«Системні технології» 5 (148) 2023 «System technologies»
DOI 10.34185/1562-9945-5-148-2023-13
УДК 004.6: 007.5: 004.94

В.В. Скалозуб, В.М. Горячкін, І.А. Терлецький, І.П. Дудник

ФОРМУВАННЯ МОДЕЛЕЙ КЛАСИФІКАЦІЇ НЕВИЗНАЧЕНИХ ДАНИХ ПРОЦЕДУРАМИ РЕДУКЦІЇ І КАППА СТАТИСТИКИ

Анотація. Стаття присвячена розвитку математичних моделей класифікації невизначених даних, представлених нечіткими величинами та коефіцієнтами упевненості CF(A). Процедури формування шаблонів діагностування використовують модифіковані мережі Хеммінга (МХН), а також методи редукції та статистики каппа Коена. При цьому визначаються граничні розмірності та склад параметрів моделі класифікації, які забезпечують встановлені ймовірнісні вимоги достовірності результатів розрахунків. Представлена процедура редукції простору моделі діагностування невизначених даних. У статті наведено постановки, математичні моделі та реалізації завдань класифікації за недетермінованими даними. Прикладом моделі класифікації за нечіткими даними являється завдання із встановлення авторів україномовних текстів. Завдання класифікації при даних у форматі CF(A) відповідає відбору кандидата. Результати чисслового моделювання дозволили встановити результативність, достовірність та ефективність запропонованих процедур формування достовірних моделей класифікації при невизначених даних.

Ключові слова: класифікація, достовірні моделі, розмірність простору, нечіткі величини, коефіцієнти CF(A), модифікована мережа Хеммінга, процедура редукції, статистика каппа Коена, україномовні тексти, визначення автора, комп'ютерне моделювання.

Вступ та постановка проблеми. Завдання та процедури класифікації та діагностування за умов неповної визначеності вихідних даних (збурені, неповні ін.) являються досить поширеними на практиці [1, 2, 4, 6]. За їх результатами формуються моделі оптимального керування різноманітними технологічними процесами, вибору раціональних заходів/виконавців тощо [1, 6, 9]. Відзначається одна з головних проблем завдань класифікації, яка полягає у встановленні властивостей достовірності та повноти моделей (формати, структура та кількість зразків тощо, які забезпечують достовірний результати). Одним із актуальних нових аспектів моделей діагностування являється узгодженість ймовірнісних вимог до достовірності результатів, розмірності параметрів шаблонів, а також кількості даних (класів) за якими виконується

діагностування. Для підвищення достовірності результатів та ефективності класифікації на основі модифікованих мереж Хеммінга (МХН) актуальними являються дослідження можливостей і застосування вимог статистики каппа Коена та методу редукції розмірності [1, 3]. Важливе значення мають постановки завдань керування сервісними системами (С&С) з неповними та неточно визначеними даними і даними в природньомовній формі, які дозволяють представити рішення на основі моделей класифікації.

Аналіз останніх досліджень і публікацій. У сучасних складних системах деякі параметри станів або контрольовані характеристики процесів можуть мати значну ступінь невизначеності [1, 5, 6, 9]. Створення інтелектуальних процедур та інформаційної технології (ІІТ) для оптимізації потоків замовлень у С&С запропоновано у [1]. В ній вибір керувань виконується шляхом реалізації завдань діагностування з урахуванням умов невизначеності на основі МХН.

Завдання розвитку моделей та інтелектуальних багатопараметричних процедур діагностування за неповними і збуреними даними являються натепер актуальними [1, 2, 6]. У статті [6] було запропоновано підхід до формування шаблонів моделей класифікації на основі МХН [1]. При тому відзначено необхідність застосування методу редукції [3] та статистики каппа Коена [4]. Ці методи разом забезпечують отримання результатів із встановленими гарантіями щодо достовірності результатів. При тому в шаблонах МНХ використовуються дані виду коефіцієнтів упевненості, CF(A) [5].

Удосконалені моделі МНХ мають суттєву відмінність від класичних моделей асоціативної пам'яті Хеммінга в завданнях класифікації за неповними та збуреними даними. В МНХ використовуються в якості моделей даних нечіткі множини ($\mu X: X \to [0; 1]$), а також коефіцієнти впевненості CF(A) із значеннями в множині [– 1; + 1]. При тому класичні моделі Хеммінга [10] представляють дані за допомогою дискретної множини {– 1; + 1}. У [1, 6] удосконалено математичні моделі та процедури класифікації МХН з урахуванням показників, які відображають неточність або природномовну структуру первинних даних. Також представлено програмні засоби і результати експериментальних досліджень. За ними встановлена надмірність параметрів-ознак шаблонів, які використовувались для завдань визначення авторства україномовних текстів [7].

Для оцінки граничної розмірності параметрів моделі класифікації «n_0», простору який синтезується, використовуються результати теореми Вапника-Червоненкіса і процедури методу граничних спрощень (МСП) [3]. Величина

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

$$n_0 = (\varepsilon * L + \ln(h))/\ln(m) \quad (1)$$

визначає розмірність простору, перевищення якої призводить до втрати гарантії досягнення заданих параметрів достовірності «ε, h»; (1 – h) оцінка ймовірності умови безпомилкового розділення випадкової і незалежної вибірки довжини «L» при заданій граничній величині помилкової класифікації «ε».

В цій статті розроблена нова процедура, яка забезпечує формування простору (1) багатопараметричної класифікації з указаними вимогам достовірності.

Мета дослідження – підвищення достовірності моделей класифікації при невизначених даних, представлених нечіткими величинами та коефіцієнтами упевненості CF(A), шляхом застосування процедур редукції і статистики каппа Коена. Отримати постановки, моделі та реалізації завдань класифікації за нечіткими даними, які забезпечують встановлення авторів україномовних текстів, реалізувати завдання вибору кандидата за даними у форматі CF(A).

Результати та основний матеріал дослідження. Головним завдання, що вирішене у роботі, було формування достовірних математичних моделей класифікації та процедури аналізу при невизначених даних певних типів на основі методів редукції та каппа статистики (ПКР). ПКР складається із таких етапів:

1) виконати оцінку показника «каппа» ступеню подібності результатів класифікації на основі моделей шаблонів з різним числом або складом параметрів,

2) при забезпеченні «подібності» результатів класифікації для різних моделей-шаблонів виконати скорочення моделі, залишити однин із шаблонів,

3) визначити і видалити найменше значимі або найбільше «подібні» між собою параметри моделі класифікації, враховуючи значення «n0».

Узагальнений варіант циклу скорочення розмірності «шаблонів» моделі при вхідному «еталон»/»вимоги». Нехай $N(t) > n_0$ число параметрів шаблону на етапі (t) розмірності $N(t =0) =$ «m». Утворюється множина конкуруючих моделей шаблонів меншої розмірності. Для цього з набору змінних шаблонів N(t) і «еталону» вимог вилучається, наприклад, параметр X_j; нові вектори для цих спрощених моделей шаблонів позначимо (X/X_j)). Такі спрощені моделі формуються для кожної змінної із числа N(t), а за ними виконується класифікація на основі МНХ.

Перевірити «подібність» оцінок за каппа статистикою результатів класифікації для усіх пар змінних шаблонів {(X/X_j)}. Для різних наборів змінних (X/X_j) шаблонів за каппа-оцінками для кожного шаблону призначаються значення «+» або «-» в залежності від результатів класифікації. На основі порівняльного аналізу для всіх змінних утворюють таблицю розбіжностей, за якою розраховують статистичні оцінки «каппа Коена» [4]:

$$K = (P_0 - P_e) / (1 - P_e), \qquad (2)$$

В (4) P_0 – ймовірнісна оцінка що показує наскільки спостережувана узгодженість краща за випадкову, а P_e – результат підрахунку максимально можливої узгодженості за винятком випадкової узгодженості конкуруючих шаблонів [4].

Для вибору параметра спрощення моделі N(t) визначається пара наборів змінних виду (X/X_j) із найбільшим значенням «K» (2) та експертна оцінка «подібності» скорочених наборів (X/X_j). Якщо величина оцінки пари (2) відповідає вимогам, то певна змінна (їх множина) може бути видалена з «шаблонів» та «еталону». Після видалення таких параметрів з моделі класифікації увесь цикл розрахунків повторюється до виконання умови («m»<= n_0).

З метою удосконалення мережі Хеммінга (МХ) при встановлення оцінок величин відстані між нечіткими елементами (μX) зразків (w_{ik}) і вхідним вектором X = {x_i: i=0...n-1} уведене нечітке відношення наступного виду

$$R(W, X) = \{\mu R(w_i, x_i)/\{w_i, x_i\}=(1 – abs(w_i-x_i))/\{w_i, x_i\}, i=0, 1, ..., n-1. \qquad (3)$$

Якщо значення ступенів приналежності величин {w_i, x_i} однакові, тоді значення μR(w_i, x_i) =1. Коли одна з величин {w_i, x_i} дорівнює 0, а інша 1, тоді значення μR(w_i, x_i) = 0, інакше μR(w_i, x_i) → [0; 1]. Відношення (3) є одною із можливостей реалізації МХН для формування нечітких моделей класифікації, а також моделей представлених коефіцієнтами упевненості CF(A) [5].

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Таблиця 1

Результати циклу процедури редукції
при формуванні моделі класифікації

K_i	L_i	X	X/{X_1}	X/{X_2}	X/{X_3}	X/{X_4}	X/{X_5}	X/{X_6}
K_1	L_{11}		+					+
	L_{12}	+		+		+		
	L_{13}				+			+
K_2	L_{21}					+		+
	L_{22}		+	+				
	L_{23}				+			
	L_{24}	+				+		
K_3	L_{31}		+	+				
	L_{32}	+					+	
K_4	L_{41}							
	L_{42}		+		+	+		
	L_{43}							
	L_{44}	+				+		+
	L_{45}					+		
K_5	L_{51}			+				
	L_{52}						+	
	L_{53}							
K_6	L_{61}			+				
	L_{62}				+	+		
	L_{63}	+						
	L_{64}				+			
K_7	L_{71}						+	
	L_{72}				+			
	L_{73}	+						+
K_8	L_{81}					+		
	L_{82}	+	+	+				

Для пояснення процедури редукції приведемо умовний приклад реалізації наведених вище завдань на основі табл. 1 результатів класифікації для шаблонів з параметрів (X_1, X_2, …, X_6). В табл. 1 позначено : K_i – класи/зразки

моделі, які можуть містити кілька екземплярів L_i; X – набір параметрів моделі класифікації на поточному етапі процесу редукції; $X/\{X_j\}$ – скорочені набори моделей класифікації та вхідних векторів на поточному етапі без множини параметрів $\{X_j\}$, які видалені з моделей; знаки «+» – визначення модифікованих шаблонів переможців (класів) при застосуванні моделі МХН. У стовпці X позначаються шаблони, які були визначені на основі X наборів параметрів. Саме серед скорочених моделей шаблонів $X/\{X_j\}$ визначається набір параметрів $\{X_j\}$, які необхідно видалити на наступному етапі процедури редукції.

Таблиця 2

Розрахунки показників каппа статистики за таблицями розбіжностей

Розрахунок за класами (KM_1, C_2)					Розрахунок за класами (KM_1, C_1)				
$K(X_1/X_2)$	0,143				$K(X_1/X_2)$	0,143			
			Так	ні				Так	ні
		так	4	1			так	4	2
		ні	2	1			ні	1	1
$K(X_1/X_3)$	-0,067		Так	ні	$K(X_1/X_3)$	-0,333		Так	ні
		так	3	2			так	4	2
		ні	2	1			ні	2	0
$K(X_1/X_4)$	0,467		Так	ні	$K(X_1/X_4)$	0,333		Так	ні
		так	4	1			так	5	1
		ні	1	2			ні	1	1
$K(X_1/X_6)$	0,25		Так	ні	$K(X_1/X_6)$	0,143		Так	ні
		так	3	2			так	4	1
		ні	1	2			ні	2	1
$K(X_3/X_4)$	0,467		Так	ні	$K(X_3/X_4)$	0,333		Так	ні
		так	4	1			так	5	1
		ні	1	2			ні	1	1
$K(X_3/X_6)$	0,750		Так	ні	$K(X_3/X_6)$	0,714		Так	ні
		так	4	1			так	5	1
		ні	0	3			ні	0	2
$K(X_4/X_6)$	0,250		Так	ні	$K(X_4/X_6)$	0,143		Так	ні
		так	0	5			так	4	2
		ні	3	0			ні	1	1

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Множини $\{X_j\}$ можуть містити не одну, а любу кількість змінних, встановлену у завданні. Прийняте в табл. 1 число параметрів моделі класифікації n=6 – умовне, а граничне значення $n_0 = 4$. Дані табл. 1 фіксують результати класифікацій за мережею МХН, які отримані при однакових вихідних постановках завдань для всіх конкуруючих варіантів, сутність яких зараз не суттєва.

Множина варіантів моделей класифікації (КМ) містить всі комбінації пар скорочених моделей $X/\{X_j\}$. Виконуються наступні передумови аналізу результатів. $КМ_1$ – оцінюється класифікація хоча б одним із шаблонів L_{ij} класу K_i, $КМ_2$ – оцінюється класифікація кожним із L_{ij}. Також досліджуються різні способи формування таблиць розбіжностей для розрахунку показників каппа (2). Відповідно першого способу, C_1, враховуються відомі попередні результати класифікації за набором X, відповідно способу C_2 такі дані не використовуються, контролюються лише отримані результати (шаблони/класи) за МХН. Результати формування таблиць розбіжностей і розрахунків показників каппа (2) за даними табл. 1 для моделі класифікації $КМ_1$, коли в якості конкуруючих варіантів розглядалися всі комбінації пар скорочених моделей $X/\{X_j\}$, показані табл. 2.

У табл. 2 приведена частина результатів розрахунків оцінок «каппа», яка дозволяє визначити першу змінну для скорочення моделі класифікації – X_6, $K(X_3/X_6) = 0,75$ (значний рівень узгодженості). Для інших не приведених у табл. 2 пар оцінки величин «каппа» були несуттєвими, подібність відсутня, як для $K(X_1/X_3)=-0,067$. Можна сказати, що «вплив» фактору X_6 на модель класифікації ураховують та опосередковано представляють фактори X_1, X_3 і X_4. Відзначається, що оцінки подібності за каппа статистикою для таблиць розбіжностей без урахування класифікації за набором X, (спосіб C_2) перевищують чи дорівнюють результати за способом C_1. У табл. 3 приведено основні результати розрахунків оцінок показників «каппа» для скороченого складу параметрів класифікації, (X_1, X_2, …, X_5), представлені за схемою табл. 2.

Найвище значення «каппа» Коена $K(X_1/X_3) = 0,53$ (помірний рівень узгодженості), тому видаляється X_3 (подібність до X_1, X_4, X_5). Відзначається також відмінність результатів стосовно етапів табл. 2 та табл. 3 за способами формування таблиць розбіжностей C_1 і C_2. А саме – щодо врахування попередньо відомих результати класифікації за набором X. У табл. 3 більш високі оцінки показників були отримані за способом C_1, а не у табл. 2. Тобто необхідно розглядати обидва способи C_1 та C_2 при виконанні скорочення числа параметрів моделі, обираючи рішення за більшими значеннями «каппа». Також

необхідна перевірка тотожності та можливості корегування набору шаблонів моделі класифікації після видалення параметрів $\{X_j\}$. У разі утворення кількох однакових шаблонів скороченої моделі $X/\{X_j\}$ у одному або у кількох класах K_i, залишається лише один. Якщо такі шаблони виникли в одному класі, залишають тільки один, а в разі належності зазначених зразків до різних класів K_i видаляються шаблони із класу з більшим числом екземплярів. Наведемо результати процедури редукції для KM_2, коли оцінюється класифікація табл. 1 кожним із 26 зразків L_{ij}.

Таблиця 3

$K(X_1/X_2)$ -0,600		Так	ні
	так	2	3
	ні	3	0

$K(X_1/X_3)$ 0,529		Так	ні
	так	3	2
	ні	0	3

$K(X_2/X_5)$ -0,412		Так	ні
	так	1	4
	ні	2	1

$K(X_3/X_4)$ 0,467		Так	ні
	так	2	1
	ні	1	4

$K(X_3/X_5)$ 0,143		Так	ні
	так	1	2
	ні	1	4

$K(X_4/X_5)$ -0,429		Так	ні
	так	0	3
	ні	2	3

Параметр X_1 табл. 4 має найвищий показник «каппа» $K(X_1/X_2)$= 0,494 (помірний рівень узгодженості) і разом з тим високі оцінки подібності для $K(X_1/X_3)$ = 0,299 та $K(X_1/X_5)$ =0,325 , які перевищують значення інших. X_1 представлений через зазначені змінні, тому X_1 видаляється з моделі. У табл. 4 також більш високі оцінки показників «каппа» отримані за способом C_1.

В якості моделі класифікації за нечіткими даними розглядається завдання із визначення авторства творів україномовних авторів (ЗАТ) [1, 7]. У ЗАТ на підставі сукупності ознак, які характеризують набори текстів і утворюють класи або зразки моделі, необхідно визначити клас, до якого належить новий текст (закодований вектор ознак). Система ознак у авторів однакова, а число наборів (творів) може бути різним. Також певні дані кодів ознак профілю можуть бути «збуреними» або навіть відсутніми. Встановлено, що формування одного окремого шаблону для кожного автору представляє певну проблему. Тому кожному автору відповідає кілька зразків/шаблонів у формі векторів нечітких величин. Відповідно [7] натепер існують системи «вимірювання» властивостей текстів із понад 60 параметрів/ознак (розмірність класифікації m>60). Вимоги до моде-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

лей класифікації ЗАТ виконанні при представленні даних у формі таблиці табл. 1.

Таблиця 4

Розрахунки m=6 за зразками ($КМ_2$, $С_2$)

$К(X_1/X_2)$ 0,425		Так	ні
	так	3	2
	ні	3	18
$К(X_1/X_3)$ -0,035		Так	ні
	так	1	4
	ні	5	16
$К(X_1/X_5)$ -0,169		Так	ні
	так	0	5
	ні	3	18
$К(X_3/X_5)$ -0,182		Так	ні
	так	0	6
	ні	3	17
$К(X_5/X_6)$ -0,169		Так	ні
	так	0	3
	ні	5	18

Розрахунки m=6 за зразками ($КМ_2$, $С_1$)

$К(X_1/X_2)$ 0,494		Так	ні
	так	14	4
	ні	2	6
$К(X_1/X_3)$ 0,299		Так	ні
	так	10	5
	ні	4	7
$К(X_1/X_5)$ 0,325		Так	ні
	так	13	5
	ні	3	5
$К(X_3/X_5)$ 0,308		Так	ні
	так	11	2
	ні	7	6
$К(X_5/X_6)$ 0,278		Так	ні
	так	14	4
	ні	4	4

У [1, 7] кожному автору україномовного тексту відповідав один узагальнений шаблон, вибір творів до формування проблематичний. Вхідний вектор (твір невідомого автора) кодується встановленим набором нечітких величин, всі зразки творів моделі мають єдину форму відображення властивостей текстів. Для прикладу дослідження авторства україномовних текстів використано дані [1]. Ознаками текстів україномовних авторів ($К_1$ – І. Багряний, $К_2$ – О. Довженко, $К_3$ – М. Жадан, $К_4$ – М. Коцюбинський, $К_5$ –- Л. Українка, $К_6$ – П. Мирний, $К_7$ – І. Франко. ($К_8$ - інші)) були такі: X_1 – математичне очікування; X_2 – середнє квадратичне відхилення; X_3 – рекурентність; X_4 – детермінізм; X_5 – середня довжина діагональних ліній; X_6 – дивергенції; X_7 – ентропія; X_8 – завмирання; X_9 – затримки; X_{10} – середня кількість слів, X_{11} – середня кількість складів, X_{12} – середня кількість літер у реченні; X_{13} – середня кількість складів та X_{14} – літер у словах. Для авторів ($К_1$, …, $К_7$) на основі десяти творів формувалися шаблони на основі розрахунку областей можливих діапазонів та середніх значень. За творів інших авторів (Т. Шевченко, М. Стельмах ін.) формувався шаблон $К_8$ - інші. Величини X_1 – X_{14} нормувалися

до інтервалу [0; 1]. Для «визначення авторства» на мережу МХН подавалися нормовані значення характеристик твору невідомого автора; Мережа МХН, як правило, вірно встановлювала клас автора твору.

Разом з тим було встановлено, що для певних вхідних творів були визначені спрощені шаблони із 4-х параметрів, які за результатами класифікації відповідали сукупності параметрів-ознак X_1 – X_{14}. Тож показана необхідність формування шаблонів моделей ЗАТ з урахуванням вимог методу спрощень [4].

Модель класифікації при даних у форматі CF(A) змістовно відповідає завданню менеджменту – відбору кандидата із зазначеної множини (ЗК). Завдання призначення ЗК [8] відоме, має багато моделей і процедур реалізації, в залежності від типу та ознак параметрів, способів їх отримування і оцінювання ін. Особливість ЗК цієї роботи наступна. Вважається, що реалізується процедура відбору «кандидата» за даних його портфоліо (переліку робіт у певних проєктах) або резюме. Кожний з кандидатів K_i має набір завдань $З_{ік}$ портфоліо P_i, які описані без попередніх вимог до структури опису $З_{ік}$. Параметри та оцінки за $З_{ік}$ можуть бути представлені і текстом/мовна-форма, відрізняться для різних кандидатів P_i. У тому числі мають оцінки у формі неточних показників ін. Остаточну структуру і значення оцінок щодо представлення Зік визначає менеджмент, який встановлює оцінки всіх ознак у формі коефіцієнтів упевненості CF(A) У підсумку шаблони різних кандидатів K_i у моделі ЗК суттєво неоднорідні, можуть мати «пропуски» в ознаках, число шаблонів кандидатів K_i різне. Вхідний «еталон» містить всі характеристики, ураховані та означені «менеджментом» при формуванні моделі ЗК за портфоліо P_i. Метою являється визначення шаблона та K_i, який в найбільше відповідає «еталону», ким виконувалося подібне завдання.

Постановка ЗК структури табл. 1 (для K_1, …, K_4,…, K_{ij}) з даними, закодованими коефіцієнтами CF(A), відзначається таким. Кандидати мають різну кількість зразків, певні зразки не мають деяких ознак (наприклад, K_{11} і K_{12} не мають даних для X_5 і X_{11}, а K_{41} не містить X_5, X_6, X_9 та X_{11}). У «еталон» включені наступні ознаки: 1) Пріоритет задачі що виконувалась. 2) Складність задачі. 3) Оцінки навичок категорії H_1. 4) Оцінки навичок категорії H_2. 5) Оцінка рівню певного фаху 6) Оцінка досвіду виконавця таких завдань. 7) Оцінка рівня визначених знань 31. 8) Завантаженість у проєкті. 9) Навичка H_9, …. 11) Оцінка навичок контролю. Засобами МХН [1, 8], а також процедурами редукції та каппа Коена, необхідно встановити шаблон індивідуальних ознак $З_{ік}$ «виконавця,

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

клас K_i» у форматі CF(A) для (X_1, ...,X_{11}), що найкраще відповідає «еталону», наприклад, (X_1=0.9, X_2=0.75, X_3= - 0.5, ..., X_6= 1, ... , X_{11}= - 0.3). Структура простору моделі класифікації формується окремо для кожного «еталону вимог», щоб забезпечити встановлену достовірність результату класифікації.

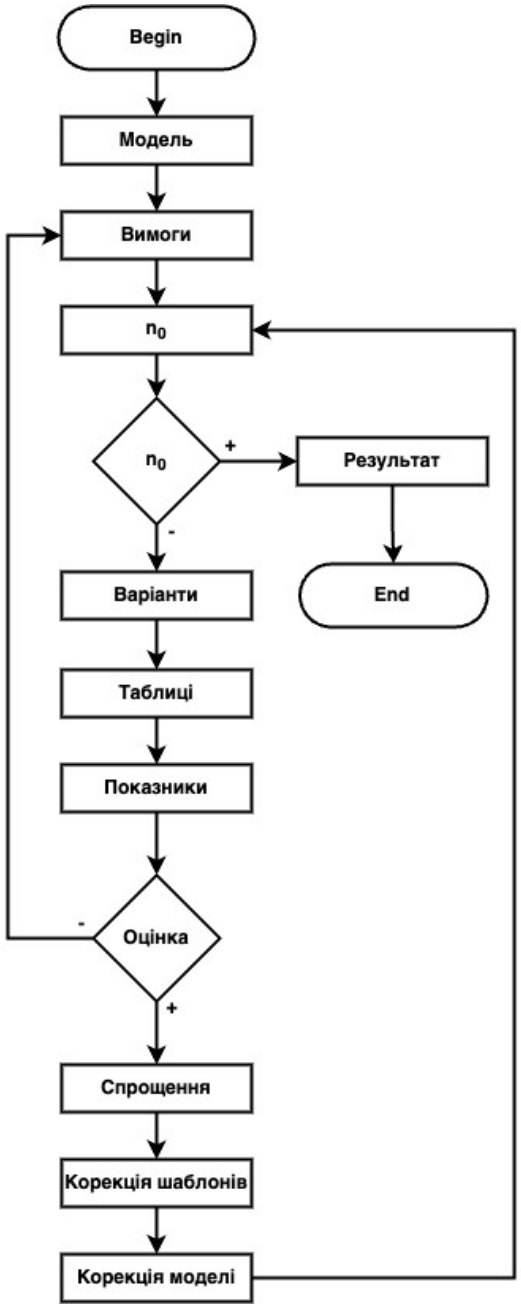

Рисунок 1 - Блок-схема алгоритму процедури редукції розмірності моделі

За схемою рис. 1 узагальнено алгоритм редукції на основі каппа статистики виконується таким чином: отримується Модель завдання класифікації за умов невизначеності даних; визначаються Вимоги щодо забезпечення точності та достовірності очікуваних результатів класифікації; за отриманими даними

розраховується гранична розмірність моделі n_0; перевіряється відповідність поточної моделі класифікації вимогам до моделі за значенням n_0; в разі виконання вимог – маємо достовірний Результат; у разі невиконання вимог починається аналіз простору і формування конкуруючих Варіантів спрощених моделей класифікації; для всіх сформованих варіантів реалізації конкуруючих моделей формуються таблиці розбіжностей; за таблицями розраховуються показники статистики каппа Коена та визначається загальна оцінка всіх варіантів; якщо така оцінка схожості конкуруючих моделей не відповідає вимогам достовірності, необхідно змінити вимоги; інакше виконується процедура спрощення (видалення певної структури параметрів моделі класифікації); при тому перевіряється необхідність і виконується корегування систем шаблонів і моделі класифікації.

Розроблено програмний комплекс, призначений для побудови математичної моделі процесів нечіткої класифікації – шаблонів баз знань для реалізації класифікації, розрахунків процесів класифікації на основі мережі МХН. Комплекс забезпечує автоматизацію обробки завдань нечіткої та CF(A) класифікації, щодо визначення класів вхідних елементів на основі їх неточно визначених ознак, прогнозування категорії «виконавця» та автору україномовного твору.

Висновки. У статті отримані удосконалені математичні моделі, алгоритми та програмні засоби, призначені для підвищення достовірності результатів класифікації при невизначених даних, представлених нечіткими величинами та коефіцієнтами упевненості CF(A). Удосконалення моделей класифікації з наведеними властивостями даних забезпечується шляхом модифікації нейронних мереж Хеммінга, а також застосування процедур редукції і статистики каппа Коена, які дозволяють сформувати математичні моделі зі встановленими ймовірнісними вимогами до результатів. В роботі запропоновані та реалізовані нові постановки, математичні моделі, а також виконано реалізацію моделей класифікації за нечіткими даними, що вирішують завдання із встановлення авторів україномовних текстів, а також завдання вибору кандидата за моделлю даних у форматі коефіцієнтів CF(A). Розроблено програмні засоби для формування моделей класифікації при невизначених даних процедурами редукції і каппа статистики.

Числові експерименти підтвердили достовірність результатів класифікації, а також ефективність запропонованої процедури редукції розмірності моделей на основі статистики каппа Коена.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ЛІТЕРАТУРА

1. Скалозуб В. В., Горячкін В. М., Клименко І. В., Терлецький І. А., Терленко А. П. Дослідження процедур мережі Хеммінга для управління сервісними системами при неточно визначених і природомовних даних. Наука та прогрес транспорту. 2022. № 3-4 (99-100). С. 33–47. DOI: https://doi.org/10.15802/stp2022/276411 (in Ukrainian)

2. Великоіваненко Г. І. Оцінювання рівня економічної безпеки на підґрунті відстані Хеммінга. 2018. URL: https://core.ac.uk/download/pdf/197269753.pdf

3. Васильев В. И. Индукция и редукция в проблемах экстраполяции. Кибернетика и вычислительная техника. 1998. Вып. 116. С. 65–81.

4. Колесник А. С., Хайрова Н. Ф. Обґрунтування використання статистики каппа Коена в експериментальних дослідженнях NLP Text Mining. Кібернетика та системний аналіз. Т. 58, № 2. 2022. С. 143–153.

5. Li Min Fu, Shortliffe E. H. The application of certainty factors to neural computing for rule discovery. IEEE Transactions on Neural Networks. 2000. Vol. 11. Iss. 3. P. 647–657. DOI: https://doi.org/10.1109/72.846736

6. Скалозуб В. В., Горячкін В. М., Терлецький І. А. Багатопараметричні інтелектуальні процедури діагностування за неповними і збуреними даними // Логістика і транспортна безпека: Проблеми та перспективи розвитку в контексті аналізу сучасних викликів і загроз: матеріали доповідей II Міжнародної науково-практичної конференції. — Дніпро: Середняк Т.К., 2023. С. 42 – 47.

7. Шинкаренко В. І., Демидович І. М. Визначення ознак авторства природномовних текстів. Штучний інтелект. 2018. № 3. С. 27–35.

8. Richard A. Brualdi. Combinatorial matrix classes. — Cambridge: Cambridge University Press, 2006. — (Encyclopedia of Mathematics and Its Applications). — ISBN 0-521-86565-4

9. Leszek Rutkowski Metody i techniki sztucznej inteligentcji. Naukove PWN, Warszava, 2005. – 520 p.

10. Haykin S. Neural networks: A Comprehensive Foundation. Prentice hall: New Jersey, 1999.1103 p.

REFERENCES

1. Skalozub V.V., Goryachkin V.M., Klymenko I.V., Terletsky I.A., Terlenko A.P. Doslidzhennia protsedur merezhi khemminha dlia upravlinnia servisnymy systemamy pry netochno vyznachenykh i pryrodomovnykh danykh. 2022. No. 3-4 (99-100). P. 33–47. DOI: https://doi.org/10.15802/stp2022/276411

2. Velykoivanenko, H.I. (2018). Otsiniuvannia rivnia ekonomichnoi bezpeky na pidgrunti vidstani Khemminha. Retrieved from https://core.ac.uk/download/pdf/197269753.pdf (in Ukrainian)
3. Vasilev, V.I. (1998). Induktsiya i reduktsiya v problemakh ekstrapolyatsii. Cybernetics and Computer Engineering, 116, 65-81. (in Russian).
4. Kolesnyk A. S., Khairova N. F. . Obgruntuvannia vykorystannia statystyky kappa Koena v eksperymentalnykh doslidzhenniakh NLP Text Mining Cybernetics and system analysis. Vol. 58, No. 2. 2022. P. 143–153.
5. Li Min Fu, Shortliffe E. H. The application of certainty factors to neural computing for rule discovery. IEEE Transactions on Neural Networks. 2000. Vol. 11. Iss. 3. P. 647–657. DOI: https://doi.org/10.1109/72.846736
6. Skalozub V.V., Goryachkin V.M., Terletskyi I.A. Bahatoparametrychni intelektualni protsedury diahnostuvannia za nepovnymy i zburenymy danymy// Logistics and transport safety: Problems and prospects of development in the context of analysis of modern challenges and threats: materials of reports II International scientific and practical conference. — Dnipro: Serednyak T.K., 2023. P. 42-47.
7. Shinkarenko V. I., Demidovych I. M. Determination of signs of authorship of natural language texts. Artificial Intelligence. 2018. No. 3. P. 27–35.
8. Richard A. Brualdi. Combinatorial matrix classes. — Cambridge: Cambridge University Press, 2006. — (Encyclopedia of Mathematics and Its Applications). — ISBN 0-521-86565-4
9. Leszek Rutkowski Metody i techniki sztucznej inteligentcji. Naukove PWN, Warsaw, 2005. – 520 p.
10. Haykin S. Neural networks: A Comprehensive Foundation. Prentice hall: New Jersey, 1999.1103 p.

Received 23.11.2023.
Accepted 24.11.2023.

Formation classification models of undetermined data by means of procedures reduction and kappa statistic

The article is devoted to the development of mathematical models for the classification of uncertain data represented by fuzzy values and certainty factors CF(A). Diagnostic pattern formation procedures use modified Hamming networks (MHN), as well as reduction methods and Cohen's kappa statistics. At the same time, the limiting dimensions and composition of the parameters of the classification model are determined, which ensure the established probabilistic requirements for the reliability of the calculation re-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

sults. The model space reduction procedure and the structure of the software complex for diagnosing uncertain data are presented. An example of a classification model based on fuzzy data is the task of identifying the authors of Ukrainian-language texts. The classification task for data in CF(A) format corresponds to candidate selection. The results of the numerical modeling made it possible to establish the effectiveness, reliability and efficiency of the proposed procedures for the formation of reliable classification models with uncertain data.

Keywords: classification, reliable models, dimensionality of space, fuzzy values, CF(A) certainty factors, modified Hamming network, reduction procedure, Cohen's kappa statistic, Ukrainian-language texts, author's definition, computer simulation.

Скалозуб Владислав Васильович - професор, каф. «Комп'ютерні інформаційні технології», Український державний університет науки і технологій, УДУНТ.
Горячкін Вадим Миколайович – зав. каф. «Комп'ютерні інформаційні технології», Український державний університет науки і технологій, УДУНТ.
Терлецький Ігор Андрійович – аспірант каф. «Комп'ютерні інформаційні технології», Український державний університет науки і технологій, УДУНТ.
Дудник Ілля Петрович – магістрант каф. «Комп'ютерні інформаційні технології», Український державний університет науки і технологій, УДУНТ.

Skalozub Vladyslav – professor, dep. "Computer and Information Technology", Ukrainian State University of Science and Technology, USUST.
Horiachkin Vadim – Head of dep. "Computer and Information Technology", Ukrainian State University of Science and Technology, USUST.
Terlitskyi Ihor – post-graduate student, Dep. "Computer and Information Technology", Ukrainian State University of Science and Technology, USUST.
Dudnyk Ilya – graduate student, Dep. "Computer and Information Technology", Ukrainian State University of Science and Technology, USUST