. Kutalik, M. Razaz, J. Baranyi // XII European Signal Processing Conference EUSIPCO. –Viena, 2004. – P. 937–940.
5. **Ritter N.** Segmentation and Border Identification of Cells in Images of Peripheral Blood Smear Slides / N. Ritter, J. Cooper // Proceedings of the Thirtieth Australasian Computer Science Conference (ACSC2007). – Ballarat, – 2007. – P. 161–169
6. **Young C.N.** A Method to Anchor Displacement Vectors to Reduce Uncertainty and Improve Particle Image Velocimetry Results / C.N.Young, D.A. Johnson, E.J. Weckman // Measurement Science and Technology. – 2004. – Vol. 15.– P. 9-20.
7. Unsupervised morphological segmentation of objects in contact / E. Martínez, X. Jové, F. Torre, E. Santamaría E. // Seizieme colloque Gretsi. – Grenoble, 1997. – P. 1379–1382.
8. **Рузова Т. А.** Модель пороговой классификации видеоизображений дисперсных образований // Зб. наук. праць НГУ. – 2007. – С. 162–167.
9. **Рузова Т.А.** Построение скелетов изображений агрегированных объектов дисперсий// Наук. вісник Нац. Гірнич.ун–ту. – 2012. – № 1. – С. 107–112.
10. Алгоритмические основы растровой машинной графики / Д.В. Иванов, А.С. Карпов, Е.П. Кузьмин, В.С. Лемпицкий и др. – Интернет-университет Информационных Технологий / Бином. Лаборатория Знаний. – 2007.– С. 283.

Надійшла до редколегії 20.06.2012

ВІДОМОСТІ ПРО АВТОРІВ

Архангельська Юлія Михайлівна – канд. техн. наук, доцент кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Байбуз Олег Григорович – д-р техн. наук, професор, завідувач кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Басова Ганна Олегівна – студентка Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Ватковський Сергій Олександрович – аспірант Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: теорія надійності.

Говоруха Володимир Борисович – д-р фіз.-мат. наук, доцент, завідувач кафедри вищої математики та інформатики Академії митної служби України.

Коло наукових інтересів: математичне моделювання соціально-економічних процесів.

Дегтярьов Андрій Андрійович – аспірант Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: обробка натуральних мов, автоматична обробка тексту, побудова систем інформаційного пошуку, оптимізація, комп'ютерна лінгвістика.

Ємел`яненко Тетяна Георгіївна – канд. техн. наук, доцент кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Зайцева Тетяна Анатоліївна – канд. техн. наук, доцент Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: обробка натуральних мов, автоматична обробка тексту, побудова систем інформаційного пошуку, оптимізація, комп'ютерна лінгвістика.

Земляна Світлана Валентинівна – канд. техн. наук, доцент кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: планування випробувань, контроль надійності.

Карпов Олег Миколайович – д-р техн. наук, професор, професор кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології розпізнавання мовних сигналів.

Лебідь Оксана Юріївна – канд. фіз.-мат. наук, доцент, доцент кафедри вищої математики та інформатики Академії митної служби України.

Коло наукових інтересів: математичне моделювання соціально-економічних процесів.

Луценко Олег Павлович – аспірант Дніпропетровського національного університету ім. Олеся Гончара

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Лучинкіна Ольга Іванівна – асистент кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології розпізнавання мовних сигналів.

Матвеєнко Анастасія Геннадіївна – студент Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: розпізнавання мови.

Мацуга Ольга Миколаївна – канд. техн. наук, доцент кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Пирогова Яна Валеріївна – аспірантка Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Приставка Пилип Олександрович – д-р техн. наук, професор, завідувач кафедри прикладної математики Національного авіаційного університету.

Коло наукових інтересів: автоматизована обробка даних, цифрова обробка зображень та відео, теорія та технології локальної сплайн-апроксимації.

Самарець Юрій Вікторович – канд. техн. наук, розробник Cognos, компанія Luxsoft.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Рузова Тетяна Олександрівна – канд. техн. наук, ст. наук. співроб. НДЛ моделювання процесів механіки рідини і газу та тепломасообміну кафедри аерогідромеханіки та енергомасопереносу ММФ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: визначення розмірів дисперсних утворень (та інших мікрооб'єктів) по їх відеозображенням (сегментація агрегованих частинок, визначення параметрів вдуву газових міхурів до рідини за їх відеозображенням та ін), просторова реконструкція і вимірювання тривимірних об'єктів.

Сидорова Марина Геннадіївна – асистент кафедри математичного забезпечення ЕОМ Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

Чебров Кирило Геннадійович – магістрант Дніпропетровського національного університету ім. Олеся Гончара.

Коло наукових інтересів: синтез мови.

Чолишкіна Ольга Геннадіївна – канд. техн. наук, доцент кафедри комп'ютеризованих засобів захисту інформації Національного авіаційного університету.

Коло наукових інтересів: локальна сплайн-апроксимація, технології автоматизованого захисту інформації.

Шубіна Ганна Сергіївна – магістрантка Дніпропетровського національного університету ім. Олеся Гончара

Коло наукових інтересів: інформаційні технології обробки статистичних даних.

ЗМІСТ

Архангельска Ю.М., Самарець Ю.В. Апроксимація неперервних випадкових процесів марковськими у системі гідрохімічного моніторингу 3

Байбуз О.Г., Сидорова М.Г. Групування пунктів гідрохімічного спостереження за схожістю хімічного складу води з урахуванням часових змін..... 12

Ватковский С.А., Емельяненко Т.Г. Нечеткие модели марковских процессов в теории надежности 18

Говоруха В.Б., Лебідь О.Ю. Проблема формального відображення змісту тексту. 29

Дегтярьов А.А., Зайцева Т.А. Симплістичний метод тематичної ідентифікації натурально-мовного тексту на основі розподілу ключових термів у тексті 35

Ємел'яненко Т.Г. Інформаційна технологія побудови прогнозів з використанням математичного апарату нечіткої логіки для фінансових часових рядів..... 45

Пирогова Я.В., Ємел'яненко Т.Г. Розробка інформаційної технології короткострокового прогнозування нечітких часових рядів 59

Земляна С.В. Обчислювальні схеми послідовного аналізу при контролі надійності з неперервною вибіркою..... 69

Карпов О.М., Чебров К.Г. Розробка алгоритмів і програм синтезу якісного мовлення за правилами 77

Луценко О.П., Байбуз О.Г. Огляд методів пошуку розладнань і перспективи їхнього застосування у технічному аналізі біржових котирувань 84

Лучинкіна О.І., Карпов О.М., Матвеєнко А.Г. Розробка алгоритмів і програм розуміння та синтезу шепітної мови 97

Мацуга О.М., Шубіна Г.С. Обчислювальні схеми ідентифікації сплайн-розподілів за ймовірнісним папером 111

Мацуга О.М., Басова Г.О. Програмне забезпечення відновлення розподілів за малими вибірками 123

Приставка П.О., Чолишкіна О.Г. Оцінка параметра експоненціального розподілу у випадку малої вибірки..... 133

Приставка П.О. Процедури небінарного subvision на основі лінійних B-сплайнів, близьких до інтерполяційних у середньому.... 142

Рузова Т.А. Модельные агрегаты дисперсий. Декомпозиция по структуре 153

Відомості про авторів 162

Наукове видання

АКТУАЛЬНІ ПРОБЛЕМИ АВТОМАТИЗАЦІЇ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

ТОМ 16

Збірник наукових праць

Українською, російською та англійською мовами

Редактор В.П. Пименов
Технічний редактор В.А. Усенко
Коректор В.П. Пименов

Підписано до друку 25.11.2012. Формат $60\times84\,{}^{1}\!/_{16}$. Папір друкарський.

Друк плоский. Ум. друк. арк. 12,09. Ум фарбовідб. 12,09. Обл.-вид. арк. 12,48. Тираж 100 пр. Вид. № 1567. Замовлення № .

Свідоцтво Державної реєстрації ДК № 289 від 21.12.2000р.
Видавництво Дніпропетровського університету, пр. Гагаріна, 72,
м. Дніпропетровськ, 49010
Друкарня ДНУ, вул. Наукова, 5, м. Дніпропетровськ, 49050