УДК 004.832.32

КРАВЕЦЬ І.О., ПИЩИТА О.О.

ДОСЛІДЖЕННЯ ЗАСТОСУВАННЯ НЕЙРОННИХ МЕРЕЖ ДЛЯ ЗАДАЧ DATA MINING

Проведено дослідження ефективності алгоритмів Data Mining, які використовують нейронні мережі. Розроблено та реалізовано для аналізу соціально-економічної інформації нейронні мережі та алгоритми, їх навчання для задач класифікації, кластерного аналізу, прогнозування даних. Запропоновано алгоритм кластеризації з визначенням як центрів кластерів, так і числа кластерів з використанням мережі Кохонена.

The efficiency of Data Mining algorithms, which are applied neuron' networks are researched. The neuron' networks and its learning algorithms for classification, the cluster analyses, and time series forecasting are developed.

Вступ

На даний час ведуться інтенсивні роботи по застосуванню нейронних мереж для вирішення задач пошуку прихованих знань в різних проблемних областях: кліматології, медицині, психології і політології, діагностиці і оптимальному управлінні в технічних системах і т.д. Нейронні мережі дозволяють вирішувати різні неформалізовані задачі (задачі, де алгоритм рішення невідомий). Дослідник при цьому одержує дуже ефективну модель проблемної області і може дуже просто моделювати різні ситуації, пред'являючи мережі різні дані і оцінювати відповідь, яку надає мережа. Нейронні мережі можуть застосовуватися і в ситуаціях, коли відомий порівнюваний по точності прогнозу метод рішення, але критичний, наприклад, час отримання результату, оскільки навчена нейронна мережа вирішує задачу, що пред'являється їй, дуже швидко [7, 9].

На українському ринку технології інтелектуальних обчислень роблять лише перші кроки. Це можна пояснити їх високою вартістю, але, як показує історія розвитку інших галузей комп'ютерного ринку України, сам по собі цей фактор навряд чи є визначальним. Скоріше тут виявляється дія деяких специфічних для України негативних факторів, що різко зменшують ефективність застосування аналітичних технологій [2]. З одного боку, існують статистичні пакети SPSS, Statistica, Matlab, в яких використані деякі алгоритми інформаційного аналізу даних, яле вони не працюють в режимі on-line і вимагають від користувача введення інформації перед кожним аналізом даних, а також знання методів статистичного та інтелектуального аналізу даних. З іншого боку існують комерційні програмні продукти, такі як SoMine, NeuroShell, NeuroScalp, Deductor (додаток до Olap), які використовують інтелектуальний аналіз даних в режимі on-line, але вони дорого коштують. Крім того, задачі інтелектуальної обробки даних мають ряд специфічних ознак (великий обсяк інформації, наявність інформації у якісному та кількісному вигляді, необхідність

автоматичного генерування гіпотез прихованих закономірностей) і потребують розробки специфічного алгоритмічного та програмного забезпечення.

Постановка задачі дослідження

Об'єктом дослідження у даної роботи є вибір або розробка структури нейронних мереж та алгоритмів роботи нейронних мереж для розв'язання задач інтелектуальної обробки даних. У даній роботі розглянуто розв'язання таких основних задач як: кластеризація, класифікація та прогнозування стосовно обробки соціально-економічної інформації.

Розв'язання задачі кластеризації за допомогою нейронних мереж.

Для розв'язання задач кластеризації найбільш доцільно використовувати мережі Кохонена. Мережа розпізнає кластери в навчальних даних і розподіляє дані до відповідних кластерів. Якщо в наступному мережа зустрічається з набором даних, несхожих ні з одним із відомих зразків, вона відносить його до нового кластеру. Якщо в даних містяться мітки класів, то мережа спроможна вирішувати задачі класифікації. Мережі Кохонена можна використовувати і в задачах, де класи відомі – перевага буде у спроможності мережі виявляти подібність між різноманітними класами[2]. Мережа Кохонена має лише два прошарки: вхідний і вихідний, який називають самоорганізованою картою. Елементи карти розташовуються в деякому просторі – як правило, двовимірному. Вхідні сигнали – вектори дійсних чисел – послідовно пред'являються мережі. Бажані вихідні сигнали не визначаються. Після пред'явлення достатнього числа вхідних векторів, синаптичні ваги мережі визначають кластери. Крім того, ваги організуються так, що топологічно близькі вузли чуттєві до схожих вхідних сигналів, що дозволяє користувачу візуалізувати дані, що неможливо зрозуміти іншим способом.

Запропоновано та розроблено алгоритм кластерізації, який використовує мережу Кохонена, знаходить центри кластерів (ваги вихідних нейронів) та отримує оптимальне число кластерів (вихідних нейронів).

Алгоритм складається із таких кроків:

1. Ініціалізація мережі. Число вихідних нейронів дорівнює 2. Ваговим коефіцієнтам мережі надаються малі випадкові значення.
2. Пред'явлення мережі нового вхідного сигналу.
3. Обчислення відстані до всіх нейронів мережі. Відстані d_j від вхідного сигналу до кожного нейрона j визначаються за формулою:

$$d_j = \sum_{i=1}^{N} (x_i(t) - w_{ij}(t))^2,$$

де $x_i(t)$ – і-ий елемент вхідного сигналу в момент часу t, $w_{ij}(t)$ – вага зв'язку від і-го елемента вхідного сигналу до нейрона j у момент часу t.

4. Вибирається нейрон-переможець j*, для якого відстань d_j найменша.
5. Налаштування ваги нейрона j* і його сусідів: робиться налаштування ваг для нейрона j* і всіх нейронів з його околу. Нові значення ваг:

$$w_{ij}(t+1) = w_{ij}(t) + r(t) * \left(x_i(t) - w_{ij}(t)\right)$$

де r(t) – швидкість навчання, що зменшується з часом (додатне число, менше одиниці).

6. Якщо навчальна вибірка не закінчена, то повернення до кроку 2. Якщо навчальна вибірка закінчена, то переходимо до наступного кроку
7. Обчислення функції мети $J(M,W) = \sum_{i=1}^{c} \sum_{j=1}^{d} d_A^2(m_j, w^{(i)})$
8. Порівняння функції мети із функцією мети попередньої епохи . Якщо функція мети зменшується, то додається ще один вихідний нейрон переходимо до пункту 2.

Якщо функція мети не зменшується, то виводимо результати. Блок-схема алгоритму представлена на рис. 1.

В алгоритмі використовується коефіцієнт швидкості навчання, який поступово зменшується. В результаті позиція центру встановлюється в певній позиції, яка задовільним чином кластеризує приклади, для яких даний нейрон є переможцем Для порівняння роботи алгоритмів кластеризації за допомогою мережі Кохонена була взята кластеризація пакетом SPSS як еталон, а також реалізований оптимізаційний алгоритм кластерізації. Дослідження показали, що результати роботи (розподіл кластерів) збігаються на 80% для мережі Кохонена та пакету SPSS і на 70% при використанні оптимізаційного алгоритму кластеризації та пакету SPSS.

Результати порівняння роботи алгоритмів кластеризації та пакету SPSS наведені на рис. 2.

Кластерний аналіз за допомогою мереж Кохонена дозволяє розглядати достатньо великий об'єм інформації і різко скорочувати, стискати великі масиви соціально-економічної інформації, робити їх компактними і наочними.

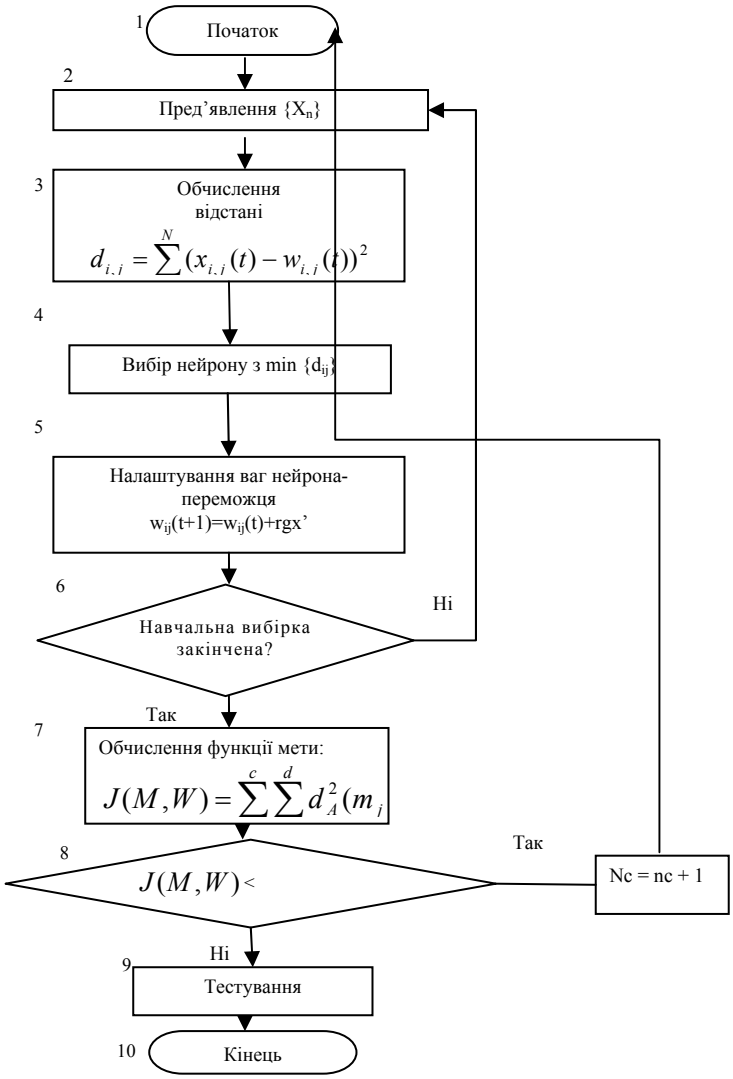

Рис. 1. Блок-схема алгоритму кластеризації з використанням мережі Кохонена

	Кл 1	Кл 2	Кл 3	Кл 4	Кл 5
Вінницьки					1
Волинськ					1
Дніпропе			1		
Донецьки					1
Житомир	1				
Закарпат				1	
Запорізьк				1	
Івано-Фр	1				
Київськи	1				
Кіровогра	1				
Луганськ	1				
Львівськ	1				
Миколаїв	1				
Одеський	1				
Полтавськ	1				
Рівненськ				1	
Сумський	1				
Тернопіль		1			
Харківськ	1				
Херсонськ	1				
Хмельниц	1				
Черкаськ	1				
Чернівець	1				
Чернігівс	1				

Мережа Кохонена

	Кл 1	Кл 2	Кл 3	Кл 4	Кл 5
Вінницьки					1
Волинськ					1
Дніпропе			1		
Донецьки	1				
Житомир	1				
Закарпат				1	
Запорізьк	1				
Івано-Фр		1			
Київськи	1				
Кіровогра	1				
Луганськ	1				
Львівськ	1				
Миколаїв	1				
Одеський	1				
Полтавськ	1				
Рівненськ	1				
Сумський	1				
Тернопіль	1				
Харківськ	1				
Херсонськ					1
Хмельниц	1				
Черкаськ	1				
Чернівець	1				
Чернігівс	1				

SPSS

	Кл 1	Кл 2	Кл 3	Кл 4	Кл 5
Вінницьки					1
Волинськ					1
Дніпропе			1		
Донецьки				1	
Житомир				1	
Закарпат				1	
Запорізьк				1	
Івано-Фр		1			
Київськи			1		
Кіровогра				1	
Луганськ	1				
Львівськ	1				
Миколаїв	1				
Одеський	1				
Полтавськ	1				
Рівненськ				1	
Сумський				1	
Тернопіль				1	
Харківськ	1				
Херсонськ					1
Хмельниц		1			
Черкаськ		1			
Чернівець					1
Чернігівс		1			

Оптимізаційний алгоритм

Рис. 2. Порівняння результатів кластерізації

Розв’язання задачі класифікації за допомогою нейронних мереж

Для розробки алгоритму класифікації було проаналізовано роботу нейронних мереж зі зворотним поширенням похибки та векторного квантування. Обидві нейронні мережі є мережі, які навчаються з вчителем. Для алгоритмів з використанням мережі зворотного поширення було обрарано нейронну мережу з 3 прошарків: вхідного прошарку, у якого число нейронів дорівнює числу ознак об’єкту, прихованого прошарку з 3 нейронів, та вихідного прошарку з 2 нейронів. Алгоритм працює наступним чином:

1. Ініціалізація мережі: вагові коефіцієнти і зсуви мережі приймають малі випадкові значення.
2. Визначення елемента навчальної множини: (вхід – вихід). Входи $\{x_1, x_2 ... x_N\}$, повинні розрізнятися для всіх прикладів навчальної множини.
3. Прямий хід. Обчислення сигналів прихованного прошарку та сигналів вихідних нейронів:

$$S_{i_m} = \sum_{i_{m-1}}^{N_{m-1}} w_{i_m j_{m-1}} y_{i_{m-1}} - b_{i_m} \qquad y_{im} = f(S_{im}),$$

$$i_m = 1, 2, ..., N_m,\ m = 1, 2, ..., L$$

де S – вихід суматора, w – вага зв’язку, y – вихід нейрона, b – зсув, i – номер нейрона, N – число нейронів у прошарку, m – номер прошарку, L – число прошарків, f – передатна функція.

4. Визначення похибки $E(W) = 0.5 * \sum_{i=1}^{N_L} (d_i - y_i)^2$

5. Якщо $E(w(t)) < \overline{\varepsilon^2_{зад}}$, то переходимо до наступного елемента навчальної виборки (пункт 2). Якщо ні, то виконується зворотний хід

6. Обчислюються частинні похідні похибки по вагам та зміщенню $\frac{\partial E}{\partial W}, \frac{\partial E}{\partial \Theta}$. Налаштовуються синоптичні ваги: $w_{ij}(t+1) = w_{ij}(t) + \gamma\, g_j * x'_J$

де w_{ij} – вага від нейрона i або від елемента вхідного сигналу i до нейрона j у момент часу t, x_i' – вихід нейрона j, r – швидкість навчання, g_j – значення похибки для нейрона j.

Якщо нейрон з номером j належить останньому прошарку, тоді

$$g_j = y_j * (1 - y_j) * (d_J - y_j)$$

де d_j – бажаний вихід нейрона j, y_j – поточний вихід нейрона j.

Якщо нейрон з номером j належить одному з прошарків з першого по передостанній, тоді $g_j = x'_j * (1 - x'_j) * \sum_K g_K * w_{jK}$

де k пробігає всі нейрони прошарку з номером на одиницю менш, ніж у того, котрому належить нейрон j.

7. Переходимо до пункту 3.

Але цей алгоритм виявив залежність кількості ітерації від початкових значень та зупинення у точках локальних мінімумах похибки. Тому був застосований алгоритм, який навчається швидше, класифікації з використанням мережі Кохонена та квантуванням навчального вектора. Під час дослідження нейронних мереж для задач класифікації було проаналізовано роботу алгоритмів нейронних мереж з зворотним поширенням похибки та мережі Кохонена з векторним квантуванням. Обидві нейронні мережі є мережі, які навчаються з вчителем. Навчання здійснюється шляхом послідовного пред'явлення вхідних векторів з одночасним налаштуванням ваг відповідно до певної процедури. У процесі навчання ваги мережі поступово стають такими, щоб кожний вхідний вектор виробляв вихідний вектор, що зменшує час на вирішення задачі.

Прогнозування часових рядів з використанням мережі зворотного поширення похибки та плаваючих вікон

Розроблений алгоритм прогнозування використанням мережі зворотного поширення похибки та плаваючих вікон з врахуванням зв'язків характеристик, що прогнозуються.

Алгоритм навчання мережі подібний до алгоритму нейронних мереж з зворотним поширенням похибки, описаному вище. Тільки вхідним вектором є значення прогнозованого показника в попередніх часових кроках. Якщо необхідно урахувати кореляцію прогнозованого показника з іншими показниками, то вхідним вектором є значення прогнозованого показника в попередніх часових кроках та значення інших показників в попередніх часових кроках.

Якщо потрібно прогнозувати показники далі, то значення вхідного вектора складаються зі значень показників зсунутих у часі вперед, що утворює ефект плаваючого вікна у часі.

Задачі прогнозування з використанням мережі зворотного поширення похибки та плаваючих вікон, який враховує зв'язки прогнозуємих характеристик, дає хороший прогноз, який становить до 10% похибки відповідно до реальних даних.

Збіжність характеристик, які прогнозуються із реальними становить 90%.

Висновки

Було проведено дослідження використання нейронних мереж для задач інтелектуальної обробки даних (Data Mining) і порівняння з іншими алгоритмами Data Mining, які не використовують нейронні мережі. В результаті досліджень було виявлено, що

- для задач класифікації більш ефективним є алгоритм класифікації з використанням мережі Кохонена та квантуванням навчального вектора, ніж алгоритм з використанням мережі зворотного поширення похибки, який виявив залежність кількості ітерації від початкових значень ваг нейронів та зупинявся у точках локальних минимумів похибки.

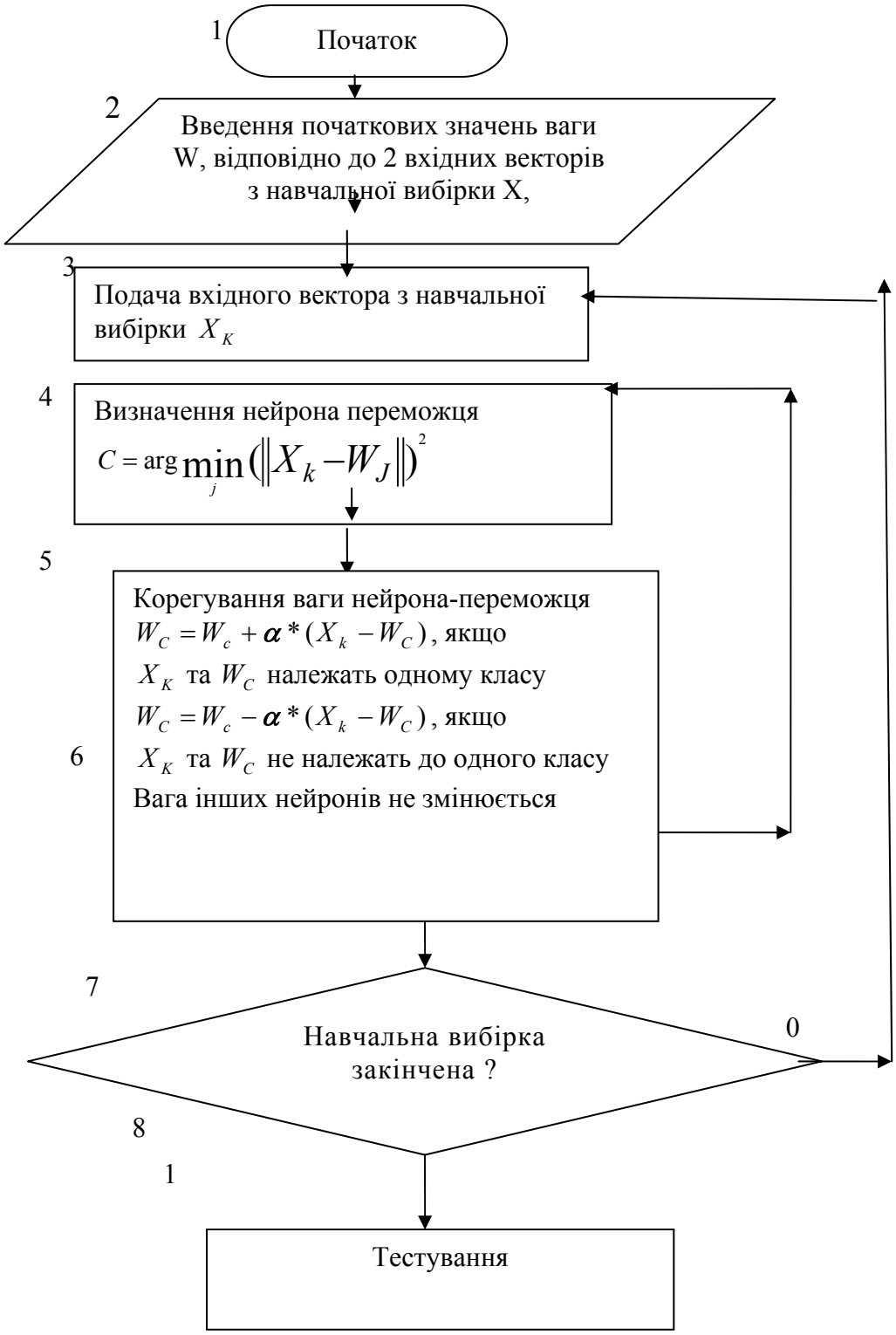

Рис. 3. Алгоритм класифікації з використанням мережі Кохонена та квантуванням навчального вектора

Рис. 4. Прогнозування ВВП (%) з використанням мережі зворотного поширення похибки та плаваючих вікон на 2007 рік, у порівнянні з реальними даними (для прогнозу використані дані за 2006 р.)

- для задач кластеризації більш доцільним є алгоритм кластеризації з використанням мережі Кохонена.
- Було запропоновано власну модифікацію алгоритму кластеризації з використанням мережі Кохонена та оптимізацією числа вихідних нейронів або кластерів;
- для задач прогнозування найбільш ефективним є алгоритм прогнозування з використанням мережі зворотного поширення похибки та плаваючих вікон, який враховує зв'язки прогнозуємих характеристик. Збіжність характеристик, які прогнозуються із реальними становить 90%.

Створена інформаційно-аналітична система, яка поєднує базу даних та інформаційний блок, може слугувати демонстраційним прикладом використання методів інтелектуального аналізу даних з використанням нейронних мереж.

Для розробки бази даних даного програмного забезпечення було обрано середовище MS Access, що забезпечує швидкий доступ і обробку великих масивів даних, не потребує спеціальної установки, бо є додатком MS Office. При розробці програмного забезпечення було використано середовище Borland Delphi 7, що відноситься до так званих систем швидкої розробки прикладних програм, і являє собою потужний генератор коду, візуальний дизайнер прикладних програм та засіб ведення баз даних з надзвичайним інтерфейсом.

ЛІТЕРАТУРА

1. Kohonen T. 1984. Self–organization and associative memory. Series in Information Sciences, vol. 8. Berlin: Springer verlag.
2. Rosenblatt R. 1959. Principles of neurodynamics. New York: Spartan Books.
3. Widrow B. 1959 Adaptive sampled–data systems, a statistical theory of adaptation. 1959. IRE WESCON Convention Record, part 4. New York: Institute of Radio Engineers.
4. Widrow B., Hoff M. 1960. Adaptive switching circuits. I960. IRE WESCON Convention Record. New York: Institute of Radio Engineers.
5. http://victoria.lviv.ua/html/oio/html/theme1.htm – Перспективні дослідження і розробки по інтелектуальних системах
6. Алексеев А.В., Круг П.Г., Петров О.М. Нейросетевые и нейрокомпьютерные технологии: Методические указания к проведению лабораторных работ по курсу "Информационные технологии". – М.: МГАПИ, 1999.

7. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.: ил.

8. Гаврилов А.В., Канглер В.М. Использование искусственных нейронных сетей для анализа данных. – СБОРНИК НАУЧНЫХ ТРУДОВ НГТУ. – 1999. – № 3(16). – 230 с.

9. Заенцев И.В. Нейронные сети: основные модели. – Воронеж, 1999. – 74 с.

10. Руденко О.Г., Бодянський Є.В. Штучні нейронні мережі: Навчальний посібник. – Харків: ТОВ "Компанія СМІТ", 2006. – 404 с.

11. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: ИНФРА-М. Финансы и статистика, 1995. – 384 с.

12. Уоссермен З.Ф. Нейрокомпьютерная техника: теория и практика. – М.: Мир – 1992. – 123 с.