УДК 621.391
DOI https://doi.org/10.26661/2413-6549-2021-1-04

НАЛАШТУВАННЯ ТА НАВЧАННЯ НЕЧІТКОЇ МОДЕЛІ ДЛЯ ЗАДАЧІ КЛАСИФІКАЦІЇ

Єгошкін Д. І.
аспірант,
асистент кафедри комп'ютерних технологій
Дніпровський національний університет імені Олеся Гончара
пр. Гагаріна, 72, Дніпро, Україна
orcid.org/0000-0002-0937-4733
KnightDanila@i.ua

Гук Н. А.
доктор фізико-математичних наук, професор,
завідувач кафедри комп'ютерних технологій
Дніпровський національний університет імені Олеся Гончара
пр. Гагаріна, 72, Дніпро, Україна
orcid.org/0000-0001-7937-1039
natalyguk29@gmail.com

Ключові слова: *нечітка класифікація, нечітка логіка, база знань, експертна система, квантільні оцінки, мова C/C++, мова JavaScript, JSON.*

У роботі розглянуто задачу класифікації об'єктів за ознаками та розроблено метод нечіткої класифікації. Для розв'язання задачі запропоновано використовувати нечітку модель представлення знань, побудовану з використанням навчальної вибірки, та систему нечіткого логічного виводу. Автоматичне формування системи нечітких логічних правил відбувається в процесі навчання. У процесі навчання відбувається налаштування параметрів моделі – нечітких границь термів. Для усунення проблеми обробки об'єктів, характеристики яких суттєво відрізняються від більшості об'єктів у вибірці та наближаються до порогових, для визначення границь термів пропонується використовувати квартильні оцінки. Модифікація класичного алгоритму нечіткої класифікації полягає в організації двохетапної процедури класифікації. Для поліпшення якості класифікації будуються допоміжні нечіткі класи, з використанням яких здійснюється відображення об'єктів в чіткі класи з навчальної вибірки. Для оцінювання якості класифікації використовуються метрики точності та повноти. Налаштування параметрів алгоритму нечіткої класифікації та розв'язання задачі нечіткої класифікації виконано з використанням набору даних іриси Фішера. Наведено порівняння результатів класифікації з використанням запропонованого в роботі двохетапного походу і класичного алгоритму нечіткої класифікації. Для зменшення впливу наявних у навчальній вибірці аномальних об'єктів на результат класифікації для визначення границь термів використовується міжквартильне середнє, що робить процедуру визначення границь термів робастною. Показано, що використання міжквартильного середнього для визначення границь термів дозволяє отримати прийнятну точність класифікації для вибірок, які містять об'єкти з аномальними характеристиками.
Досліджено вплив способу розбиття вибірки на навчальну і тестову, а також вплив розміру навчальної вибірки на якість класифікації.
Запропонований підхід є придатним для обробки даних в умовах обмеженої вибірки, часткового чи повного перекриття класів об'єктів, наявності об'єктів з нетиповими значеннями характеристик.

SETTING UP AND TRAINING A FUZZY MODEL FOR A CLASSIFICATION PROBLEM

Yehoshkin D. I.
Postgraduate Student,
Assistant at the Department of Computer Technology
Oles Honchar Dnipro National University
Gagarin avenue, 72, Dnipro, Ukraine
orcid.org/0000-0002-0937-4733
KnightDanila@i.ua

Huk N. A.
Doctor of Sciences in Physics and Mathematics, Professor,
Head of the Department of Computer Technology
Oles Honchar Dnipro National University
Gagarin avenue, 72, Dnipro, Ukraine
orcid.org/0000-0001-7937-1039
natalyguk29@gmail.com

Key words: *fuzzy classification, fuzzy logic, knowledge base, expert system, quantile estimates, C/C++ language, JavaScript language, JSON.*

The paper considers the problem of classifying objects by attributes and developed a method of fuzzy classification. To solve the problem, it is proposed to use a fuzzy model of knowledge representation, built using a training sample and a fuzzy logical inference system. The automatic formation of a system of fuzzy logical rules occurs in the learning process. In the process of training, the parameters of the model are adjusted – the fuzzy boundaries of the terms. To eliminate the problem of processing objects, the characteristics of which differ significantly from the majority of objects in the sample and approach the threshold, it is proposed to use quartile estimates to determine the boundaries of terms.

A modification of the classical fuzzy classification algorithm consists of organizing a two-stage classification procedure. To improve the quality of classification, auxiliary fuzzy classes are constructed, with the use of which objects are mapped into crisp classes of the training set. To estimate the quality of the classification, metrics of accuracy and completeness are used. The settings of the fuzzy classification algorithm and the solution of the fuzzy classification problem are performed using the Fisher Iris dataset. The comparison of classification results using the proposed two-stage approach and the classical fuzzy classification algorithm is presented. To reduce the influence of anomalous objects present in the training set on the classification result, the midhinge is used to determine the boundaries of terms, which makes the procedure for determining the boundaries of terms robust. It is shown that the use of the midhinge for determining the boundaries of terms obtains acceptable classification accuracy for samples containing objects with anomalous characteristics.

The influence of the method of dividing the sample into training-testing and the influence of the size of the training sample on the quality of classification is investigated.

The proposed approach is suitable for data processing in conditions of a limited set, in a partial or complete overlap of object classes, the presence of objects with atypical values of characteristics.

Вступ. Однією із задач аналізу даних є задача класифікації. Її розв'язок дозволяє розбити множину об'єктів на заздалегідь створені групи (класи) на основі аналізу їх формального опису, що дозволяє прискорити процес подальшої обробки даних. При класифікації кожен об'єкт спостереження відноситься до певної групи або номінальної категорії на основі деякої якісної властивості або сукупності властивостей.

Задачам класифікації присвячено багато робіт. До числа поширених методів розв'язання задачі класифікації відносяться: нейронні мережі; логістична і пробіт-регресія; дерева рішень; метод найближчого сусіда; машини опорних векторів; дискримінантний аналіз.

Традиційні підходи, засновані на апараті математичної статистики або імітаційному моделюванні, не дозволяють будувати адекватні моделі в умовах обмеженості часових, обчислювальних і матеріальних ресурсів. Тому при розв'язанні багатьох практичних задач, пов'язаних з класифікацією об'єктів, широко застосовуються моделі і методи штучного інтелекту з використанням технологій інтелектуального аналізу даних.

Так, у роботі [1] для розв'язання задачі класифікації застосовується апарат нечіткої логіки. Для класифікації об'єктів з дискретним виходом застосовується нечіткий логічний вивід типу Сугено та розширюються функціональні можливості пакету Fuzzy Logic Toolbox системи MATLAB. Взаємозв'язок між чотирма вхідними і однією вихідної змінної описується трьома нечіткими правилами, які з точністю 85% здатні класифікувати вибірку «Іриси Фішера».

У роботі [2] досліджується застосування нечітких моделей в задачах класифікації на основі представлення даних у вигляді нечітких градацій. Введено класи об'єктів, еталонні зразки в кожному класі, а кожен клас характеризується розподіленої областю значень нечітких критеріїв. Початкові дані для класифікації є суперечливими, а об'єкти, еталони і класи перетинаються, тому побудова розв'язку задачі в класичній постановці викликає значні труднощі. Класифікація розглядається як різновид задачі прийняття рішень, в якій за допомогою узагальнення нечітких фактів, що характеризують властивості, стан або зміну станів об'єктів, здійснюється вибір найкращого класу для кожного об'єкту. Проведено порівняння різних мір узгодженості об'єктів з класами, досліджено їх вплив на результати класифікації.

У роботі [3] розвивається метод гірської кластеризації Ягера-Фільова на випадок нечітких представлень простору станів і ознак, який заснований на обчисленні щільності розподілу інтегральних характеристик об'єктів в нечіткому просторі станів. Із використанням нечіткої відстані Хеммінга визначено приналежність об'єкту при перетині областей нечіткого розподілу ознак.

У роботі [4] розглядається актуальна на даний момент проблема генерації набору нечітких правил для системи нечіткого виводу Мамдані на основі числових даних, отриманих у процесі навчання системи керування. Запропонований підхід базується на алгоритмах чіткої і нечіткої кластеризації – алгоритмі гірської кластеризації і алгоритмі Густафсона-Кесселя.

У побудові систем класифікації на основі нечіткої логіки важливим є правильна побудова системи правил та налаштування параметрів моделі. Зазвичай для виконання таких робіт залучають групи експертів. Однак використання результатів спостережень та автоматична побудова системи логічного виведення на їх основі дозволить мінімізувати людський фактор, скоротити час та витрати на налаштування моделі.

У даній роботі розвивається підхід щодо автоматизації побудови системи класифікації з використанням навчальної вибірки та моделі нечіткої логіки. Метод пропонує автоматичне формування системи правил на основі аналізу об'єктів навчальної вибірки, налаштування параметрів моделі в процесі навчання, врахування наявності аномальних об'єктів у навчальної вибірці.

Для покращення якості класифікації використовуються допоміжні нечіткі класи приналежності, за допомогою яких здійснюється подальше відображення об'єктів в чіткі класи з навчальної вибірки.

Постановка задачі. Розглядається множина об'єктів деякої предметної області. Об'єкти множини X характеризуються деякими ознаками K. Існує скінченна множина класів Z, серед яких необхідно розподілити об'єкти.

Для побудови системи класифікації використовуються результати спостережень за об'єктами, ознаки вимірюються та подаються числовими значеннями, для кожного з об'єктів визначено відповідний клас, до якого він належить. Такий набір даних формує навчаючу вибірку виду:

$$X^n = \{(x_i, z_j)\},\ i = \overline{1,N},\ j = \overline{1,J}$$

де (x_i, z_j) – пара об'єкт-клас; $x_i \in X^n$ – об'єкт; z_j – клас, до якого належить об'єкт x_i, $z_j \in Z$. Кожен об'єкт x_i характеризується вектором ознак $\overline{K} = (k_1, k_2, ..., k_l), k_l \in R$.

Навчаюча вибірка для скінченого числа об'єктів описує відображення $F^*: X \to Z$, за допомогою якого можна визначити приналежність певного об'єкту до певного класу об'єктів.

Необхідно побудувати відображення $F: X \to Z$, за допомогою якого можливо класифікувати довільний об'єкт з множини X.

У класичній постановці задача класифікації передбачає, що в результаті розбиття будуть отри-

мані детерміновані класи об'єктів, і кожен об'єкт належить тільки одному класу. Однак на практиці такий похід у деяких випадках призводить до аналітичної невизначеності, тому в роботі розвивається підхід, заснований на представленні об'єктів і класів об'єктів у вигляді нечітких даних, який передбачає розробку системи нечіткого логічного виводу для розв'язання задачі класифікації та налаштування параметрів моделі в процесі виконання процедури класифікації.

Математична модель задачі. Для опису об'єктів предметної області будемо використовувати нечітку модель представлення знань, а для виконання процедури класифікації – систему нечіткого логічного виводу.

Будемо вважати, що кожна ознака об'єкта описується лінгвістичною змінною. Для кожної ознаки формується терм-множина, елементами якої є нечіткі змінні, задаються границі термів.

Для відображення чітких вхідних значень ознак k_l – в нечіткі множини вводяться функції приналежності M_{lt} виду:

$$M_{lt}(k_l,a_{lt},b_{lt},c_{lt},d_{lt})=\begin{cases}0, & k_l\le a_{lt}\\ \dfrac{k_l-a_{lt}}{b_{lt}-a_{lt}}, & a_{lt}\le k_l\le b_{lt}\\ 1, & b_{lt}\le k_l\le c_{lt}\\ \dfrac{d_{lt}-k_l}{d_{lt}-c_{lt}}, & c_{lt}\le k_l\le d_{lt}\\ 0, & d_{lt}\le k_l\end{cases} \quad (1)$$

де M_{lt} – функції приналежності ознаки k_l терму t; $a_{lt},b_{lt},c_{lt},d_{lt}$ – числові параметри, які визначають границі термів та впорядковані відношенням $a_{lt}\le b_{lt}\le c_{lt}\le d_{lt}$.

Для підвищення якості класифікації нечітка модель представлення знань потребує налаштування. Для визначення границь термів будемо використовувати дані навчальної вибірки. Аналіз великих обсягів спостережень демонструє, що в результатах можуть бути присутніми нетипові спостереження, значення яких не можуть бути описані загальними закономірностями. Наявність таких даних може бути пов'язано з похибкою у вимірюваннях, аномаліями в розподілі даних або факторами, які не були враховані при побудові моделі. Значення, що суттєво відхиляються, можуть істотно погіршити налаштування процедури класифікації, оскільки класифікатор буде намагатися пояснити нетипові спостереження.

У роботі для обчислення границь термів пропонується використовувати квартильні оцінки, а саме міжквартильне середнє (MH – Midhinge), значення якого обчислюється за навчальною вибіркою. Застосування такого підходу робить процедуру визначення границь термів робастною.

Для визначення межквартильного середнього набір даних з навчальної вибірки впорядковується та ділиться на чотири частини, потім обчислюється середнє значення між першим і третім квартилями [5]:

$$MH=\frac{Q1+Q3}{2}, \quad (2)$$

$$Q1=x_{\min Q1}+\Delta x_{Q1}\frac{\frac{N}{4}-S_{Q1-1}}{n_{Q1}},$$

$$Q3=x_{\min Q3}+\Delta x_{Q3}\frac{\frac{3N}{4}-S_{Q3-1}}{n_{Q3}},$$

де Q1, Q3 – медіани h найменших і h найбільших значень відповідно; $x_{\min Q1}$, $x_{\min Q3}$ – нижні границі інтервалів, що містять перший і третій квартиль; $\Delta x_{Q1},\Delta x_{Q3}$ – ширина інтервалів; S_{Q1-1}, S_{Q3-1} – накопичена частота інтервалу, що передує даному – може бути визначена, як сума всіх попередніх частот до поточної; n_{Q1}, n_{Q3} – частота попадання значень навчальної вибірки в інтервали, що містять перший і третій квартилі відповідно.

Границі термів визначаються відносно елементів навчальної вибірки у такий спосіб (рівняння 3), де MH – міжквартильне середнє; $t=\overline{1,T}$, $T\ge 2$, T – кількість елементів терм-множини лінгвістичної змінної.

У роботі пропонується модифікація класичного алгоритму нечіткої класифікації, яка полягає в організації двохетапної процедури класифікації. Із застосуванням введеної модифікації спочатку виконується проміжна класифікація об'єктів всередині системи, в результаті чого будуються класи нечітких об'єктів [6]. На другому етапі, виходячи з нечітких класів Y, відбувається класифікація стосовно чітких класів Z за допомогою навчальної вибірки та введених метрик. Такий підхід дозволяє уникнути помилкових результатів класифікації у випадку, коли класи розташовані поблизу один від одного, а можливо, і перетинаються.

Для організації першого етапу класифікації вводяться проміжні нечіткі класи Y нечітких об'єктів. Під нечіткими класами розуміються класи з властивостями, які є загальними для нечітких об'єктів, що належать різним класам. Тобто об'єкт $x_i\in X$ може одразу належати до декількох класів з певною мірою істинності. Відображення об'єкта $x_i\in X$ в клас Y можна представити функціональною залежністю від вектору ознак $(k_1,k_2,...,k_l)$:

$$Y=f(k_1,k_2,...,k_l), \quad (4)$$

де f – дійсна функція чітких значень $(k_1,k_2,...,k_l)$.

$$\underline{\bar{k}}_t = \begin{cases} \left(\min\limits_{i=1}\left(k_i\right), \max\limits_{i=1}\left(k_i\right)\right), \ t=1, T=2 \\ \left(\min\limits_{i=1}\left(k_i\right), \max\limits_{i=1}\left(k_i\right)\right), \ t=T, T=2 \\ \left(\min\limits_{i=1}\left(k_i\right), \min\limits_{i=1}\left(k_i\right) + 2\cdot\dfrac{2\cdot\left|\mathrm{MH}\left(k_i\right)_{i=1} - \min\left(k_i\right)_{i=1}\right|}{T+1}\right), t=1, T>2 \\ \left(\min\limits_{i=1}\left(k_i\right) + \dfrac{(t-1)\cdot 2\cdot\left|\mathrm{MH}\left(k_i\right)_{i=1} - \min\left(k_i\right)_{i=1}\right|}{T+1}, \max\limits_{i=1}\left(k_i\right) - \dfrac{(T-t)\cdot 2\cdot\left|\mathrm{MH}\left(k_i\right)_{i=1} - \max\left(k_i\right)_{i=1}\right|}{T+1}\right), 1<t<T, T>2 \\ \left(\max\limits_{i=1}\left(k_i\right) - 2\cdot\dfrac{2\cdot\left|\mathrm{MH}\left(k_i\right)_{i=1} - \max\left(k_i\right)_{i=1}\right|}{T+1}, \max\limits_{i=1}\left(k_i\right)\right), t=T, T>2 \end{cases}, \quad (3)$$

Рівняння 3

У роботі залежність (4) зображується системою нечітких логічних правил. Для забезпечення повноти системи правил потрібна наявність хоча б одного правила для кожного терму вихідної змінної, а при формуванні їх умовних частин складаються всі можливі комбінації термів вхідних змінних.

Система правил будується у вигляді:

$$\begin{aligned}
&\Pi_1 : Якщо\ k_1 \epsilon A_{11} \wedge k_2 \epsilon A_{21} \wedge k_3 \epsilon A_{31} \wedge \ldots \wedge k_l \epsilon A_{l1}\ \ TO\ \ Y = Y_1, \\
&\Pi_2 : Якщо\ k_1 \epsilon A_{12} \wedge k_2 \epsilon A_{21} \wedge k_3 \epsilon A_{31} \wedge \ldots \wedge k_l \epsilon A_{l1}\ \ TO\ \ Y = Y_m, \\
&\Pi_3 : Якщо\ k_1 \epsilon A_{11} \wedge k_2 \epsilon A_{22} \wedge k_3 \epsilon A_{31} \wedge \ldots \wedge k_l \epsilon A_{l1}\ \ TO\ \ Y = Y_m, \\
&\Pi_4 : Якщо\ k_1 \epsilon A_{11} \wedge k_2 \epsilon A_{21} \wedge k_3 \epsilon A_{32} \wedge \ldots \wedge k_l \epsilon A_{l1}\ \ TO\ \ Y = Y_m, \\
&\ldots \\
&\Pi_{n+1} : Якщо\ k_1 \epsilon A_{11} \wedge k_2 \epsilon A_{21} \wedge k_3 \epsilon A_{31} \wedge \ldots \wedge k_l \epsilon A_{l2}\ \ TO\ \ Y = Y_m, \\
&\Pi_{n+2} : Якщо\ k_1 \epsilon A_{12} \wedge k_2 \epsilon A_{22} \wedge k_3 \epsilon A_{31} \wedge \ldots \wedge k_l \epsilon A_{l1}\ \ TO\ \ Y = Y_{m+1}, \\
&\Pi_{n+3} : Якщо\ k_1 \epsilon A_{12} \wedge k_2 \epsilon A_{21} \wedge k_3 \epsilon A_{32} \wedge \ldots \wedge k_l \epsilon A_{l1}\ \ TO\ \ Y = Y_{m+1}, \\
&\ldots \\
&\Pi_{2n+1} : Якщо\ k_1 \epsilon A_{12} \wedge k_2 \epsilon A_{21} \wedge k_3 \epsilon A_{31} \wedge \ldots \wedge k_l \epsilon A_{l2}\ \ TO\ \ Y = Y_{m+1}, \\
&\ldots \\
&\Pi p : Якщо\ k_1 \epsilon A_{1T} \wedge k_2 \epsilon A_{2T} \wedge k_3 \epsilon A_{3T} \wedge \ldots \wedge k_l \epsilon A_{lT}\ \ TO\ \ Y = Y_M,
\end{aligned} \quad (5)$$

де $p = \overline{1, P}$ – номер правила в базі правил; P – загальна кількість правил; A_{lt} – нечіткий терм, що характеризує ознаку k_l для правила p; $t = \overline{1, T}$, T – кількість термів для ознаки k_l; Y_m – мітка проміжного класу, до якого належить об'єкт $\mathrm{x_i} \in \mathrm{X}$; $m = \overline{1, M}$, M – кількість класів Y.

Загальна кількість логічних правил визначається кількістю всіх можливих комбінацій лінгвістичних значень A_{lt} і відповідає умові:

$$P \le \prod_{l=1}^{\mathrm{L}} \mathrm{T}_l$$

де L – розмірність вектора ознак $\overline{K}: (k_1, k_2, ..., k_l)$; T_l – кількість термів для ознаки k_l.

Метод розв'язання. Класифікація здійснюється на основі алгоритму нечіткого логічного виводу. На відміну від класичного підходу, формування системи правил виконується автоматично на етапі навчання. З використанням навчальної вибірки об'єкти якої мають однакову вимірність в просторі ознак, формулюється система правил (5).

Для виконання ідентифікації об'єкта застосовано алгоритм нечіткого логічного виводу [7], модифікований введенням додаткового механізму порівняння чіткого вихідного значення Y^* та ступеню приналежності M^*, отриманого для об'єкта $X^* = (k_1, k_2, ..., k_l)$, і вихідних значень Y_i^* та M_i^* для об'єктів $X^n = \{(\mathrm{x_i}, \mathrm{z_j})\}$, наявних в базі знань.

Для виконання процедури фазифікації вхідних змінних k_l в нечіткі множини A_{lt} будемо використовувати операцію [8]:

$$A_{lt} = \int_{\underline{k}_t}^{\overline{k}_t} (M_{lt}(k_l)/k_l)dk \quad (6)$$

де $\underline{k}_t, \overline{k}_t$ – границі термів.

Ступінь приналежності вхідного об'єкта $X^* = (k_1, k_2, ..., k_n)$ нечітким класам Y_m з бази знань (5) описується системою нечітких логічних рівнянь:

$$M_{Y_m}(X) = \bigvee_{p=\overline{1,P}} \bigwedge_{l=\overline{1,L}} \left[M_{lt}(k_l) \right], \; t = \overline{1,T} \qquad (7)$$

де оператори $\vee$ та $\wedge$ відповідають виконанню логічних операцій «АБО» та «І» відповідно. У роботі використані їх реалізації у вигляді знаходження max та min .

Нечітка множина $\tilde{Y}$ проміжних класів, що відповідає вхідному об'єкту X визначається у вигляді:

$$\tilde{Y} = \underset{m=\overline{1,M}}{agg}\left(\int_{\underline{Y}}^{\overline{Y}} imp(M_{Y_m}(X), M_{Y_m}(Y)/Y)dY \right) \qquad (8)$$

де imp – операція імплікації, agg – операція агрегування, які реалізовані операцією знаходження min та max відповідно.

Чітке значення виходу Y^* визначається в результаті дефазифікації нечіткої множини $\tilde{Y}$ за методом центру тяжіння:

$$Y^* = \int_{\underline{Y}}^{\overline{Y}} Y \cdot M_{\tilde{Y}}(Y)dY \Big/ \int_{\underline{Y}}^{\overline{Y}} M_{\tilde{Y}}(Y)dY \qquad (9)$$

Ступінь приналежності M^* вхідного об'єкта $X^* = (k_1, k_2, ..., k_l)$ визначається в результаті дефазифікації нечіткої множини $\tilde{Y}$ за методом центру тяжіння:

$$M^* = \int_{\underline{Y}}^{\overline{Y}} C \cdot M_{\tilde{Y}}(Y)dY \Big/ \int_{\underline{Y}}^{\overline{Y}} M_{\tilde{Y}}(Y)dY \qquad (10)$$

де C – константа, значення якої обирається експериментально в залежності від типу функції приналежності та елементів навчальної вибірки.

Для організації процедури порівняння визначимо відстань між об'єктом X^* та об'єктами навчальної вибірки $X^n = \{(\mathrm{x_i}, \mathrm{z_j})\}$. Відстань визначається на основі обраної метрики в просторі характеристик. Для оцінки міри близькості елементів використовується Евклідова відстань:

$$d((Y^*, M^*), (Y_i^*, M_i^*)) = \sqrt{(Y^* - Y_i^*)^2 + (M^* - M_i^*)^2} \qquad (11)$$

$$d((Y^*, M^*), (Y_i^*, M_i^*)) < \varepsilon \qquad (12)$$

Після виконання процедури дефазифікації обчислюється $d((Y^*, M^*), (Y_i^*, M_i^*))$ та перевіряється умова (12).

Далі будуємо множину $\overset{d}{Y}$, до якої відносяться об'єкти навчальної вибірки $X^n = \{(\mathrm{x_i}, \mathrm{z_j})\}$ схожі з об'єктом X^*, якщо $d((Y^*, M^*), (Y_i^*, M_i^*)) < \varepsilon$, то $(\mathrm{x_i}, \mathrm{z_j}) \in \overset{d}{Y}$.

Клас, до якого належить вхідний об'єкт X^*, визначається у такий спосіб:

$$\max_{j=1}^{M} \left(\underset{\mathrm{z_j}}{card}(\overset{d}{Y}) \right), \qquad (13)$$

де $\underset{\mathrm{z_j}}{card}(\overset{d}{Y})$ – потужність множини $\overset{d}{Y}$.

Вказану послідовність дій можна описати наступним алгоритмом.

Алгоритм:

Крок 0. Ініціалізація. Задати навчальну вибірку $X^n = \{(\mathrm{x_i}, \mathrm{z_j})\}$; значення ε – похибка системи; кількість нечітких класів Y.

Крок 1. За допомогою навчальної вибірки $X^n = \{(\mathrm{x_i}, \mathrm{z_j})\}$, побудувати функції приналежності (1) та розрахувати границі термів лінгвістичних змінних A_{lt} за формулою (3).

Крок 2. Побудувати систему правил у вигляді (5).

Крок 3. За допомогою механізму нечіткого логічного виведення (6) – (10) розрахувати ступені приналежності об'єктів навчальної вибірки X^n нечітким множинам $\tilde{Y}$, визначити чітке значення Y^* та M^*, зберегти ці значення разом з навчальною вибіркою $X^n = \{(\mathrm{x_i}, \mathrm{z_j}, \tilde{Y}_i, Y_i^*, M_i^*)\}$.

Крок 4. Для вхідного об'єкту X^* на базі сформованих продукційних правил за допомогою (6) – (8) розрахувати ступені приналежності об'єкта нечітким множинам $\tilde{Y}^*$. За формулами (9)-(10) отримати чітке значення Y^* та M^*.

Крок 5. З множини об'єктів $X^n = \{(\mathrm{x_i}, \mathrm{z_j}, \tilde{Y}_i, Y_i^*, M_i^*)\}$ обрати лише ті, для яких $\tilde{Y}_i = \tilde{Y}^*$.

Крок 6. Обчислити відстані між об'єктом X^* та обраними об'єктами з навчальної вибірки $X^n = \{(\mathrm{x_i}, \mathrm{z_j}, \tilde{Y}_i, Y_i^*, M_i^*)\}$ за формулою (11), перевірити виконання умови (12).

Якщо нерівність (12) є вірною, побудувати множину Y та перейти до кроку 8, інакше перейти до кроку 7.

Крок 7. Об'єкт неможливо класифікувати.

Крок 8. Для побудованої множини $\overset{d}{Y}$, знайти потужність для кожного з класів $\mathrm{z_j}$ з $X^n = \{(\mathrm{x_i}, \mathrm{z_j}, \tilde{Y}_i, Y_i^*, M_i^*)\}$.

Крок 9. Отримати результат ідентифікації з умови (13).

Розроблений алгоритм, було реалізовано у вигляді програмного забезпечення з використанням мов C/C++ та JavaScript, а також текстового формату обміну даними JSON. Для розробки використовувалися платформи NetBeans IDE, WhiteStarUML, GitHub, WebGL, Chrome, Mozilla Firefox, Opera.

Аналіз результатів класифікації. Запропонований підхід до класифікації об'єктів за допомогою нечіткої логіки та елементів навчальної вибірки був протестований для даних з відомої задачі класифікації Iris Data Set – Іриси Фішера [9]. Задача класифікації передбачає визначення належності ірису до одного з 3 типів рослин: Setosa, Versicolor, Virginica.

Навчальна вибірка $X^n = \{(\mathrm{x_i}, \mathrm{z_j})\}$ складається зі 150 об'єктів $\mathrm{x_i}$, кожен з яких належить одному з трьох класів $\mathrm{z_j}$. Кожен клас містить по 50 елементів вибірки. В ролі ознак об'єкта використовуються: довжина чашолистка – sepalLength, ширина чашолистка – sepalWidth, довжина пелюстки – petalLength, ширина пелюстки – petalWidth. Вихідна змінна Y характеризує проміжні класи об'єктів: клас I, клас II, клас III, кількість проміжних нечітких класів Y співпадає з кількістю вихідних класів класифікації, які зазначено у вибірці.

Для лінгвістичних змінних, які описують кожну з ознак, вводяться терми low, mid, high. Для кожної вхідної та вихідної змінної вводяться функції приналежності виду (1), обчислюються границі термів за формулами (3).

Далі формується система правил (5), передумови яких складені з усіх можливих комбінацій значень нечітких вхідних змінних (всього 81 правило).

Для апробації запропонованого підходу до класифікації об'єктів навчальна вибірка $X^n = \{(x_i, z_j)\}$ була поділена на дві частини: навчальну Q – 120 ірисів и тестову V – 30 ірисів.

Розглянемо етапи виконання процедури класифікації об'єкту X =(6.3, 2.7, 4.9, 1.8), $X \in V$. На першому етапі класифікації з використанням сформованих продукційних правил та шляхом виконання процедур (6) – (8) було розраховано ступені приналежності об'єкта нечітким множинам $\tilde{Y}^*$. За формулами (9) – (10) отримано дефазифіковані значення $Y^* = 60.35$, $M^* = 0.46$. Визначено, що за значенням $Y^* = 60.35$, об'єкт X належить до проміжних класів II та III. Далі за допомогою (11), (12) побудовано множину $\overset{d}{Y}$ і визначено клас, до якого належить об'єкт, за допомогою (13):

$$\max\left(\underset{\text{setosa}}{card}(\overset{d}{Y}), \underset{\text{versicolor}}{card}(\overset{d}{Y}), \underset{\text{virginica}}{card}(\overset{d}{Y}) \right) =$$

$$= \max(0,3,7) = 7 \Rightarrow \underset{\text{virginica}}{card}(\overset{d}{Y})$$

За результатами виконаних розрахунків отримано, що об'єкт належить класу Virginica, що відповідає навчальної вибірці [9].

Порівняння результатів класифікації з використанням запропонованого в роботі двохетапного підходу і класичного алгоритму нечіткої класифікації [1] наведено на рис. 1.

У класичному алгоритмі використовувалась модель нечіткого логічного виводу типу Сугено з чотирма вхідними і однією вихідної змінною, функції приналежності побудовано за середнім значеннями. При використанні класичного підходу об'єкт X було віднесено до класу Versicolor, результат фазифікації вихідної змінної, отриманий методом центру тяжіння, наведено на рис. 1 а.

Результат класифікації за допомогою запропонованого підходу наведено на рис. 1. b). Можна бачити, що навколо об'єкту X (на рис. 1б позначений трикутником) є 3 об'єкти класу Versicolor (на рис. 1б позначені квадратами) та 7 об'єктів класу Virginica (на рис. 1б позначені кругами), що за (13) свідчить про належність об'єкту X до проміжного класу III з більшою ймовірністю, ніж до класу II. Проведений аналіз дозволяє визначити належність об'єкту X до класу Virginica, що співпадає з результатом класифікації за вибіркою [9].

У таблиці 1 подано деякі результати класифікації елементів тестової вибірки V.

Для оцінки якості класифікації вводяться метрики:

accuracy – частина правильних відповідей моделі:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN};$$

precision – точність:

$$\text{precision} = \frac{TP}{TP + FP};$$

recall – повнота:

$$\text{recall} = \frac{TP}{TP + FN};$$

f1-score – f-міра:

$$\text{f1-score} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}},$$

де TP- True Positive; FP – False Positive; FN – False Negative; TN – True Negative.

Із використанням введених метрик для результату класифікації, наведеного в таблиці 1, отримані наступні значення: accuracy=0,87; precision= 0,87; recall = 0,83; f1-score=0,84.

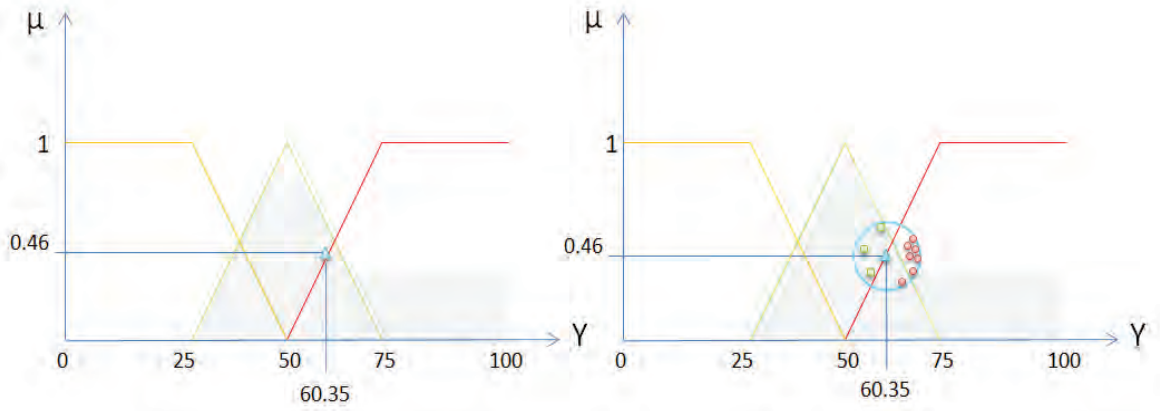

Рис. 1: a) Класичний алгоритм нечіткої класифікації; b) Розроблений метод класифікації

На рис. 2 наведено розподіл об'єктів навчальної вибірки в двовимірному просторі на три проміжні класи, можна зазначити, що проміжні класи об'єктів навчальної вибірки добре розрізняються.

Розглянемо вплив запропонованого способу робастного визначення границь термів у випадку появи в навчальній вибірці $X^n = \{(x_i, z_j)\}$ об'єктів з аномальними значеннями, ознаки об'єктів, доданих до вибірки, наведені в табл. 2, аномальні значення виділені сірим кольором.

У таблиці 3 представлені результати визначення границь термів, які обчислені з використанням МН (3) і з застосування АМ (Arithmetic mean) – середнього арифметичного значення для вибірок, які не містять і містять аномальні об'єкти.

Можна відзначити, що при додаванні в навчальну вибірку аномальних об'єктів, різниця між границями термів обчислених за допомогою МН і АМ змінюється в діапазоні від 0.015 до 3.67. Найбільші відмінності спостерігаються при визначенні границь термів лінгвістичної

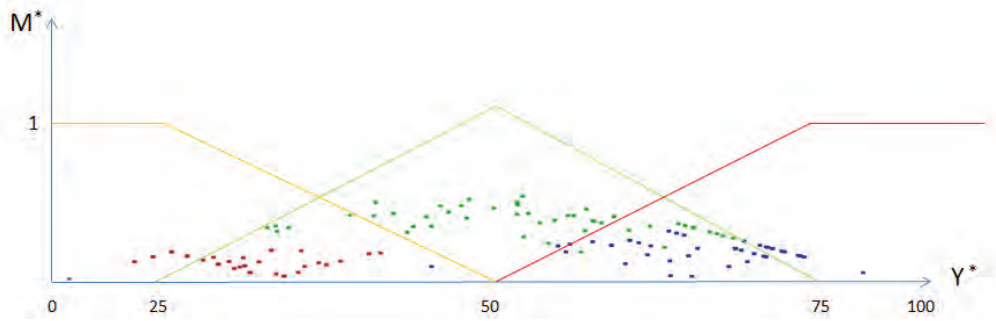

Рис. 2. Загальний розподіл об'єктів

Таблиця 1

X	Вхідні параметри				Дефазифіцировані значення Y* та M^*	Результат 1-го етапу класифікації (проміжної класифікації)	$card^{d}_{z_j}(Y)$			Результат класифікації	Дійсне значення
	sepalLength	sepalWidth	petalLength	petalWidth			z_j = Setosa	z_j= Versicolor	z_j= Virginica		
1	4.9	3.1	1.5	0.2	31.25, 0.45	Клас_I Клас_II	8	0	0	Setosa	Setosa
10	5.1	3.3	1.7	0.5	33.81, 0.46	Клас_I Клас_II	7	0	0	Setosa	Setosa
14	5.8	2.7	3.9	1.2	48.97, 0.38	Клас_I Клас_II	0	5	0	Versicolor	Versicolor
20	5.6	3	4.5	1.5	55.49, 0.4	Клас_II Клас_III	0	5	0	Versicolor	Versicolor
21	7.7	2.6	6.9	2.3	63.06, 0.49	Клас_II Клас_III	0	0	5	Virginica	Virginica
26	6.3	2.5	5	1.9	59.10, 0.47	Клас_II Клас_III	0	0	3	Virginica	Virginica
27	6	2.7	5.1	1.6	57.81, 0.44	Клас_II Клас_III	0	4	6	Virginica	Versicolor
30	6.3	2.5	4.9	1.5	56.28, 0.45	Клас_II Клас_III	0	6	3	Virginica	Versicolor

Таблиця 2

sepalLength	sepalWidth	petalLength	petalWidth	Клас
0.3	0.1	0.1	0.8	Setosa
0.9	0.3	510.1	0.8	Virginica

змінної petalLength (в табл. 3 виділено сірим кольором).

Покажемо вплив робастного визначення границь термів МН на результат класифікації. На рис. 3, рис. 4 наведено розподіл об'єктів між проміжними класами, який отримано для термів, границі яких визначалися з використанням АМ (рис. 3) і МН (рис. 4).

У випадку, коли границі термів обчислювались з використання АМ (рис. 3), отримані класи об'єктів Setosa (об'єкти позначено трикутником), Versicolor (об'єкти позначено квадратами), Virginica (об'єкти позначено колом) погано розрізняються. У випадку застосування МН для обчислення границь термів (рис. 4) класи добре

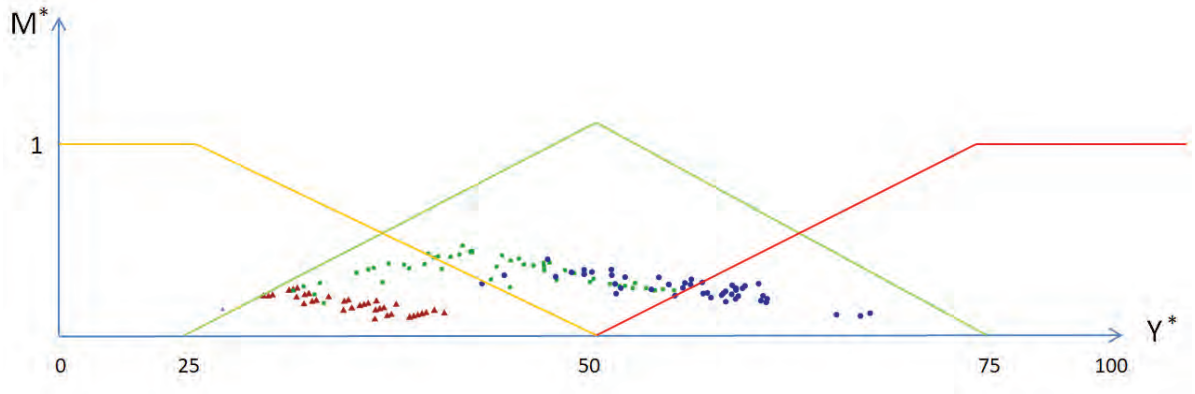

Рис. 3. Розподіл об'єктів між проміжними класами, обчислений з використанням АМ

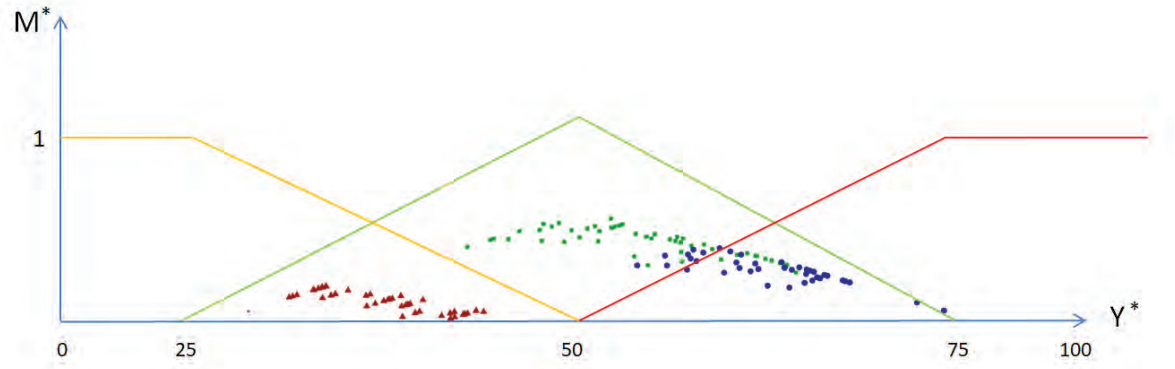

Рис. 4. Розподіл об'єктів між проміжними класами, обчислений з використанням МН

Таблиця 3

Лінгвістичні змінні	**Лінгвістичні значення**	**Границі термів (МН)**	**Границі термів (АМ)**	**Границі термів (МН) при додаванні аномальних об'єктів**	**Границі термів (АМ) при додаванні аномальних об'єктів**
sepalLength	low	[4.3, 5.75]	[4.3, 5.85]	[0.3, 5.75]	[0.3, 5.7]
	mid	[5.025, 6.025]	[5.07, 6.88]	[3.025, 6.825]	[3.01, 6.80]
	high	[5.75, 7.9]	[5.85, 7.9]	[5.75, 7.9]	[5.7, 7.9]
sepalWidth	low	[2.0, 3.099]	[2.0, 3.06]	[0.1, 3.099]	[0.1, 2.98]
	mid	[2.55, 3.75]	[2.53, 3.73]	[1.59, 3.75]	[1.54, 3.69]
	high	[3.099, 4.4]	[3.06, 4.4]	[3.099, 4.4]	[2.98, 4.4]
petalLength	low	[1.0, 3.35]	[1.0, 3.76]	[0.1, 3.30]	[0.1, 6.97]
	mid	[2.175, 5.125]	[2.38, 5.33]	[1.7, 256.7]	[3.54, 258.54]
	high	[3.35, 6.9]	[3.76, 6.9]	[3.30, 510.1]	[6.97, 510.1]
petalWidth	low	[0.1, 1,05]	[0.1, 1,2]	[0.1, 1.05]	[0.1, 1.19]
	mid	[0.575, 1.775]	[0.65, 1.85]	[0.575, 1.775]	[0.65, 1.85]
	high	[1,05, 2.5]	[1,2, 2.5]	[1.05, 2.5]	[1.19, 2.5]

Таблиця 4

	Метод визначення границь термів	accuracy	precision	recall	f1-score
Звичайна вибірка	МН	0,87	0,87	0,83	0,84
	АМ	0,86	0,85	0,82	0,83
Вибірка з аномальними об'єктами	МН	0,84	0,84	0,81	0,82
	АМ	0,40	0,38	0,35	0,36

Таблиця 5

Розмір навчальної вибірки (навчальна/ тестова)	accuracy	precision	recall	f1-score
130/20	0,87	0,87	0,83	0,84
120/30	0,87	0,87	0,83	0,84
100/50	0,83	0,83	0,80	0,81
75/75	0,77	0,77	0,75	0,75

розрізняються, незначні перетинання присутні для об'єктів Versicolor (позначено квадратами), Virginica (позначено колом).

У таблиці 4 наведені значення метрик якості результатів класифікації з використанням різних способів обчислення границь термів. З аналізу таблиці видно, що використання міжквартильного середнього для визначення границь термів дозволяє отримати прийнятну точність класифікації для вибірок, які містять об'єкти з аномальними характеристиками.

У роботі досліджувалося питання впливу способу розбиття вибірки на навчальну і тестову, а також розміру навчальної вибірки на якість класифікації. Результати порівняння наведені в табл. 5.

З аналізу результатів випливає, що розмір навчальної вибірки впливає на результат класифікації, зі зменшенням обсягу навчальної вибірки якість класифікації погіршується.

Висновки. У роботі для розв'язання задачі класифікації запропоновано використовувати нечітку модель представлення знань, побудовану на базі навчальної вибірки, та систему нечіткого логічного виводу. Генерація системи нечітких логічних правил відбувається автоматично, налаштування моделі здійснюється в процесі навчання з використанням навчальної вибірки. Для зменшення впливу наявності аномальних об'єктів в навчальній вибірці на результат класифікації для визначення границь термів використовується міжквартильне середнє. Побудовано двохетапну процедури класифікації з використанням допоміжних нечітких класів об'єктів, які потім відображаються в чіткі вихідні класи. Наведено порівняння результатів класифікації з використанням запропонованого в роботі двохетапного походу і класичного алгоритму нечіткої класифікації. Виконано числовий аналіз впливу параметрів моделі та розмірів вибірки на якість класифікації.

Запропонований підхід дозволяє мінімізувати участь експерта під час формування системи правил та налаштування моделі нечіткої класифікації.

Надалі планується застосування запропонованого підходу до класифікації наборів даних, що містять об'єкти зі значною кількістю ознак та важко класифікуються з використанням класичних підходів.

ЛІТЕРАТУРА

1. Штовба С.Д. Классификация объектов на основе нечеткого логического вывода. *Exponenta Pro – Математика в приложених*. 2004. № 1. С. 68–69.
2. Романов В.Н. Применение нечетких моделей в задачах классификации. *Альманах современной науки и образования. Тамбов: Грамота*. 2014. № 5-6(84). С. 108-112. ISSN 1993-5552.
3. Кучеренко Е.И., Глушенкова И.С., Глушенков С.А. Нечеткое разбиение объектов на основе критериев плотности. *Радіоелектроніка, інформатика, управління*. 2016. № 1(36). Зр. 32-39. doi:10.15588/1607-3274-2016-1-4.
4. Пташко Е.А., Ухоботов В.И. Автоматическая генерация нечетких правил для управления мобильным роботом с гусеничным шасси на основе числовых даннях. *Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и інформатика*. 2017. Vol. 6, №. 3. Pp. 60–72.
5. Edward R. Tufte. The Visual Display of Quantitative Information. Second Edition. Graphics Press, Box 430, Cheshire, Connecticut, 2007. pp. 191.
6. Терлецкий Д.А., Провотар А.И. Нечеткие объектно-ориентированные динамические сети. *Кибернетика и системный анализ.* Киев, 2015. том 51. № 1. С. 40–47.

7. Ротштейн А.П. Интеллектуальные технологии идентификации: нечеткая логика, генетические алгоритмы, нейронные сети. URL : http://matlab.exponenta.ru/fuzzylogic/book5/index.php.
8. Гук Н.А., Єгошкін Д.І., Сірик С.Ф. Алгоритм класифікації на базі нечіткої логіки з розширюваною кількістю виводів. *Питання прикладної математики і математичного моделювання* : Зб. наук. пр. Дніпро. 2018. Вип. 18. С. 67–76.
9. Richard O. Duda, Peter E. Hart, David G. Stork Pattern Classification, 2nd Edition. Wiley-Interscience, 2001. 688 p.

REFERENCES

1. Shtovba S.D. (2004) Klassifikatsiya ob"yektov na osnove nechetkogo logicheskogo vyvoda. *Exponenta Pro – Matematika v prilozheniyakh*. no. 1, pp. 68-69.
2. Romanov V.N. (2014) Primeneniye nechetkikh modeley v zadachakh klassifikatsii. *Al'manakh sovremennoy nauki i obrazovaniya. Tambov: Gramota.* No. 5-6 (84), pp. 108-112. ISSN 1993-5552.
3. Kucherenko Ye.I., Glushenkova I.S., Glushenkov S.A. (2016) Nechetkoye razbiyeniye ob"yektov na osnove kriteriyev plotnosti. *Radíoyelektroníka, ínformatika, upravlínnya*. No. 1 (36), pp. 32-39. doi:10.15588/1607-3274-2016-1-4
4. Ptashko Ye.A., Ukhobotov V.I. (2017) Avtomaticheskaya generatsiya nechetkikh pravil dlya upravleniya mobil'nym robotom s gusenichnym shassi na osnove chislovykh dannykh. *Vestnik Yuzhno-Ural'skogo gosudarstvennogo universiteta. Seriya: Vychislitel'naya matematika i informatika*. Vol. 6, no. 3, pp. 60-72.
5. Tufte E.R. (2007) *The Visual Display of Quantitative Information. Second Edition*. Graphics Press, Box 430, Cheshire, Connecticut.
6. Terletskiy D.A., Provotar A.I. (2015) Nechetkiye ob"yektno-oriyentirovannyye dinamicheskiye seti. *Kibernetika i sistemnyy analiz*. Vol. 51, no. 1, pp. 40-47.
7. Rotshteyn A.P. "Intellektual'nyye tekhnologii identifikatsii: nechetkaya logika, geneticheskiye algoritmy, neyronnyye seti". URL: http://matlab.exponenta.ru/fuzzylogic/book5/index.php
8. Huk N.A., Yehoshkin D.I., Siryk S.F. (2018) Alhorytm klasyfikatsiyi na bazi nechitkoyi lohiky z rozshyryuvanoyu kil′kistyu vyvodiv. *Pytannya prykladnoyi matematyky i matematychnoho modelyuvannya*. Vol. 18, pp. 67-76.
9. Duda R.O., Hart P.E., Stork D.G. (2001) *Pattern Classification, 2nd Edition*. Wiley-Interscience.