УДК 519.23

DOI: http://dx.doi.org/10.20535/2219-3804172017123925

Н. А. Яремчук[1], *професор кафедри, к.т.н.*, **Р. С. Семенюк**[2], *аспірант*

ВИКОРИСТАННЯ НЕЧІТКОЇ КЛАСИФІКАЦІЇ ВИДУ РОЗПОДІЛУ ДЛЯ ВИБІРОК МАЛОГО ОБ'ЄМУ

En Existing methods for improving the measurement result accuracy with small volume samples are analyzed. The fuzzy identification method (proposed by Y. M. Klikushyn) which permits identification distributions under condition of small samples was selected for distribution identification.

The further research of features and development of fuzzy identification in small volume samples was conducted, the scope of fuzzy identification usage was expended and a research on probability of dispensation differentiation was carried out in the work.

The reliability of classifying samples using the generation of reference samples was defined in the work. The research of reliability on one sample showed that the probability of changing linguistic code is quite large. It is recommended to average several samples to increase the likelihood of accurate identification. Number of samples can be determined for a given probability.

The research of fuzzy classification included the development of finding fuzzy estimates algorithm for samples of different sizes, finding the necessary and sufficient number of estimates and choice of choice principles of used estimates from the possible range. The correlations for fuzzy estimates based on the specified volume were obtained. Usage of these ratios allowed expanding the scope of fuzzy classification and creating a library of linguistic codes which greatly simplifies its procedure. Recommendations on the samples number required for achieving necessary classification probability were made.

Ru В статье рассмотрено применение метода нечеткой классификации вида распределения одной или нескольких выборок малого объема, основанного на непараметрическом оценивании, для построения контрольных карт точности и стабильности технологического процесса. Нечеткая классификация позволяет определить эффективную оценку центра распределения и вид контрольной карты.

Вступ

На практиці часто виникають задачі опрацювання результатів вимірювань за малих об'ємах вибірок. Отже невизначеність результату вимірювання доволі велика. У зв'язку із цим постає задача підвищення точнос-

[1] *НТУУ «Київський політехнічний інститут ім. Ігоря Сікорського», кафедра інформаційно вимірювальної техніки*
[2] *НТУУ «Київський політехнічний інститут ім. Ігоря Сікорського», кафедра інформаційно вимірювальної техніки*

ті результату вимірювання. Тому пошук нових методів, що дозволяють підвищити точність при опрацюванні результатів вимірювань є актуальним.

Як показали дослідження, проведені у роботах [1, 2], знання виду розподілу має важливе значення під час опрацюванні результатів вимірювань.

Знання виду розподілу дозволяє:

– у значній мірі підвищити точність оцінок, що є результатами вимірювань за рахунок вибору найбільш ефективних [4];
– значно точніше визначати коефіцієнт охоплення для розширеної невизначеності [1, 2] замість вибору оцінки зверху.

Таким чином знання розподілу дозволяє зменшити і стандартну, і розширену невизначеність.

Так у роботі [1] наведено ефективні оцінки (медіана, середина розмаху), знайдені статистичним опрацюванням вибірок, розподілених за іншими, ніж нормальний, розподілами. Якщо розподіл генеральної сукупності, із якої отримують практичну вибірку скінченного обсягу, є інший ніж нормальний, то середнє значення не обов'язково буде найкращою оцінкою результату вимірювання.

Зокрема, у разі розподілів із обмеженими граничними значеннями, наприклад, рівномірного, особливо арксинусного розподілів генеральної сукупності за обсягів вибірки значно ефективнішою оцінкою результату (із меншою дисперсією за заданого обсягу вибірки) є середина розмаху вибірки.

У табл. 1 наведені значення відносної ефективності E_n середини розмаху і медіани відносно середнього значення вибірки [1]:

$$E_n = \frac{u(x)}{u(x_c)}, \qquad (1)$$

де $u(x)$ – стандартна невизначеність середнього арифметичного,

$u(x_c)$ – стандартна невизначеність інших оцінок центру розподілу

***Таблиця* 1.**

Відносна ефективність середини розмаху та медіани із вибірок деяких розподілів

Розподіл/оцінка	E_n	$n=9$
Арксинусний/середина розмаху	$\frac{n}{\pi^2}\sqrt{\frac{n}{5}}$	1,22
Лапласа/медіана	$\sqrt{2}$	1,41

Під час оцінювання якості вимірювання важливою є не тільки стандартна невизначеність, але й розширена. Оскільки форма розподілу Лапласа і особливо рівномірного та арксинусного розподілів істотно відрізняєть-

ся від нормального, то і зв'язок між розширеною і стандартною невизначеністю середини розмаху вибірок є іншим ніж у разі нормального розподілу вибірки. Цей зв'язок, як відомо, кількісно відображається коефіцієнтом охоплення, а також ефективною кількістю ступенів свободи. Якщо розподіл не відомий, то потрібно брати оцінку зверху, тобто максимальне значення. Існуючі методи ідентифікації виду розподілу або не можуть бути використанні для малих вибірок, або вони можуть використовуватись лише для конкретного розподілу [1, 2]. Тому для ідентифікації розподілу у роботі обрано метод нечіткої ідентифікації, запропонований Клікушиним Ю. М. [2], який дозволяє ідентифікувати розподіли за умов малих вибірок. Оскільки нечіткі оцінки залежать від об'єму вибірки або кількості порядкових статистик, метод дозволяє ідентифікувати розподіли у разі фіксованого об'єму вибірки у 9 або 17 елементів.

Даний метод побудовано наступним чином. У якості теоретичної бази для формування процедур ідентифікації використана теорія нечітких множин, зокрема поняття лінгвістичної змінної (ЛЗ), а також введене поняття центр розподілу [2, 3]. Нечітка оцінка (НО) представляє собою середнє арифметичне певних частин впорядкованої вибірки, тим самим вказуючи, у якій частині діапазону концентрується більшість значень.

Для ідентифікації використовуються шість нечітких оцінок, чутливих до концентрації впорядкованих значень. Нечіткі оцінки, що представляють комбінації певних порядкових статистик, ранжуються, кодуються і за певним набором коду ідентифікують розподіл. Як вказано у роботі [1], у разі нечіткої ідентифікації можна розрізнити рівномірний, нормальний, арксинусоїдний розподіли і розподіл Лапласа.

Апробація нечіткої ідентифікації, проведена автором, показала доцільність використання даного метода. Але процедура нечіткої ідентифікації, що наведена в роботі [2], розроблена тільки для фіксованого об'єму вибірки ($n=9, n=17$). Крім того відсутній аналіз вірогідності розрізнення розподілів.

Постановка задачі

Провести подальше дослідження особливостей і удосконалення нечіткої ідентифікації у вибірках малого об'єму, розширити область застосування нечіткої ідентифікації й провести дослідження вірогідності розрізнення розподілів.

Генерація опорних вибірок для верифікації і дослідження нечіткої ідентифікації

Для дослідження нечіткої ідентифікації обрано метод генерації опорних вибірок. Найпростішим і найбільш фундаментальним методом, на ос-

нові якого генеруються подібні величини, є метод зворотної функції [4], за яким для генерації випадкової величини із розподілу X генерується випадкове число r і вирішується рівняння $r = F(x)$ щодо значення $x = F^{-1}(R)$.

Метод зворотної функції: $R = F^{-1}(x)$.

Так, наприклад, для арксинусного розподілу у діапазоні $-a < x < a$. Зі щільністю

$$\frac{1}{\pi\sqrt{a^2 - x^2}}.$$

Функція розподілу ймовірності має наступний вигляд:

$$F(x) = \begin{cases} 0\_\text{при}\_ x < -a \\ \frac{1}{2} + \frac{1}{\pi}\arcsin\frac{x}{a}\_\text{при}\_ -a < x < a \\ 1\_\text{при}\_ x > a \end{cases}.$$

У свою чергу зворотне перетворення дорівнює: $a \cdot \sin\pi\left[\frac{1}{n-1} - 0{,}5\right] = x_i \;\; i \in (0, n-1)$, де n – кількість членів опорної вибірки, що створює дискретну решітку за генерацією опорних вибірок.

Вибірка, отримана зворотнім перетворенням арксинусного розподілу при $n = 9$: –3; –2,62; –1,95; –1; 0; 1; 1,95; 2,62; 3. Враховуючи наступні розрахунки вірогідності, вибірка розраховується для інтервалу [0,1].

У результаті отримуємо:0; 0,063; 0,175; 0,33; 0,5; 0,66; 0,825; 0,937; 1.

Згідно із імен нечітких оцінок (формули для їх розрахунку наведено у [2] і нижче), розраховуємо НО для отриманої вибірки:

$$\text{"168"} = \left(x_{(1)} + \left(x_{(6)} + x_{(8)}\right)/2\right)/2 = 0{,}399;$$

$$\text{"267"} = \left(x_{(2)} + \left(x_{(6)} + x_{(7)}\right)/2\right)/2 = 0{,}403;$$

$$\text{"249"} = \left(\left(x_{(2)} + x_{(4)}\right)/2 + x_{(9)}\right)/2 = 0{,}598;$$

$$\text{"348"} = \left(\left(x_{(3)} + x_{(4)}\right)/2 + x_{(8)}\right)/2 = 0{,}595;$$

$$\text{"1478"} = \left(\left(x_{(1)} + x_{(4)}\right)/2 + \left(x_{(7)} + x_{(8)}\right)/2\right)/2 = 0{,}523;$$

$$\text{"2369"} = \left(\left(x_{(2)} + x_{(3)}\right)/2 + \left(x_{(6)} + x_{(9)}\right)/2\right)/2 = 0{,}4745,$$

де $x_{(1)} ... x_{(9)}$ – порядкові статистики.

Ранжована вибірка значень НО: 0,399; 0,403; 0,4745; 0,523; 0,595; 0,598.
Медіана ранжованої вибірки: *med* =(0,4745+0,523)/2=0,49875

Кодування виконується за правилом: якщо значення НО менше за медіану, то даному номеру присвоюється код 1, інакше –0.

Виходячи із наведеного правила, отриманий за наведеним вище порядком НО лінгвістичний код (ЛК) буде мати наступний вигляд: 110001.

У результаті проведених розрахунків для інших розподілів отримуємо наступні результати:

***Таблиця* 2.**

Результати визначення опорних вибірок, нечітких оцінок і лінгвістичних кодів вибірки

Закон розподілу	Вибірка, наведена для інтервалу [0, 1]	Нечіткі оцінки	Лінгвістичний код
Нормальний	0; 0,119; 0,247; 0,372; 0,5; 0,622; 0,748; 0,876; 1	0,249; 0,356; 0,535; 0,580; 0,481; 0, 564	110010
Коші	0; 0,124; 0,252; 0,378; 0,5; 0,627; 0,753; 0,878; 1	0,342; 0,378; 0,593; 0,567; 0,465; 0,582	110010
Рівномірний	0; 0,248; 0,3875; 0,4468; 0,5; 0,553; 0,6125; 0,752; 1	0,326; 0,415; 0,674; 0,585; 0,453; 0,547	110100
Арксинусний	0; 0,063; 0,175; 0,33; 0,5; 0,66; 0,825; 0,937;1	0,399; 0,403; 0,598; 0,595; 0,523; 0,4745	110001

Головною перевагою нечіткої ідентифікації є її швидкодія. Із іншої сторони, у неї є і один важливий недолік. Як бачимо із результатів, наведених у таблиці, даний вид ідентифікації не дозволяє отримати однозначний результат у розподілах одного класу (наглядно це видно по ЛК нормального розподілу і розподілу Коші).

Нечітка ідентифікація реалізується порівнянням лінгвістичного коду вибірки із кодами опорних вибірок. Якщо код вибірки не співпадає із кодами опорних вибірок, знаходиться найближчий розподіл за коефіцієнтом еквівалентності. У цьому випадку висновок набуває вигляду «розподіл вибірки близький до обраного за коефіцієнтом еквівалентності».

Проведений аналіз на опорних вибірках показав, що у дійсності нечітка ідентифікація дозволяє розрізняти наступні класи розподілів: одномодальний симетричний, одномодальний несиметричний, двухмодальний, рівномірний.

Тому надалі метод, що досліджуємо назвемо нечіткою класифікацією із визначенням класу розподілу.

Але у разі нечіткої класифікації (розподілу за класами) можна реалізувати для вибору більш ефективної оцінки і коефіцієнту охоплення [3]. Наприклад, у [3] для класу одномодальних симетричних розподілів рекомендується формула для коефіцієнту охоплення:

$$t = 1{,}62 \cdot \left[3{,}8 \cdot (\varepsilon - 1{,}6)^{2/3} \right]^{\lg \lg [1/(1-P_\partial)]}.$$

Для розподілів класу Шапо (рівномірний розподіл):

$$t = 1{,}56 \cdot \left[1{,}12 + (\varepsilon - 1{,}8)^{0{,}58} / \sqrt{10} \right]^{\lg [0{,}1/(1-P_\partial)]}.$$

Для класу гостровершинних двухмодальних розподілів (арксинусний розподіл):

$$t = 1{,}23 \cdot \left[1 + \sqrt{\frac{\varepsilon - 1}{2{,}5}} \lg \frac{0{,}175}{1 - P_\partial} \right].$$

Розширення можливостей нечіткої класифікації для різних об'ємів вибірки

Нечітка ідентифікація, що продемонстрована в роботі [2], розроблена тільки для фіксованого об'єму вибірки ($n = 9, n = 17$). Тому у даній роботі проведене подальше дослідження особливостей і удосконалення нечіткої класифікації у разі вибірок малого об'єму.

За різних об'ємах вибірки буде змінюватись і кількість нечітких оцінок. Тому постає питання формування алгоритму знаходження НО.

Під час підбору нечітких оцінок необхідно дотримуватися кількох умов:

- оцінки повинні бути лінгвістично еквівалентними, тобто середнє значення їх індексів має дорівнювати медіані;
- оцінки повинні бути незалежними, тобто такими що б їх не можна отримати лінійною комбінацією інших НО;
- оцінки повинні бути асиметричними, тобто такі, у яких індекси елементів розташовуються несиметрично щодо медіани.
- оцінки повинні бути попарно симетричними, тобто у наборі оцінок кожна оцінка повинна мати собі пару, дзеркально відображає її щодо медіани.

Таким чином опрацювання результатів вимірювань проводяться за наступними етапами:

- нечітка класифікація форми розподілу;
- вибір відповідно до класу оцінки центру розподілу, що приймають як результат вимірювання;
- обчислення стандартної невизначеності обраної оцінки;
- обчислення коефіцієнту охоплення відповідно класу розподілу.

Розглянувши вибірки розміром у одинадцять, тринадцять і п'ятнадцять елементів, була помічена закономірність у тому, що елементи

цих оцінок підбираються за однаковим принципом і жорстко прив'язані до індексу медіани і розміру вибірки.

***Таблиця* 3.**

Рекомендації для імен НО при різних об'ємах вибірок

Імена оцінок за умови різних об'ємів вибірок N		
N=11	N=13	N=15
1–8–9	2–6–13	2–7–15
3–4–11	3–6–12	3–7–14
1–7–10	3–5–13	3–6–15
2–5–11	1–9–11	1–10–13
2–7–9	2–8–11	2–9–13
3–5–10	1–8–12	1–9–14

У самих «широких» оцінок ліві індекси дорівнюють один одному, центральні знаходяться на однаковій відстані від медіани, а праві приймають значення максимального індексу або знаходяться на рівній відстані від нього. Це дозволяє вивести формули шести оцінок, необхідних для знаходження розподілу. За цими формулами можна обчислити НО для будь-якого числа вимірів.

На основі отриманих оцінок отримано бібліотеку всіх лінгвістичних кодів для розподілів, яка значно спростить процедуру класифікації форми розподілу.

***Таблиця* 4.**

Лінгвістичні коди при різних об'ємах вибірок

Найменування розподілу	Лінгвістичні коди за умови різних об'ємів вибірок		
	N=11	N=13	N=15
Нормальний	111000	111000	111000
Коші	111000	111000	111000
Рівномірний	101100	011100	011100
Лапласа	110010	110001	110001
Арксинусний	000111	000111	000111

Достовірність класифікації

Значення порядкових статистик підлягають статистичним змінам. Математичною моделлю цих змін може бути бета-розподіл, що відповідає розподілу порядкових статистик [4].

Якщо $\xi_1, \xi_2, ..., \xi_n$ незалежні і рівномірно розподілені на відрізку [0, 1] і $\xi_{(1)}, \xi_{(2)}, ..., \xi_{(n)}$ – впорядковані за зростанням величини, де $\xi_{(k)}$ $(k = 1, ..., n)$

– k-та порядкова статистика, то щільність ймовірності $f_{(k)}(x)$ k-ї порядкової статистики має β – розподіл із $\alpha = k,\ \beta = n - k + 1.$

$$D_{(\xi)} = \frac{\alpha \cdot \beta}{(\alpha + \beta)^2 \cdot (\alpha + \beta + 1)}.$$

Тому дисперсія порядкових статистик для нечіткої класифікації розраховується за формулою:

$$D_k = \frac{k \cdot (n - k + 1)}{(n+1)^2 \cdot (n+2)}. \qquad (2)$$

Достовірність класифікації визначається за можливими змінами значень порядкових статистик, що приводить до змін НО.

Дисперсія нечітких оцінок із урахуванням (2) розраховується за сумою порядкових статистик. Наприклад, для імені НО 1478 вибірки із рівномірним розподілом

$D_{1478} = \frac{1}{16}\left[D_1 + D_4 + D_7 + D_8\right]$, де за формулою (2):

$D_1 = 10^{-2}$, $D_4 = 2{,}7 \cdot 10^{-2}$, $D_7 = 2{,}3 \cdot 10^{-2}$, $D_8 = 1{,}8 \cdot 10^{-2}$.

Тоді $D_{1478} = 0{,}49 \cdot 10^{-2}$.

Отримані результати для інших НО наведені у табл. 5.

***Таблиця* 5.**

НО і їх дисперсії для вибірки із рівномірним розподілом

Імена	168	267	2369	*med*	1478	348	249
Оцінки	0,399	0,403	0,4745	0,499	0,523	0,595	0,598
Дисперсії	$0{,}61 \cdot 10^{-2}$	$0{,}76 \cdot 10^{-2}$	$0{,}49 \cdot 10^{-2}$	–	$0{,}49 \cdot 10^{-2}$	$0{,}76 \cdot 10^{-2}$	$0{,}61 \cdot 10^{-2}$

Для оцінки вірогідності класифікації визначається критична відстань, тобто відстань h_{kp} між НО, найближчими до медіани.

Для отриманої таблиці:

$$h_{kp} = \text{НО}(1478) - med = med - \text{НО}(2369).$$

Так як НО складається із 3-х або 4-х порядкових статистик, то розподіл нечіткої оцінки можна прийняти приблизно нормальним з дисперсією, що наведена в табл. 3.

Тоді вірогідність зміни лінгвістичного коду визначатиметься як:

$$P = P_1 \cdot P_2,$$

де P_1 – ймовірність знаходження $\text{НО}(1478)$ за областю $\text{НО}(1478) - med$ від $\text{НО}(1478)$, а P_2 – ймовірність знаходження $\text{НО}(2369)$ за областю $med - \text{НО}(2369)$ від $\text{НО}(2369)$, P – ймовірність одночасного знаходження оцінок у наведених областях.

Ймовірність правильної класифікації розподілу: $1-P$, де за умов наближеності розподілу НО до нормального, для НО із табл. 5:

$$P_1 = 0,5 - \Phi\left(\frac{h_{kp}}{\sqrt{D_{1478}}}\right).$$

За умов симетрії оцінок (рис.1):

$$P_2 = P_1.$$

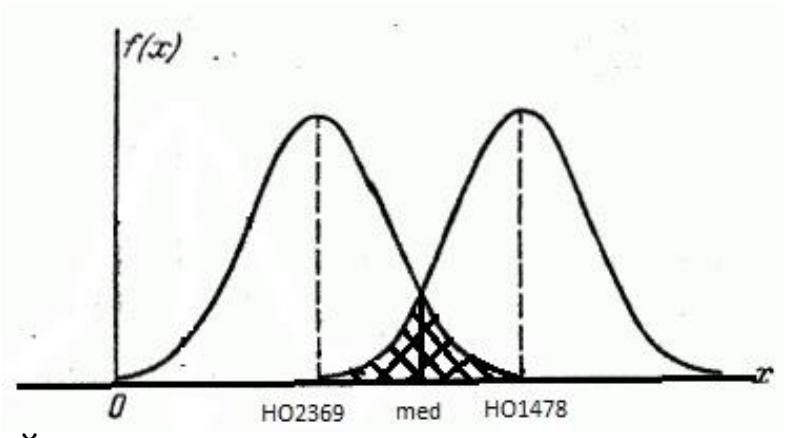

Рис. 1. Ймовірність одночасного знаходження оцінок у наведених областях

За наведеними на рис. 1 оцінками, отримаємо:

$$P_1 = 0,37,\ P = P_1 \cdot P_2 = 0,14 \cdot 100\% = 14\%.$$

Ймовірність правильної класифікації:

$$(1-P)\cdot 100\% = (1-0,14)\cdot 100 = 86\%.$$

Результати оцінки достовірності класифікації, що були проведені для рівномірного та інших розподілів, наведені у табл. 6.

***Таблиця* 6.**

Ймовірність правильної ідентифікації класу розподілу

Найменування розподілу	Клас розподілу	Ймовірність зміни лінгвістичного коду	Ймовірність правильної ідентифікації
Нормальний	Одномодальний симетричний	18%	82%
Коші	Одномодальний симетричний	17%	83%
Рівномірний	Розподіл типу "Шапо"	14%	86%
Лапласа	Лівосторонній	13%	87%

Найменування розподілу	Клас розподілу	Ймовірність зміни лінгвістичного коду	Ймовірність правильної ідентифікації
	несиметричний		
Арксинусний	Двохмодальний симетричний	10%	90%

Розрахунки показали, що ймовірність зміни лінгвістичного коду за умов однієї вибірки доволі висока. Для зменшення дисперсії НО у роботі [2] рекомендується проводити нечітку класифікацію за декількома вибірками малого об'єму.

Проведений авторами аналіз вірогідності дозволяє визначити необхідну кількість вибірок за умов заданої вірогідності класифікації.

Якщо задати, наприклад, ймовірність правильної класифікації $P_k = 95\%$, то можна визначити необхідну кількість вибірок.

Ймовірність зміни лінгвістичного коду: $P = 1 - P_k$,

$$P_1 = \sqrt{1 - P_k},$$

$$\sqrt{1 - P_k} = 0,5 - \Phi\left(\frac{h_{kp}}{\sqrt{D_{1478}}}\right),$$

$$\Phi\left(\frac{h_{kp}}{\sqrt{D_{1478}}}\right) = 0,28,$$

$$n_b = \frac{med}{D_{1478}} \approx 5.$$

Результати розрахунків кількості вибірок, що були проведені для рівномірного та інших розподілів, наведені у табл. 7.

***Таблиця* 7.**

Рекомендована кількість вибірок для різних розподілів за заданою вірогідністю класифікації

Назва розподілу	Кількість вибірок	
	$P_k = 95\%$	$P_k = 99\%$
Нормальний	6	16
Коші	6	16
Рівномірний	5	15
Лапласа	4	14
Арксинусний	4	14

Висновки

У роботі було проведено дослідження можливостей використання методу нечіткої ідентифікації для знаходження розподілу малих вибірок. Знання класу розподілу дозволяє знайти точкові оцінки, найбільш відповідні для даного розподілу і які відповідають вимогам спроможності, незміщеності й ефективності і таким чином зменшити стандартну невизначеність результату вимірювань. Знання розподілу дозволяє знайти значення коефіцієнту охоплення, який теж відповідає визначеному розподілу, та зменшити розширену невизначеність, не використовуючи оцінку зверху.

Проведене дослідження показало, що нечітка ідентифікація дозволяє визначити лише класи розподілів: одномодальний симетричний, одномодальний несиметричний, двухмодальний, рівномірний,тобто провести нечітку класифікацію.

Визначена достовірність віднесення вибірок до певного класу за допомогою генерації опорних вибірок. Дослідження достовірності за однією вибіркою показало, що ймовірність зміни лінгвістичного коду досить велика. Для підвищення ймовірності правильної ідентифікації рекомендується усереднення декількох вибірок. Кількість вибірок може бути визначена за заданою вірогідністю.

Дослідження нечіткої класифікації включало у себе розробку алгоритму знаходження нечітких оцінок для вибірок різного розміру, знаходження необхідної і достатньої кількості оцінок й принципи вибору використовуваних оцінок із усього набору можливих. Отримано співвідношення для нечітких оцінок у залежності від заданого об’єму. Використання цих співвідношень дозволило розширити сферу нечіткої класифікації і створити бібліотеку лінгвістичних кодів, яка значно спростить її процедуру. Отримано рекомендації по кількості вибірок для отримання необхідної вірогідності класифікації.

Список використаної літератури

1. *Дорожовець М.* Опрацювання результатів вимірювань, Навч. посібник. – Львів: Видавництво Національного університету «Львівська політехніка», 2007. – 624 с.
2. *Кликушин Ю. Н.* Представление случайных сигналов с помощью принадлежностных спектров. – Омский государственный технический университет: «Журнал радиоэлектроники», №2, 2000.
3. *Новицкий П. В.* Оценка погрешностей результатов измерений. / П. В. Новицкий, И. А. Зограф //1985. – 248 с.
4. Справочник по теории вероятностей и матиматической статистике / В. С. Королюк, Н. И. Портенко, А. В. Скороход, А. Ф. Турбин – М.: Наука, 1985. – 640 с.