

Стрельченко І.І.

ВИКОРИСТАННЯ НЕЙРОННИХ МЕРЕЖ В ЗАДАЧАХ ДИНАМІЧНОЇ КЛАСТЕРИЗАЦІЇ ЕКОНОМІЧНИХ СИСТЕМ

Державний вищий навчальний заклад «Київський національний економічний університет імені Вадима Гетьмана»

В роботі розглянуто сучасні підходи до класифікації економічних об'єктів на основі теорії нейронних мереж. Досліджено переваги та недоліки використання радіально-базисних мереж, мереж Кохонена, векторного квантування та зустрічного розповсюдження. Обґрунтовано використання мереж Кохонена в умовах невеликого обсягу навчальної вибірки та відсутності початкової системи класів для порівняння. Розроблено концептуальні положення та процедуру застосування коефіцієнта конкордації для вибору оптимальної топології нейронної мережі. Отримані теоретичні положення дістали практичної реалізації в системі MatLab. Розроблений програмний модуль протестований на статистичних даних Міжнародного валютного фонду, що характеризують протікання фінансової кризи в період з 2007–2009 рр.

Ключові слова: кластеризація, макроекономічний індикатор, фінансова криза, мережа Кохонена, коефіцієнт конкордації.

Вступ

В задачах класифікації багатовимірних даних ефективним є використання нейронних мереж. Технологія нейронних мереж дозволяє, зокрема, визначити приналежність точок у просторі розмірністю $n > 3$, що є надзвичайно важливим в задачах дослідження соціально-економічних, біологічних, інформаційних систем.

Для вирішення задач класифікації найбільшого розповсюдження на сьогодні дістали нейронні мережі з наступною архітектурою:

– радіально-базисні мережі (англ. термін Radial Basis Function Network, RBF-мережа); – у загальному випадку під радіально базисною нейронною мережею розуміється двохшарова мережа без зворотних зв'язків, яка містить прихований шар радіально симетричних нейронів. Радіальні базисні нейронні мережі складаються з більшої кількості нейронів, ніж стандартні мережі з прямою передачею сигналів і навчанням методом зворотного поширення помилки, але на їх створення потрібно значно менше часу. Ці мережі найбільш ефективні, коли є велика кількість навчальних векторів. [1];

– мережі Кохонена (англ. термін Kohonen Network) – це один з різновидів нейронних мереж, які використовують неконтрольоване навчання. При такому способі моделювання навчальна множина складається лише зі значень вхідних змінних, в процесі навчання немає по-

рівнювання виходів нейронів з еталонними значеннями. Можна сказати, що така мережа вчиться розуміти структуру даних. Всі подані на вхід сигнали заданої навчальної множини самоорганізована мережа в процесі навчання розділяє на класи, будуючи так звані топологічні карти [2];

– мережі векторного квантування (англ. термін LVQ – learning vector quantization) – засновані на навчальному векторному квантуванні та являть собою шар Кохонена, що навчається із вчителем. Побудова LVQ-мережі передбачає попереднє визначення кількості кластерів n , кількості класів m ($n \gg m$) та приналежність кожного кластера до певного класу [2];

– мережі зустрічного розповсюдження (Counterpropagation Network) – становлять собою мережі, які складаються з двох шарів: шару Кохонена та шару Гроссберга [3–4]. Мережі зустрічного розповсюдження поступаються за точністю багатшаровим персептронам, проте швидко навчаються та володіють низкою корисних властивостей [5].

Проте, коли існують об'єктивні обмеження на обсяг навчальної вибірки, а в структурі вихідних даних містяться дані лише описові характеристики об'єктів, що підлягають класифікації, та немає еталонної системи класів, до яких має бути віднесений кожен з них – від трьох з чотирьох наведених типів нейронних мереж досліднику доведеться відмовитись. Та узяти до

роботи нейронну мережу Кохонена.

Постановка проблеми

На основі статистичних даних Міжнародного валютного фонду провести кластеризацію економічних систем за рівнем подібності динаміки обраних макроекономічних індикаторів в часі. В якості інструментарію використати нейронну мережу типу Кохонена. Врахувати можливість вибору нейронної мережі з оптимальною топологією.

Метою даного дослідження є побудова нейронної мережі Кохонена з оптимальною топологією, що розбиватиме довільну вибірку з характеристик економічних систем за рівнем подібності зміни обраних макроекономічних індикаторів в часі.

Аналіз та дослідження публікацій

Штучні нейронні мережі є універсальним інструментарієм для ефективної побудови моделей практично будь-яких нелінійних структур [6]. Вони дозволяють вирішувати завдання розпізнавання образів, оптимізації, ідентифікації, управління динамічними об'єктами [7–8].

Мережі Кохонена – це один з різновидів нейронних мереж, що принципово відрізняються від інших, оскільки використовують неконтрольоване навчання. Під час такого навчання навчальна вибірка складається лише зі значень вхідних змінних, в процесі навчання немає порівняння виходів нейронів з еталонними значеннями. Можна сказати, що така мережа навчається розуміти структуру даних.

Ідея нейронної мережі Кохонена належить фінському вченому Тойво Кохонену [2]. Основний принцип роботи мереж – введення в правило навчання нейрона інформації щодо його розташування.

Науковці, які займаються дослідженням мереж Кохонена приходять до неоднозначних висновків щодо їх ефективності порівняно з іншими алгоритмами. Аббас [9] експериментально довів, що мережа Кохонена є більш ефективною порівняно з іншими алгоритмами кластеризації майже в усіх випадках. В інших роботах встановлено, що даний тип нейронних мереж менш схильний до локальних мінімумів, ніж алгоритм k-середніх [10]. Натомість існують дослідження, які демонструють, що застосування мереж Кохонена в задачах кластеризації не перевершує показники ієрархічного методу кластеризації та методу k-середніх та іноді виявляється менш ефективним. [11].

Проте в рамках даного дослідження мережі Кохонена володіють набором переваг, порівняно з іншими методами кластеризації, зокрема:

- доведено – ефективно працюють в умовах невеликого обсягу навчальної вибірки;
- реалізують концепцію «навчання без учи-

теля»;

- мають інтуїтивно зрозумілу та високоякісну графічну візуалізацію.

Викладення основного матеріалу

Кластеризація – це поділ досліджуваної безлічі об'єктів на групи «схожих» об'єктів, які називаються кластерами [12]. Завдання кластеризації принципово відрізняється від завдання класифікації. Рішенням задачі класифікації є віднесення кожного з об'єктів до одного із заданих визначених класів. У задачі кластеризації відбувається віднесення об'єкта до одного з невизначених класів. Розбиття об'єктів на кластери здійснюється при одночасному їх формуванні.

Змістовний аналіз отриманих кластерів дозволяє виявити нові закономірності в сукупності досліджуваних об'єктів, допомагає виявити аномалії. Відносячи новий об'єкт до одного з кластерів, можна прогнозувати поведінку об'єкта, оскільки його поведінка буде схожою з поведінкою інших об'єктів, що увійшли до кластера.

Формально задача кластеризації описується наступним чином [13]. Існує множина об'єктів $I = \{i_1, i_2, \dots, i_n\}$, кожен з яких характеризується вектором x_j , $j=1, 2, \dots, n$ параметрів: $x = \{x_{j1}, x_{j2}, \dots, x_{jm}\}$. Кластеризація передбачає побудову відображення початкової множини I на множину кластерів C :

$$I \rightarrow C, C \{c_1, c_2, \dots, c_k, \dots, c_g\}, \quad (1)$$

де c_k – кластер, що містить «подібні» об'єкти з множини I :

$$c_k = \{i_j, i_p \mid S \in I, I_p \in I \text{ та } d(i_j, i_p) < \sigma\}, \quad (2)$$

де σ – величина, яка визначає ступінь подібності для включення об'єктів до одного кластера; $d(i_j, i_p)$ – міра близькості між об'єктами або відстань між кластерами.

У випадку, коли відстань між двома об'єктами i_j та i_p – $d(i_j, i_p)$ менша за визначене значення σ , об'єкти вважаються близькими (схожими) і розміщуються в одному кластері. У протилежному випадку – об'єкти i_j та i_p є відмінними та розташовані у різних кластерах. Умову (2) називають гіпотезою компактності кластерів.

Для визначення відстані між кластерами існує декілька варіантів. В мережі Кохонена зазвичай використовують евклідову відстань. Евклідова відстань між векторами – це евклідова норма різниці векторів або довжина відрізка, яка поєднує точки i_j та i_p .

$$d(i_j, i_p) = \sqrt{\sum_{i=1}^m (i_j - i_p)^2} = \|i_j - i_p\|, \quad (3)$$

Один цикл навчання мережі Кохонена наведений на рис. 1 [14].

Процес рішення задачі кластеризації на основі мережі Кохонена є неоднозначним. По-перше, не існує однозначно найкращого критерію якості кластеризації. Відома ціла низка досить розумних критеріїв, а також низка алгоритмів, які не мають чітко вираженого критерію, але здійснюють досить розумну кластеризацію. Всі вони можуть давати різні результати. По-друге, число кластерів, як правило, невідомо заздалегідь і встановлюється відповідно до деякого суб'єктивного критерію. По-третє, результат кластеризації істотно залежить від початкової матриці ваг та може істотно відрізнитись для однієї навчальної вибірки при різних прогонах машинної реалізації алгоритму.

В межах даного дослідження запропоноване та експериментально протестоване можливе рішення перерахованих вище проблем.

Розглянемо наступну задачу. Існує вибірка з 65 країн, що характеризуються різним рівнем розвитку економіки. Під час проходження світової фінансової кризи (2007–2009 рр.) зафіксовані щоквартальні дані наступних макроекономічних індикаторів: валовий внутрішній продукт, зовнішні зобов'язання резидентів країни перед нерезидентами, валютний курс, обсяг золотовалютних резервів, вартість державних облігацій.

Якщо розглянути розмірність даних, то кожна точка характеризується п'ятьма координатами – значеннями макроекономічних індикаторів, кожен з яких у свою чергу має глибину в дванадцять часових значень. Загалом така система характеристик вкладається у шість вимірів: час і п'ять макроекономічних показників.

Потрібно розділити початкову сукупність на групи, в межах яких динаміка обраних індикаторів протягом зазначеного періоду буде подібною.

Для визначення оптимальної топології мережі Кохонена, що реалізує найкращу кластеризацію вихідних даних, розроблено та реалізовано в системі MatLab наступну покрокову процедуру:

1. Визначається максимально допустима кількість нейронів (N) у прихованому шарі за формулою (4)

$$N_{nv} > \frac{N}{\%e}, \quad (4)$$

де N_{nv} – розмір навчальної вибірки ($N_{nv}=65$); $\%e$ – відсоток помилки навчання ($\%e=0,1$).

2. Будується нейронна мережа Кохонена із заданою структурою ($N=1, \dots, 6$).

Для кожного отриманого кластеру обчислюють ранговий коефіцієнт конкордації (5–6):

$$W = \frac{12S}{m^2(n^3 - n)}, \quad (5)$$

де n – кількість об'єктів, що підлягають ранжуванню; m – кількість вибірок; а змінна S – знаходиться за формулою

$$S = \sum_{i=1}^n (R_i - \bar{R})^2, \quad (6)$$

де R_i – сума рангів i -го об'єкта ($R_i = \sum_{j=1}^m R_{ij}$); \bar{R} – середнє суми рангів за всіма об'єктами.

3. Визначається показник оптимальності топології поточної нейронної мережі за сумою всіх отриманих коефіцієнтів конкордації.

4. Виконується порівняння поточного показника оптимальності з отриманим під час попереднього прогону алгоритму. Більше значення показника вказує на нейронну мережу з оптимальною топологією.

Результатом моделювання за розробленою покроковою процедурою є нейронна мережа Кохонена з топологією оптимальною за обраним критерієм. Вона поділяє навчальну вибірку на шість кластерів за подібністю зміни п'яти макроекономічних індикаторів.

На рис. 2–3 наведені основні результати побудови нейронної мережі Кохонена з оптимальною топологією.

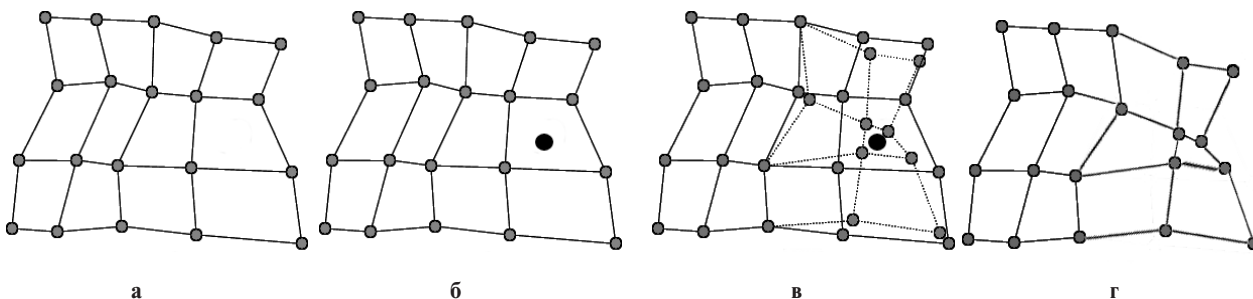


Рис. 1. Приклад навчання карти Кохонена для одного вхідного прикладу: а) вихідна нейронна мережа із вагами розподіленими випадковим чином; б) розташування навчального прикладу в координатах карти Кохонена; в) навчання нейронної мережі у відповідності із навчальним прикладом; г) навчена нейронна мережа після першого навчального прикладу

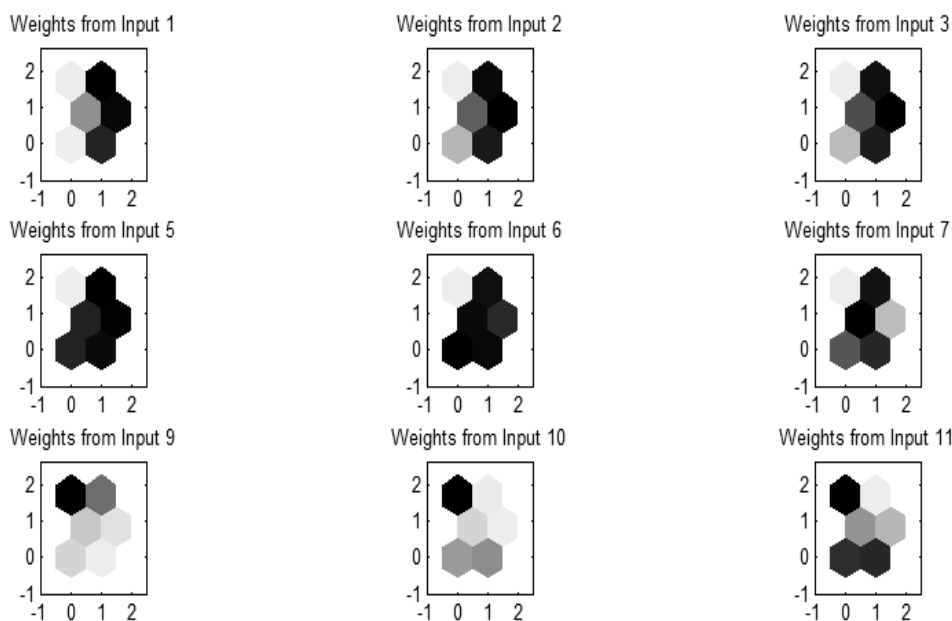


Рис. 2. Діаграми розподілу ваг нейронів мережі Кохонена за напрямком і силою зв'язків між входами

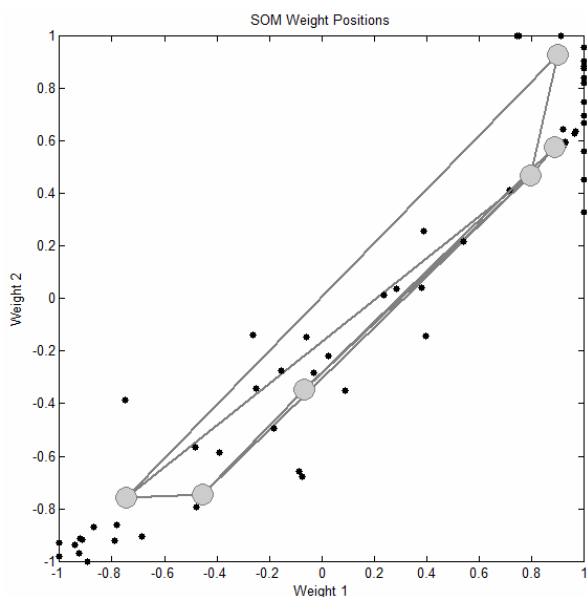


Рис. 3. Діаграма розташування нейронів карти Кохонена з оптимальною топологією, що відповідає кластерній структурі вихідної вибірки

Зокрема, рис. 3 дає двовимірну візуалізацію розташування нейронів навченої мережі та точок вихідних даних у цьому просторі.

Найбільша група країн була віднесена до кластеру номер 1. Найменший за розміром кластер включає чотири країни. Зазвичай до занадто малих кластерів потрапляють об'єкти, досліджувані характеристики яких мають аномальні значення. Таким чином, робимо висновок, що поведінка основних макроекономічних індикаторів країн даного кластеру демонструють нетипову поведінку під час проходження світової

фінансової кризи 2007–2009 рр.

Отримані кластери дозволяють виконати детальний аналіз особливостей поведінки економік різних країн під час світової фінансової кризи 2007–2009 рр.

Наприклад, порівняння діаграм входів кластерів номер 1 та 6, що розташовані далеко один від одного в топологічній структурі навченої мережі та мають протилежні кольори практично на кожному вході прихованого шару (рис. 2), дає підставу припустити, що динаміка досліджуваних макроекономічних індикаторів суттєво відрізнялась на більшості часових проміжків та мала протилежну направленість. На підтвердження такого висновку наведемо перелік країн, що потрапили до першого та шостого кластеру відповідно:

1. Євросона, Естонія, Литва, Чехія, Данія, Ізраїль, Сінгапур, Швейцарія, Сполучені Штати Америки, Бруней, Малайзія, М'янма, Філіппіни, Таїланд, Албанія, Боснія і Герцеговина, Болгарія, Хорватія, Латвія, Македонія, Чорногорія.

2. Камбоджа, Фіджі, Монголія, Шрі-Ланка, В'єтнам, Казахстан, Білорусь, Киргизька Республіка, Росія, Таджикистан, Україна.

Висновки

Результатом здійсненого дослідження є розроблені концептуальні положення побудови нейронної мережі Кохонена з оптимальною топологією для розв'язання задачі кластеризації економічних систем. В якості критерію оптимальності запропоновано використати систему рангових коефіцієнтів конкордації. Розроблена та протестована процедура знаходження нейрон-

ної мережі, що виконує найкращу кластеризацію навчальної вибірки за обраним критерієм. Написаний програмний модуль, що реалізує процес побудови нейронної мережі Кохонена з оптимальною топологією, та розбиває довільну вибірку з характеристик економічних систем за рівнем подібності зміни обраних макроекономічних індикаторів в часі.

Запропоновані концептуальні положення є універсальними та можуть бути використані в задачах динамічної кластеризації економічних, соціальних, медичних об'єктів, для машинного навчання.

СПИСОК ЛІТЕРАТУРИ

1. Дюк В., Самійленко А. Data mining: учебный курс. – СПб: Питер, 2001. – 368 с.
2. Kohonen T. Self-organizing maps. 3d ed. – New-York: Springer, 2001. – 501 p.
3. Уоссермен Ф. Нейрокомпьютерная техника: Теория и практика. – М.: Мир, 1992. – 240 с.
4. Бодянский Е.В., Руденко О.Г. Искусственные нейронные сети: архитектура, обучение, применение. – Харьков: ТЕЛЕТЕХ, 2004. – 372 с.
5. Grossberg S. Some networks that can learn, remember and reproduce any number of complicated space-time patterns, I // Journal of Mathematics and Mechanics. – 1969. – Vol.19. – No. 1. – P.53-91.
6. Нейрокомпьютеры и их применение: Нейронные сети: история развития теории. Кн. 5: Учеб. пособие для вузов / Под общей ред. А.И. Галушкина, Я.З. Цыпкина. – М.: ИПРЖР, 2001. – 840 с.
7. Уоссермен Ф. Нейрокомпьютерная техника. – М.: Мир, 1992. – 184 с.
8. Хайкин С. Нейронные сети: Полный курс. – М.: Издательский дом Вильямс, 2008. – 1103 с.
9. Ossama Abu Abbas. Comparisons Between of Data Clustering algorithms // The International Arab Journal of Information Technology. – 2008. – Vol.5. – No. 3.
10. Fernando Vazgo, Victor Lobo, Marco Painho. Self-organizing maps as substitutes for k-means clustering // Computational Science – ICCS. – 2005. – P.9-28.
11. Sueli A. Mingoti, Joab O. Lima Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms // European Journal of Operational Research 174. – 2006. – P.1742-1759.
12. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. – М.: ДМК Пресс, 2015. – 400 с.
13. Bishop C.M. Pattern Recognition and Machine Learning. – Springer, Series: Information Science and Statistics, 2006. – 740 p.
14. Стрельченко І.І., Савіна С.С., Доценко Н.В. Багатокритеріальні моделі в інтелектуальних системах прийняття рішень: навч. посібник. – К.: ДВНЗ «КНЕУ ім. В. Гетьмана», 2011. – 340 с.

Надійшла до редакції 10.08.17
Рецензент: д.е.н., проф. Дубницький В.І.

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧАХ ДИНАМИЧЕСКОЙ КЛАСТЕРИЗАЦИИ ЭКОНОМИЧЕСКИХ СИСТЕМ

Стрельченко И.И.

В работе рассмотрены современные подходы к классификации экономических объектов на основании теории нейронных сетей. Исследованы преимущества и недостатки использования радиально-базисных сетей, сетей Кохонена, векторного квантования и встречного распространения. Обосновано использование сетей Кохонена в условиях небольшого объема обучающей выборки и отсутствия исходной системы классов для сравнения. Разработаны концептуальные положения и процедура использования коэффициента конкордации для выбора оптимальной топологии нейронной сети. Полученные теоретические положения получили практическую реализацию в системе MatLab. Разработанный программный модуль был протестирован на статистических данных Международного валютного фонда, которые характеризуют протекание финансового кризиса в период с 2007 по 2009 гг.

Ключевые слова: кластеризация, макроекономический индикатор, финансовый кризис, сеть Кохонена, коэффициент конкордации.

THE USE OF NEURAL NETWORKS IN THE TASKS OF ECONOMIC SYSTEMS DYNAMIC CLUSTERING

Strelchenko I.I.

The paper considers modern approaches to the classification of economic objects based on the theory of neural networks. The advantages and disadvantages of using radial-basis networks, Kohonen networks, vector quantization and counter propagation are explored. It justified the use of Kohonen networks in a small volume of training sample and the lack of initial classes for the comparison. Conceptual provisions and the procedure for using the concordance coefficient for selecting the optimal neural network topology have been developed. The theoretical position gained practical implementation in MatLab system. The developed software module was tested on the statistical data of the International Monetary Fund, which characterize the course of the financial crisis in the period from 2007 to 2009.

Keywords: clusterization, macroeconomic indicator, financial crisis, Kohonen network, coefficient of concord.