УДК 004.89:369

ІНТЕЛЕКТУАЛЬНА СИСТЕМА КЛАСИФІКАЦІЙНОГО ПРОГНОЗУВАННЯ УСПІШНОСТІ СОЦІАЛЬНИХ ПРОЕКТІВ

О. В. Федоришин, В. Д. Карпуша, О. М. Теліженко,
Сумський державний університет,
вул. Римського-Корсакова, 2, м. Суми, 40007, Україна

У статті викладено науково-методичний підхід до оцінювання ефективності реалізації, успішності впровадження, доцільності реорганізації, актуальності злиття соціально-орієнтованих кампаній, стартапів, проектів, програм тощо. Реалізовано інтелектуальну систему класифікації стартапів за рівнем успішності. При цьому розглянуто основні аспекти формування вхідного математичного опису системи, особливості її функціонування в режимі навчання та екзамену, а також основні критерії оцінки ефективності інтелектуальної системи в інформаційному розумінні. Виконано класифікаційне прогнозування успішності страхових стартапів на основі застосування парадигми штучних нейронних мереж за алгоритмом зворотного поширення помилки. Для підвищення достовірності класифікації оптимізовано структуру нейронної мережі.

***Ключові слова**: класифікаційне прогнозування, успішність стартапу, страхова компанія, вхідний математичний опис, штучні нейронні мережі, алгоритм зворотного поширення помилки, багатошаровий персептрон.*

ВСТУП

Сучасні інформаційні системи аналізу даних страхової компанії містять в своєму складі інтелектуальні елементи, що дозволяють оцінювати потенційні можливості клієнтів щодо їх участі в різних програмах страхування та надання нових страхових продуктів. Використання інтелектуальних систем дозволяють класифікувати клієнтів за певними характеристиками, наприклад, ступенем їх довіри до страхової компанії, визначати їх зацікавленість в збільшенні обсягу страхових послуг. Інтелектуальний аналіз даних (Data Mining) та класифікаційне прогнозування є поширеними інструментами для спектра інтелектуальних інформаційних технологій, що використовуються при цьому, містить як класичні технології експертних знань, так і сучасні розробки в сфері нечітких множин, кластер-аналізу та штучних нейронних мереж. Можливість формувати бази знань та використовувати класифіковані дані дозволяє ще на етапі планування оцінити ступінь успішності проекту та своєчасно внести зміни в його структуру.

В статті запропоновано підхід до створення інтелектуальної системи прогнозування успішних стартапів на прикладі інтелектуального аналізу даних страхової компанії з елементами штучного інтелекту для оцінки активності клієнтів компанії щодо участі в спеціальній програмі страхування «Mobile Home Policy».

ОГЛЯД ІСНУЮЧИХ МЕТОДІВ. ПОСТАНОВКА ЗАДАЧІ

Дослідження складних явищ та процесів неможливо без використання сучасних інструментів аналізу. Використання ідей машинного навчання та теорії розпізнавання образів [1] дозволяє реалізувати інтелектуальні системи [2], здатні аналізувати дані подібно людині [3]. Сучасний стан

рівня автоматизації українських підприємств знаходиться на перехідному етапі[4]. Вважається, що необхідність автоматизації того або іншого напряму діяльності залежить від розвитку того сектору, де застосовуються прикладні методи аналізу та управління. Data Mining [5], або ж knowledge discovery є одним з підходів, що використовується для початкового аналізу даних. Використання даного підходу дозволяє вирізняти інформацію від даних, виявляти закономірності, схожість у розподілі, тощо. Data Mining є одним з різновидів низькорівневого аналізу даних універсальної природи, що робить його незамінним для різноманітних прикладних задач [6].

Багато алгоритмів автоматичної класифікації та їх функціоналів якості розглянуті в [7]. Ці функціонали й алгоритми характеризуються різною трудомісткістю і потребують ресурсів високопродуктивних комп'ютерів. Різноманітні процедури класифікації входять до складу практично всіх сучасних пакетів прикладних програм для статистичного опрацювання багатовимірних даних. Класифікаційні процедури ієрархічного типу призначені для наочного уявлення про стратифікаційну структуру всієї досліджуваної сукупності об'єктів. Ці процедури засновані на послідовному об'єднанні кластерів і на послідовній розбивці. Найбільше поширення одержали процедури послідовного об'єднання кластерів, котрий називають висхідним, індуктивним, композиційним або ж дендрологічним.

Нечіткі методи мають цілий ряд переваг перед класичними методами аналізу: можливість нечіткого кластер–аналізу (віднесення одного об’єкта відразу до кількох кластерів), здатність самостійно вибирати структуру даних (динамічно змінювати функцію приналежності для кожного кластеру окремо) та інші.

Найбільш перспективним методом прогнозування є використання нейронних мереж. Нейронні мережі мають більш гнучку структуру. Для зміни структури у рамках визначеної архітектури нейронної мережі достатньо регулювати кількість шарів та нейронів, додаткові переваги надає можливість зміни активаційної функції. Лише ці незначні перетворення надають можливість повністю змінити структуру мережі, що дозволить максимально пристосувати обрану архітектуру, яка розв’язується і в свою чергу дозволить мінімізувати похибку навчання мережі (підвищити точність прогнозування). Ще одна серйозна перевага нейронних мереж полягає в тому, що побудова нейромережевої моделі відбувається адаптивно під час навчання, без участі експерта.

Метою роботи є розробка інформаційного та програмного забезпечення інтелектуальної системи прогностичної класифікації активності клієнтів страхової компанії щодо участі в спеціальній програмі страхування «Mobile Home Policy» на основі аналізу їх соціодемографічних параметрів та поточних страхових зобов’язань.

Основними завданнями роботи є:

– формування вхідного математичного опису інтелектуальної системи;

– вибір типу та структури штучної нейронної мережі, що здатна навчатися «з учителем»;

– розробка та реалізація алгоритмів оптимізації функціональних параметрів нейронної мережі;

– Оцінка впливу на достовірність класифікацію такого параметру навчання, як передатна функція персептронів ШНМ;

– перевірка працездатності розробленого блоку інтелектуальної системи на задачі визначення потенційних учасників спеціальної програми страхування «Mobile Home Policy».

АЛГОРИТМ НАВЧАННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ

Для реалізації інтелектуальної системи прогнозування обрано нейромережеву парадигму, а саме багатошаровий персептрон Розенблатта [8]. Нейромережа з даною топологією використовується для вирішення ряду прикладних задач типу Data Mining [9], крім того, можливе використання оптимально конфігурованої нейромережі для оцінювання станів об'єкта, що досліджується. Для підвищення достовірності розпізнавання доцільно використати алгоритм зворотного поширення помилки [10]. Новизною застосування ШНМ даної топології є оптимізація передатної функції [11] - функціоналу, що на основі значень вхідних синапсів(ваг) нейрона, формує його вихідний сигнал. Для багатошарового персептрона вихідний сигнал внутрішнього прошарку нейронів є вхідними даними для нейронів зовнішнього рівня.

Синтез ШНМ, крім задання її структури, також полягає в навчанні. Навчання являє собою конфігурування нейронів шляхом почергової модифікації ваги синапсів. Період коригування всіх нейронів нейромережі називають епохою або ж циклом навчання. Перед початком навчання, перш за все, необхідно вагу синапсів ініціалізувати початковими значеннями ваг. При введенні запам'ятованого стимулу (результат класифікації) з'являється реакція синапсів зовнішнього шару - формується вихідний сигнал, значення якого тлумачать як результат класифікації. Коли утворена хвиля досягає зовнішніх нейронів, знаходять величину помилки – різниці між отриманим та бажаним значенням реакції нейромережі [12]. Помилка класифікації є основним аргументом оцінки достовірності класифікації даних за допомогою нейромережі.

Даний алгоритм використовує певну зовнішню ланку, що надає ШНМ, крім вхідних, і цільові вихідні образи, які створюються на попередньому етапі для кожного вхідного образу експертами. Тому такі алгоритми називаються алгоритмами навчання з учителем [13].

Оцінка реакції прихованих нейронів здійснюється обчисленням зваженого значення помилки класифікації, знайденої для шару нейронів. За вагову функцію використовують поточні значення синапсів нейронів зовнішнього шару, котрі йдуть від прихованих нейронів до ефекторів - вихідного шару. Помилка поширюється в зворотному напрямку, коректуючи ваги за вище приведеним алгоритмом. Якщо прихованих шарів декілька – перелік помилок наводять для кожного, починаючи з шару ефекторів.

Цільовою функцією помилки нейромережі, що мінімізується, є величина:

$$E(\bar{w}) = \frac{1}{2}\sum_{j}\sum_{p}\left(y_j^{N}(\bar{p}) - d_{jp}\right)^2, \qquad (1)$$

де $y_j^{N}(\bar{p})$ - реальний вихідний стан нейрона j вихідного шару N ШНМ при подачі на її входи p-ого образу; d_{jp} - ідеальний (цільовий) стан цього нейрона.

Підсумок ведеться за нейронами вихідного шару і за всіма оброблюваними ШНМ образами. Мінімізація здійснюється за методом градієнтного спуску, тобто ваги модифікуються таким чином:

$$\Delta w_{ij}^{(n)} = -\eta \cdot \frac{\partial E}{\partial w_{ij}}, \tag{2}$$

де w_{ij} - ваговий коефіцієнт синапсу, що з'єднує і-тий нейрон шару n-1 з j-тим нейроном шару n, η – коефіцієнт швидкості навчання (0<η<1). Похідну з (2) можна подати у вигляді

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j} \cdot \frac{dy_j}{ds_j} \cdot \frac{\partial s_j}{\partial w_{ij}}, \tag{3}$$

де y_j - вихідний сигнал нейрона j, s_j - зважена сума його вхідних сигналів (аргумент активаційної функції).

Очевидно, $\partial s_j / \partial w_{ij} = y_i^{(n-1)}$.

dy_j / ds_j вказує на те, що похідна активаційної функції за її аргументом повинна бути визначена по всій осі абсцис. (Наприклад, якщо за активаційну функцію взяти гіперболічний тангенс, то $dy_j / ds_j = 1\text{-}s^2$).

$$\frac{\partial E}{\partial y_j} = \sum_k \frac{\partial E}{\partial y_k} \cdot \frac{dy_k}{ds_k} \cdot \frac{\partial s_k}{\partial y_j} = \sum_k \frac{\partial E}{\partial y_k} \cdot \frac{dy_k}{ds_k} \cdot w_{jk}^{(n+1)}. \tag{4}$$

Зазначимо, що сума ваг виконується серед нейронів шару (n+1).

Позначимо

$$\delta_j^{(n)} = \frac{\partial E}{\partial y_j} \cdot \frac{dy_j}{ds_j}, \tag{5}$$

тоді

$$\delta_j^{(n)} = \left[\sum_k \delta_k^{(n)} \cdot w_{jk}^{(n+1)} \right] \cdot \frac{dy_j}{ds_j}. \tag{6}$$

Для вихідного шару розраховуємо:

$$\delta_j^{(N)} = y_j^{(N)} \cdot -d_j \cdot \frac{dy_j}{ds_j}. \tag{7}$$

Тоді (2) набуває вигляду:

$$\Delta w_{ij}^{(n)} = -\eta \cdot \delta_j^{(n)} y_i^{(n-1)}. \tag{8}$$

Алгоритм навчання багатошарового персептрона за методом зворотного поширення помилки наступний:

– На вхід нейромережі подається навчальна вибірка-сукупність вхідних даних (реалізація) та результат класифікації(належність до певного класу).

– Розрахунок $\delta^{(N)}$ для вихідного шару нейромережі за (7).

– Розрахунок приросту ваги Δw для вихідного шару за (6) та (8).

– Розрахунок $\delta^{(n)}$ і Δw за (6) та (8) для решти шарів n=N-1..1.

– Корекція синаптичних ваг $w_{ij}(k) = w_{ij}(k-1) + \Delta w_{ij}(k)$.

– Оцінка помилки, перейти на перший пункт у випадку істотного значення помилки.

На першому кроці у нейромережу подаються вся навчальна вибірка у вигляді реалізацій та результату класифікації.

РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ

Як приклад застосування інтелектуальної системи розглянуто прогнозування активності клієнтів компанії щодо участі в спеціальній програмі страхування «Mobile Home Policy». Формування вхідного математичного опису для інтелектуальної системи виконувалося за даними розміщеними на репозитарії даних для машинного навчання (Machine Learning Repository) Центру машинного навчання та інтелектуальних систем університету Каліфорнії.

В роботі застосовуються дані розділу Insurance Company Benchmark (COIL 2000) Data Set. Навчальна вибірка у вигляді матриць типу «об'єкт-властивість» формується для двох класів клієнтів компанії: таких, що приймають учать в спеціальних програмах страхування, та таких, що відмовляються від участі. Словник ознак складається з 85 характеристик клієнтів, наприклад, їх вік, посада, сімейний стан, наявність інших страхових полісів тощо. Навчальна вибірка складалася з 4000 реалізацій, з яких 238 відповідали позитивно на пропозицію щодо участі в спеціальній програмі страхування «Mobile Home Policy».

Програмна реалізація системи виконувалася в середовищі для інженерних та наукових розрахунків MATLAB з використанням спеціалізованого пакету для формування, навчання та тестування штучних нейронних мереж Neuro Net Toolbox.

На рис. 1 наведено графік динаміки зміни значення середньоквадратичної помилки в процесі навчання нейромережі з такими параметрами та передатною функцією нейронів прихованого та вихідного прошарку у вигляді логістичної функції:

$$f(s) = \frac{1}{1+e^{-s}}, \qquad f(s)\,log\,sig(s) = \frac{1}{1+e^{-s_0}}, \tag{9}$$

де s – результат роботи суматора штучного нейрону.

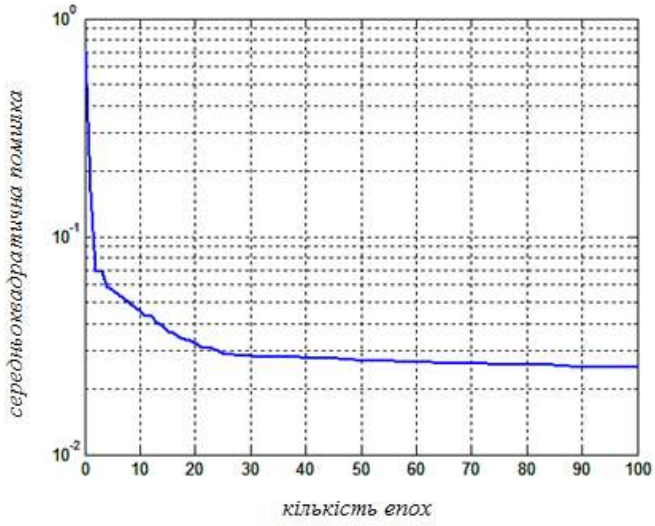

Рисунок 1 – Динаміка зміни значення середньоквадратичної помилки при навчанні ШНМ з логістичною передатною функцією

Аналіз рисунку показує, що нейронній мережі не вдалося побудувати безпомилковий класифікатор, оскільки значення помилки більше нуля (E = 0,0249859).

Ймовірність правильного визначення клієнтів, реакція яких на рекламну кампанію позитивна складає D_1 = 99,73%, а ймовірність правильної класифікації тих, хто не прийме участь D_2 = 59,66%. З метою підвищення цих значень був запропонований m-сценарій, який змінював такий важливий параметр як тип передатної функції. Змінимо передатну функцію нейронів прихованого та вихідного прошарків на лінійну

$$f(s) = purelin(s) = s, \qquad f(s) = s, \tag{10}$$

де s – результат роботи суматора штучного нейрону.

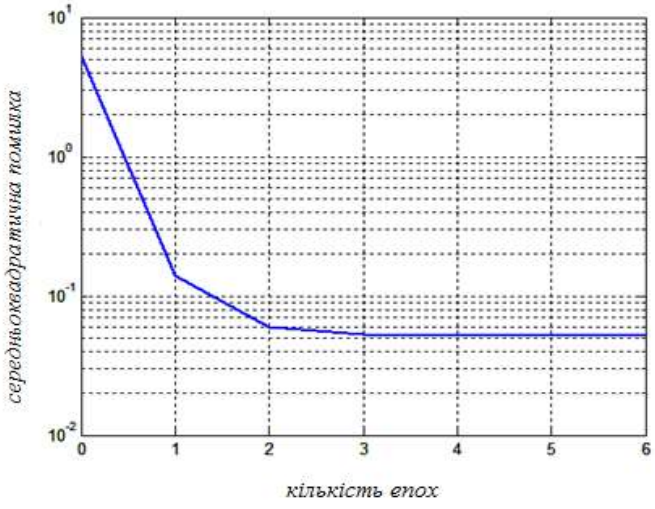

Рисунок 2 - Динаміка зміни значення середньоквадратичної помилки при навчанні ШНМ з лінійною передатною функцією

Аналіз рис. 2 показує, що значення середньоквадратичної помилки складає $E = 0{,}052218$. При цьому точність розпізнавання реалізацій першого класу $D_1 = 99{,}21\%$, а для другого класу $D_2 = 0{,}97\%$. Це доводить те, що класи клієнтів є лінійно-нероздільними. Отже, такий вид передатної функції для даної системи застосовувати не можна.

Найбільш ефективним було застосування передатної функції у вигляді гіперболічного тангенсу

$$f(s) = tansig(s) = \frac{2}{1 + e^{-2s}} - 1, \qquad f(s) = \frac{2}{1 + e^{-2s}} - 1, \tag{11}$$

де s – результат роботи суматора штучного нейрону.

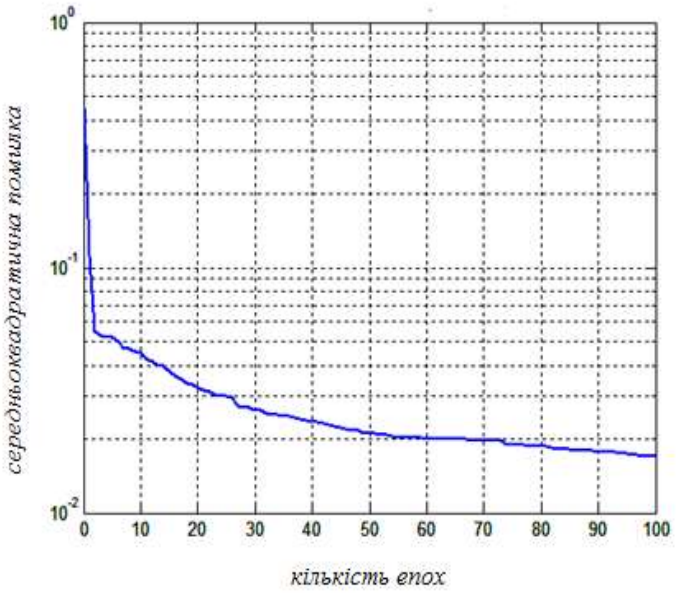

Рисунок 3 – Динаміка зміни значення середньоквадратичної помилки при навчанні ШНМ з передатною функцією у вигляді гіперболічного тангенсу

В даному випадку, достовірність її рішень щодо першого класу клієнтів складає $D_1 = 98{,}37\ \%$, а другого класу – $D_2 = 75{,}14\ \%$. Значення середньоквадратичної помилки складає $E = 0{,}0169204$.

Аналогічним чином виконувалося перенавчання штучної нейромережі для різних варіантів передатних функцій прихованого та вихідного прошарку.

Результати перенавчання наведені в таблиці 1.

Аналіз табл. 1 показує, що оптимальною, в інформаційному розумінні, є нейромережа, що використовує передатні функції у вигляді гіперболічного тангенсу.

Таблиця 1 – Вибір передатних функцій нейромережі

		Передатні функції вихідного прошарку					
		Гіперболічний тангенс		Лінійна		Логістична	
Передатні функції прихованого прошарку	Гіперболічний тангенс	E=0,0169		E=0,0464		E=0,0465	
		D_1=98,37%	D_2=75,14%.	D_1=99,10%	D_2=20,17%.	D_1=99,95%	D_2=21,01%.
	Лінійна	E=0,0522		E=0,0522		E=0,0462	
		D_1=*99,21*%	D_2=*0.97*%.	D_1=*99,21*%	D_2=*0,97*%.	D_1=99,84%	D_2=23,95%.
	Логістична	E= 0,0179		E=0,0306		E=0,0250	
		D_1=99,89%	D_2=73,11%.	D_1=99,87%	D_2=42,02%	D_1=99,73%	D_2=59,66%.

ВИСНОВКИ

В роботі запропоновано підхід до оцінювання успішності стартапів та їх класифікаційного прогнозування. Побудовано в рамках нейромережної парадигми інтелектуальну систему оцінювання ефективності проектів та стартапів. Досліджено вплив на достовірність класифікації такого параметра навчання, як тип передатної функції, при цьому. Проведено оптимізацію типу передатної функції нейронів різних прошарків нейромережі. При цьому достовірність правильної класифікації складає D_1*=98,37 %* для першого класу та D_2*=75,14 %* для другого класу відповідно. Значення середньоквадратичної помилки складає *E=0,0169.* Практичну цінність отриманих результатів можна отримати впровадженням інтелектуальної системи в існуючі комплекси для класифікації та аналізу даних, застосувати для оцінки соціальних, економічних, політичних, програм та стартапів, оцінювати вихід на ринок нових продуктів та якість ребрендингу існуючих, тощо. Оскільки достовірність класифікації не є безпомилковою, для підвищення якості класифікації необхідно оптимізувати структуру нейромережі шляхом збільшення кількості вхідних нейронів (ознак розпізнавання) та їх контрастування, збільшення кількості прихованих прошаків нейронів, оптимізувати передатну функцію окремо для кожного нейрона, а не для цілого шару, розробити гібридний алгоритм навчання штучних нейронних мереж з використанням підходів та принципів інших технологій машинного навчання та розпізнавання образів.

SUMMARY

SUCCESSFULNESS FORECASTING INTELLIGENCE SYSTEM FOR SOCIAL PROJECTS CLASSIFICATION

O. Fedoryshyn, V. Karpusha, O. Telizhenko,
Sumy State University,
2, Rimskogo-Korsakova St., 40007, Sumy, Ukraine

The article presents the scientific and methodical approach to performance evaluation of, implementation, successful application, reorganization feasibility, merging relevance of social oriented companies, startups, projects, programs, etc. The intelligence system of startup classification according to the level of success is implemented. Herewith the basic aspects of

formation of the input mathematical representation of a system, specifics of its functioning in the mode of study and examination, and basic criteria of intelligent system performance evaluation in the information concept are considered. The forecasting classification of the successfulness of the insurance startups was carried out on the basis of the paradigm of artificial neural networks' application under the back propagation of error algorithm. To improve the reliability of the classification the structure of neural network is optimized.

Keywords: *classification forecasting, startup successfulness, the insurance company, input mathematical representation, artificial neural network, back propagation of error algorithm, multilayer perceptron.*

СПИСОК ЛІТЕРАТУРИ

1. Bishop C. M. Neural Networks and Pattern Recognition. – Oxford Press. 1995.
2. Башмаков А. И. Интеллектуальные информационные технологии: учеб. пособие / А. И. Башмаков, И. А. Башмаков. — М. : Изд-во МГТУ им. Н. Э. Баумана, 2005. — 304 с
3. Калініна І. О. Дослідження нейромережевих методів у задачах прогнозування / І. О. Калініна // Наукові праці. – К., 2009. – Вип.93, Т.106.
4. Береза А. М. Інформаційні системи і технології в економіці: навчально-методичний посібник для самостійного вивчення. – 2002. – 278 с.
5. Дюк В. Data Mining: учебный курс / В. Дюк, А. Самойленко. – СПб. :Питер, 2001. – 386 с.
6. Ian H. Witten . Hall Data Mining: Practical Machine Learning Tools and Techniques. — 3rd Edition / Ian H. Witten, Eibe Frank and Mark A. — Morgan Kaufmann, 2011. — P. 664.
7. Методы нейроинформатики: отв. за выпуск М.Г. Доррер. –Красноярск: КГТУ, 1998. – 205 с
8. Соколов Е. Н. Нейроинтеллект: от нейрона к нейрокомпьютеру / Е. Н. Соколов, Г. Г. Вайтнявичус. – М. : Наука, 1989. – С. 283.
9. Уоссермен Ф. Нейрокомпьютерная техника / Ф . Уоссермен. – М. : Мир,1992.
10. Kohonen T. Self-organization and Associative Memory / T. Kohonen. – Berlin: Springer-Verlag, 1989.
11. Minsky M. L, Papert S. 1969. Perseptrons. Cambridge, MA: MIT Press. (Русский перевод: Минский М. Л., Пейперт С. Перцептроны. – М. : Мир. – 1971.)
12. Fausett L. V. Fundamentals of Neural Networks: Architectures, Algorithms and Applications / L. V. Fausett. – Prentice Hall, 1994.
13. Hecht-Nielsen, Robert. Counter-Propagation Networks // IEEE First International Conference on Neural Networks. – 1987. –Volume II.

Надійшла до редакції 21 травня 2014 р.