РОЗДІЛ 1

ЗАГАЛЬНА ХАРАКТЕРИСТИКА ТА ВЛАСТИВОСТІ НЕЙРОННИХ МЕРЕЖ

Нейроінформатика – це розділ штучного інтелекту, який вивчає принципи подання та обробки інформації у штучних нейронних мережах.

Штучні нейронні мережі (НМ) – математичні моделі, а також їх програмні або апаратні реалізації, побудовані за принципами подання й обробки інформації у *біологічних нейронних мережах* – мережах нервових кліток живого організму.

1.1 Біологічний нейрон

Нервова система людини складається з клітин, які називаються *нейронами*, і має надзвичайну складність: близько 10^{11} нейронів беруть участь у близько 10^{15} передавальних *зв'язках*, що мають довжину метр і більше. Кожен нейрон має багато якостей, спільних з іншими клітинами, але його унікальною здатністю є прийом, обробка і передача електрохімічних сигналів по нервовим шляхам, що утворюють комунікаційну систему мозку.

На рис. 1.1 показана спрощена схема *біологічного нейрона*. *Дендрити* (деревоподібні відростки на входах нейрона) йдуть від *тіла нейрона* до інших нейронів, де вони приймають сигнали в *точках з'єднання*, які називаються *синапсами*.

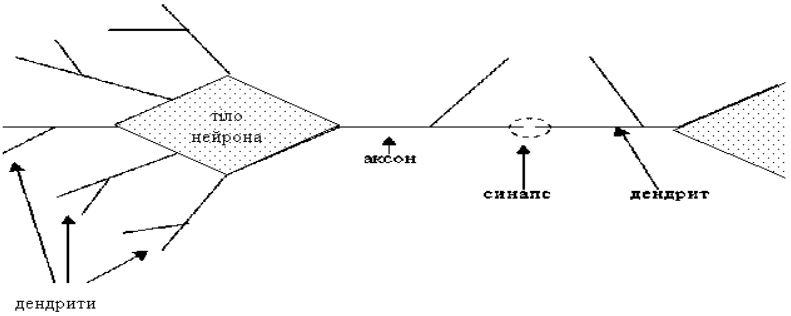


Рисунок 1.1 – Спрощена схема біологічного нейрона

Прийняті синапсом *вхідні сигнали* підводяться до тіла нейрона. Тут вони підсумовуються, причому одні входи прагнуть збудити нейрон, інші – перешкодити його збудженню. Коли сумарне *збудження* в тілі нейрона перевищує деякий *поріг*, нейрон збуджується, посилаючи по *аксону* (відросток на виході нейрона) сигнал іншим нейронам.

У цієї основної функціональної схеми багато ускладнень і виключень, проте більшість штучних НМ моделюють лише ці прості властивості.

1.2 Моделі штучних нейроелементів

Штучний нейрон (*формальний нейрон*, *нейроподібний елемент*) – це примітивний обчислювальний пристрій (або його модель), що має кілька входів і один вихід, і є основним обчислювальним елементом НМ.

Схему штучного нейрона зображено на рис. 1.2.

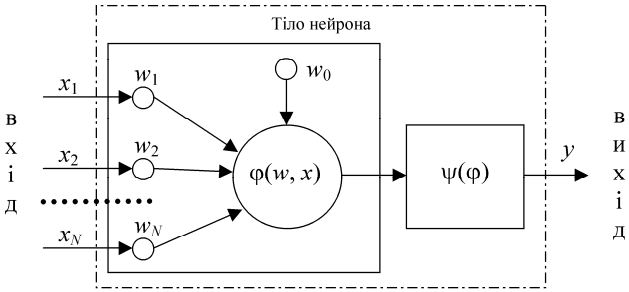


Рисунок 1.2 – Формальний нейрон

На вхід одношарового нейрона надходить *вхідний вектор* – набір вхідних сигналів $x=\{x_j\}$, $j = 1, 2, ..., N$, де N – кількість входів.

Кожний вхідний сигнал x_j зважується (масштабується певним чином) відносно зіставленої йому *ваги зв'язку* (*вагового коефіцієнта*) w_j, яка моделює перетворення сигналу у синапсі (міжнейронному контакті).

Дискримінантна (*вагова*, *постсинаптична*) *функція* нейрона φ поєднує зважені вхідні сигнали, отримуючи *постсинаптичний потенціал* та подає його значення до *функції активації*

(*передатної функції*) ψ, яка видає скалярне значення, що видається на *виході нейрона*. Таким чином, формальний нейрон реалізує скалярну функцію векторного аргументу, моделюючи перетворення вхідних сигналів у синапсах та тілі біологічного нейрона.

Отже, *математична модель функціонування штучного нейрона* описується співвідношенням: $y = \psi(\varphi(w, x))$, де x – вектор вхідних аргументів (сигналів); y – значення на виході нейрона; ψ – функція активації; φ – дискримінантна функція; $w=\{w_j\}$ – вектор, що містить значення вагових коефіцієнтів w_j і значення *зсуву* (*порогове значення*) w_0.

Набір вагових коефіцієнтів нейрона w моделює його *пам'ять*. Тому нейрони можна розглядати як запам'ятовуючі пристрої. У той же час нейрони можуть розглядатися як примітивні процесори, що здійснюють обчислення значення функції активації на основі значення дискримінантної функції вхідних сигналів і ваг.

Дискримінантні функції, як правило, використовують такі:

– *зважена сума*: $\varphi(w,x) = \sum_{j=1}^{N} w_j x_j + w_0$;

– *зважений добуток*: $\varphi(w,x) = \prod_{j=1}^{N} w_j x_j = w_0 \prod_{j=1}^{N} x_j$;

– *евклідова відстань*: $\varphi(w,x) = \sum_{j=1}^{N} (w_j - x_j)^2$.

Функція активації $\psi(x)$, де x – аргумент функції активації, повинна бути обмеженою (на інтервалі значень $x \in (-\infty, a]$ приймати значення y^0, на інтервалі $x \in [b,+\infty)$ – приймати значення y^1, на інтервалі $x \in (a, b)$ приймати значення y: $y^0 \le y \le y^1$, де y^0 та y^1 – деякі постійні мінімальне та максимальне значення, a та b – деякі константи, причому: $a \le b$) і монотонною (на інтервалі $x \in (a, b)$ $\Delta\psi(x) = \psi(x+\Delta x) - \psi(x)$) і не повинна змінювати знак при $\Delta x > 0$ та x, $x+\Delta x \in (a, b)$.

Як функція активації, як правило, застосовуються такі функції:

– *лінійна* приймає значення в діапазоні $(-\infty;+\infty)$:

$$\psi(x) = cx,$$

де c – константа, як правило, c =1;

– *лінійна біполярна з насиченням:*

$$\psi(x) = \begin{cases} 1, x > a_2; \\ Kx, a_1 \le x \le a_2; \\ -1, x < a_1, \end{cases}$$

де K, a_1, a_2 – константи;

– *лінійна уніполярна з насиченням:*

$$f(x) = \begin{cases} 1, x \ge \dfrac{1}{2a}; \\ ax + 0{,}5, |x| < \dfrac{1}{2a}; \\ 0, x \le -\dfrac{1}{2a}, \end{cases}$$

де a – константа.

– *порогова (Хевісайда)* приймає значення в діапазоні [0;1]:

$$\psi(x) = \begin{cases} 1, x \ge 0; \\ 0, x < 0; \end{cases}$$

– *порогова знакова (біполярна)* приймає значення в діапазоні [–1; 1]:

$$\psi(x) = \begin{cases} 1, x \ge 0, \\ -1, x < 0; \end{cases}$$

– *сигмоїдна логістична (уніполярна)* приймає значення в діапазоні (0;1):

$$\psi(x) = \frac{1}{1 + e^{-cx}},$$

де c – константа, як правило, $c = 1$.

– *гіперболічний тангенс* (*біполярна*):

$$\psi(x) = \tanh(Kx) = \frac{e^{Kx} - e^{-Kx}}{e^{Kx} + e^{-Kx}},$$

де K – константа.

– *косинусоїдальна з насиченням (уніполярна):*

$$\psi(x) = \begin{cases} 1, x \geq \dfrac{\pi}{2}; \\ \dfrac{1}{2}\left(1 + \cos\left(x - \dfrac{\pi}{2}\right)\right), \mid x \mid < \dfrac{\pi}{2}; \\ 0, x \leq -\dfrac{\pi}{2}; \end{cases}$$

– *синусоїдальна з насиченням* (*біполярна*):

$$\psi(x) = \begin{cases} 1, x \geq a; \\ \sin x, \mid x \mid < a; \\ -1, x \leq -a, \end{cases}$$

де a – константа.

– *радіально-базисна* (*Гауса*) приймає значення в діапазоні $(0; +\infty)$:

$$\psi(x) = e^{-cx^2},$$

де c – константа;

– *winner-take-all* (WTA – *переможець отримує усе*):

$$\psi^{(\mu,i)}(\{\varphi^{(\mu,j)}\}) = \begin{cases} 1, \varphi^{(\mu,i)} \leq \min\limits_{j}\{\varphi^{(\mu,j)}\}; \\ 0, \text{інакше.} \end{cases}$$

Незважаючи на те, що лінійні функції є найбільш простими, їх застосування обмежене, в основному, найпростішими НМ, що не містять прихованих шарів, в яких, крім того, існує лінійна залежність між вхідними і вихідними змінними. Такі мережі мають обмежені можливості. Багатошарова ж лінійна мережа може бути замінена еквівалентною одношаровою.

Проте використання лінійних активаційних функцій не позбавлене сенсу, у багатошарових НМ для розширення можливостей мережі застосовують нелінійні функції активації.

При побудові НМ часто доводиться працювати як з активаційною функцією, так і з її першою похідною. У цих випадках необхідним є використання як активаційної монотонної диференційованої та обмеженої функції. Особливо важливу роль відіграють такі функції при моделюванні нелінійних залежностей між вхідними і вихідними змінними. Це так звані *сигмоїдальні функції*. Функція $f(\bullet)$ називається сигмоїдальною, якщо вона є монотонно зростаючою, диференційованою і задовольняє умові:

$$\lim_{\lambda \to -\infty} f(\lambda) = k_1, \lim_{\lambda \to \infty} f(\lambda) = k_2, k_1 < k_2 .$$

Сигмоїдальну функцію за аналогією з електронними системами можна вважати нелінійною підсилювальною характеристикою штучного нейрона. Центральна область такої функції, що має великий коефіцієнт підсилення, вирішує проблему обробки слабких сигналів, а області з падаючим посиленням на позитивному і від'ємному кінцях слугують для обробки сильних збуджень. Таким чином, нейрон функціонує з великим посиленням у широкому діапазоні рівнів вхідного сигналу.

Функція *Softmax* (нормована експоненційна функція) – це узагальнення логістичної функції, що "стискує" K-вимірний вектор $\mathbf{z}$ із довільним значеннями компонент до K-вимірного вектора $\sigma(\mathbf{z})$ з дійсними значеннями компонент в області [0, 1] що в сумі дають одиницю:

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, j = 1,..., K.$$

Функція *ReLU* (Rectified Linear Units – ненасичуваний лінійний модуль) визначається формулою:

$$\psi(\varphi) = \max(0, \varphi).$$

1.3 Класифікація та види моделей нейромереж

У загальному випадку *нейронна мережа* являє собою сукупність нейронів, зв'язаних певним чином. Основними відмінностями нейромережевих моделей є способи зв'язку

нейронів між собою, функції нейроелементів, а також механізми та напрямки розповсюдження сигналів по мережі.

Конкретний вид виконуваного НМ перетворення інформації обумовлюється характеристиками нейроподібних елементів і особливостями архітектури мережі: топологією міжнейронних зв'язків, вибором певних підмножин нейроподібних елементів для введення і виведення інформації, способами навчання мережі, наявністю або відсутністю конкуренції між нейронами, напрямком і способами управління і синхронізації передачі інформації між нейронами.

Класифікація НМ:

– за *типом вхідної інформації* виділяють: *аналогові* (використовують інформацію у формі дійсних чисел) та *бінарні* (виконавчі) НМ (оперують з інформацією, описаною в двійковому вигляді);

– за *типом функції активації нейронів* виділяють мережі: *неперервні* (аналогові, дійсні, диференційовані – мережі, кожен елемент яких реалізує неперервно диференційовану функцію), *дискретні* (бінарні, порогові, недиференційовані – мережі, кожен елемент яких реалізує недиференційовану функцію), *дискретно-неперервні* (містять елементи з диференційованими і недиференційованими функціями);

– *за типом графа міжнейронних зв'язків* виділяють: *мережі без циклів* (*ациклічні мережі*); *мережі з циклами* (поділяються на *рівноважні мережі з циклами* і *мережі з обмеженими циклами*);

– *за типом структур нейронів* виділяють: *гомогенні* (*однорідні*) НМ (складаються з нейронів одного типу з єдиною функцією активації) та *гетерогенні* (*неоднорідні*) НМ (містять нейрони з різними функціями активації);

– *за кількістю шарів нейронів* виділяють: *одношарові* НМ (містять один шар нейронів) та *багатошарові* НМ (БНМ – містять більше одного шару взаємопов'язаних нейронів);

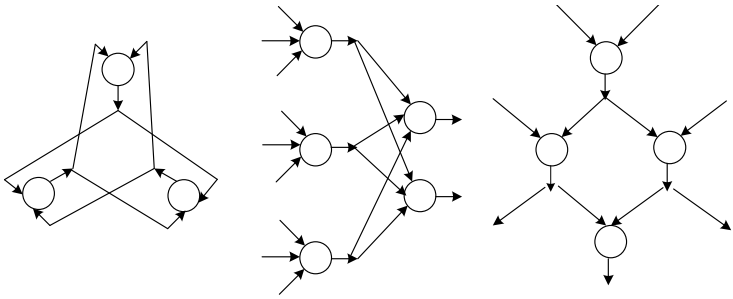
– *за типом дискримінантної функції* виділяють мережі; *зважені* (із вагами зв'язків) та *без ваг зв'язків*;

– *за принципом синтезу* виділяють: *мережі, що навчаються* (графи міжнейронних зв'язків та ваги входів змінюються при виконанні методу навчання) та мережі, що конструюються (кількість і

тип нейронів, граф міжнейронних зв'язків, ваги входів нейронів визначаються при створенні НМ, виходячи з розв'язуваної задачі);

– *за топологією зв'язків* виділяють (див. рис. 1.3): *повнозв'язні* (*інтерактивні*) мережі (усі нейрони пов'язані за принципом «кожний з кожним»: кожен нейрон передає свій вихідний сигнал іншим нейронам, включаючи самого себе, і вихідні сигнали мережі можуть бути всі або деякі вихідні сигнали нейронів після декількох тактів функціонування мережі; всі вхідні сигнали подаються всім нейронам), *неповнозв'язні* (*шаруваті*, *багатошарові*, *ієрархічні*) мережі (мають зазвичай шарувату організацію. У них різняться латеральні (бічні) зв'язки, які охоплюють нейрони одного шару, і проекційні (аферентні), що з'єднують шари нейронів. Частина нейронів має додаткові зовнішні входи, які утворюють рецепторне поле. Нейрони першого шару отримують вхідні сигнали, перетворять їх і через точки галуження передають нейронам наступного шару і так далі до останнього шару, який видає вихідні сигнали. У цьому випадку інформація рухається по висхідній від вхідного до вихідного прошарку і кожен наступний шар забезпечує як би більш високий рівень її обробки, ніж попередній. Число нейронів у кожному шарі може бути будь-яким і ніяк заздалегідь не пов'язано з кількістю нейронів в інших шарах) *ієрархічно інтерактивні* мережі (шари різного рівня пов'язані двосторонніми зв'язками і наявні зв'язки між елементами одного шару, але структура зв'язків не є однорідною, як в повністю інтерактивній мережі); *слабозв'язні* (*з локальними зв'язками*) мережі (нейрони розташовуються у вузлах прямокутної або гексагональної решітки, кожен нейрон зв'язаний з чотирма (окіл фон Неймана), шістьма (окіл Голео) або вісьмома (окіл Мура) своїми найближчими сусідами);

– *за характером зв'язків* виділяють: *мережі прямого поширення* (*мережі без зворотних зв'язків*, feedforward – сигнал по мережі проходить тільки в одному напрямку: від входу до виходу) та *рекурентні мережі* (*із зворотними зв'язками*, *зі зворотним поширенням інформації*, feedforward / feedback – характеризуються як прямим, так і зворотним поширенням інформації між шарами НМ). Серед рекурентних мереж, у свою чергу, виділяють мережі: *релаксаційні* (циркуляція інформації відбувається до тих пір, поки не перестануть змінюватися вихідні



а *б* *в*

Рисунок 1.3 – НМ з різною топологією зв'язків:

а – повнозв'язна мережа; *б* – шарувата мережа; *в* – слабозв'язана мережа

значення НМ – стан рівноваги) та *шаруваті мережі із зворотними зв'язками* (у них відсутній процес релаксації). Серед шаруватих мереж із зворотними зв'язками виділяють: *шарувато-циклічні мережі* (відрізняються тим, що шари замкнуті у кільце: останній шар передає свої вихідні сигнали першому; всі верстви рівноправні і можуть, як отримувати вхідні сигнали, так і видавати вихідні), *шарувато-повнозв'язні мережі* (складаються з шарів, кожен з яких представляє собою повнозв'язну мережу, а сигнали передаються як від шару до шару, так і всередині шару; в кожному шарі цикл роботи розпадається на три частини: прийом сигналів з попереднього шару, обмін сигналами всередині шару, вироблення вихідного сигналу і передача до подальшого шару), *повнозв'язно-шаруваті* (за своєю структурою аналогічні шарувато-повнозв'язним, але функціонуючи по-іншому: у них не розділяються фази обміну всередині шару і передачі наступного шару; на кожному такті нейрони всіх шарів приймають сигнали від нейронів як свого шару, так і наступних), *рециркуляційні мережі* (характеризуються прямим $y = f(x)$, і зворотним $x = f(y)$ перетворенням інформації; в рециркуляційних мережах навчання проводиться без вчителя, тобто вони самоорганізуються у процесі роботи) та *мережі з боковими зворотними зв'язками* (laterally connected);

– *за характером поділу зв'язків* виділяють: *немонотонні мережі* (не дозволяють визначити, як вплине зміна будь-якого

внутрішнього параметра мережі на вихідний сигнал) та *монотонні мережі* (кожен шар мережі крім останнього поділяється на два блоки: збудливий і гальмуючий. При цьому всі зв'язки в мережі влаштовані так, що елементи збудливої частині шару збуджують елементи збудливою частини наступного шару і гальмують елементи наступного шару. Аналогічно, які гальмуючі елементи збуджують гальмуючі елементи і гальмують збуджуючі елементи наступного шару (назви «гальмуючий» і «збуджуючий» відносяться до впливу елементів обох частин на вихідні елементи). Для нейронів монотонних мереж необхідна монотонна залежність вихідного сигналу нейрона від параметрів вхідних сигналів. Відзначимо, що для мереж з сигмоїдними елементами вимога монотонності означає, що ваги всіх зв'язків повинні бути не негативні);

– *за типом методу навчання* (характером налаштування синапсів) виділяють: *мережі з динамічними зв'язками і ітеративним навчанням, заснованому на принципі корекції помилок* (Базуються на запропонованій фізіологами моделі повторення шляхів нервового збудження. Основний метод навчання – зворотне поширення помилки широко використовується в сучасних нейрокомп'ютерів. Головною його перевагою є асимптотична збіжність процесу навчання, гарантуюча потенційну досяжність необхідної точності реакції НМ. Недолік методу полягає в необхідності багаторазового повторення ітерацій навчання на всьому обсязі запам'ятовуваних даних. Велика витрата часу й обчислювальних ресурсів при навчанні обмежували можливість застосування в адаптивних системах реального часу методів, заснованих на методі зворотного поширення), *мережі з фіксованими зв'язками і неітеративним навчанням* (Вони використовують методи прямого обчислення значення матриці зв'язків, які забезпечують запам'ятовування інформації без повторення, що прискорює процес навчання у порівнянні з ітеративними методами. Однак вони характеризується надмірною кількістю нейронів і низьким обсягом інформації, що запам'ятовується. Тому вони поки не отримали широкого поширення і застосовуються, в основному, в дослідницьких проектах);

– *за характером навчання* виділяють: *мережі з контрольованим* (*піднаглядним*, *з вчителем*, *з супервізором*) *навчанням* (коли відомим є вихідний простір рішень НМ; при навчанні порівнюють заздалегідь відомий вихід з отриманими значеннями) та *мережі з неконтрольованим* (*без нагляду*, *без вчителя*, *без супервізора*) *навчанням* (коли НМ формує вихідний простір рішень тільки на основі вхідних впливів, навчається, не знаючи заздалегідь правильних вихідних значень, але групуючи «близькі» вхідні вектори так, щоб вони формували один і той же вихід мережі; неспостережне навчання використовується, зокрема, при вирішенні задачі кластеризації) та *мережі зі змішаним навчанням* (коли частина ваг визначається при спостеріганні, а частина – при неспостережному навчанні; навчання здійснюється шляхом пред'явлення прикладів, що складаються з наборів вхідних даних у сукупності з відповідними результатами, при спостережному навчанні і без останніх при неспостережному);

– *за методом навчання* виділяють мережі: з неітеративним навчанням, з методом зворотного поширення помилки, з конкурентним навчанням, з використанням правила Хебба, з гібридним навчанням (в них застосовуються різні методи навчання) та інші;

– *за типом часу функціонування* виділяють мережі: *з неперервним часом* (аналогові мережі), *з дискретним асинхронним часом* (у кожен момент часу лише один нейрон змінює свій стан) та *з дискретним часом, що функціонують синхронно* (стан змінюється відразу у цілої групи нейронів);

– *у залежності від врахування попереднього стану мережі* виділяють мережі: *статичні* (не мають у своїй структурі ні зворотних зв'язків, ні динамічних елементів, а вихід залежить від заданої множини на вході і не залежить від попередніх станів мережі), *нечіткі* (поєднують в собі БНМ і нечітку систему, їх перший шар нейронів реалізує етап введення нечіткості (фаззіфікації), другий шар відображає сукупність нечітких правил, третій шар реалізує дефаззіфікацію – функцію приведення до чіткості), *нетрадиційні* НМ;

– *за типом розв'язуваних задач* виділяють мережі для: обробки та фільтрації даних, категоризації і таксономії даних, пошуку закономірностей у даних, заповнення прогалин у таблицях даних,

візуалізації та картографування даних, розпізнавання (класифікації) образів, непараметричної апроксимації залежностей за точковими даними, побудови баз даних великої ємності з швидким асоціативним пошуком інформації за неповними вхідними даними, розробки пристроїв асоціативної пам'яті, побудови експертних систем, яких навчають, на прикладах, здобуття знань з даних, адаптивного управління складними об'єктами і процесами, трудомістких задач оптимізації (типу задач про комівояжера і т. п.), шифрування, дешифрування, стиснення інформації, перекладу тексту;

– *за областю застосування* виділяють мережі для: розпізнавання сигналів, мови, зображень і тексту, технічної та біомедичної діагностики, моделювання залежностей в природничих науках і техніці, соціально-економічного прогнозування, управління в техніці та економіці, побудови інформаційно-пошукових систем, криптографії.

Для використання на практиці НМ реалізують у вигляді ***нейрокомп'ютера*** – обчислювальної системи, архітектура якої спеціалізована на виконанні операцій, адекватних структурі НМ.

Нейрокомп'ютери якісно відрізняються від усіх попередніх поколінь ЕОМ тим, що в них відсутні заздалегідь створені методичні програми і що вони, аналогічно людському мозку, здатні навчатися на окремих прикладах. У звичайних ЕОМ елементи схем з'єднані послідовно, кожен елемент з'єднаний тільки з двома-трьома елементами, так що сигнал обробляється поетапно, крок за кроком. Однак у НМ елементи мають множину паралельних з'єднань, причому кожен елемент з'єднаний майже з кожним. Через це вхідний сигнал поширюється по всій мережі, і всі елементи мережі працюють паралельно, реалізуючи, як кажуть, масовано-паралельні обчислення. Цим пояснюється можливість вирішувати складні обчислювальні задачі в реальному часі, справлятися з непередбаченими ситуаціями і навіть синтезувати знання з даних майже без участі людини.

Виділяють три *рівні нейрокомп'ютерів*:

– рівень 1 – імітаційна модель НМ на звичайній ЕОМ;

– рівень 2 – плата або приставка до персонального комп'ютера;

– рівень 3 – повнофункціональний нейрокомп'ютер на мікрочіпах.

1.4 Властивості штучних нейромереж

Переваги НМ і нейрокомп'ютерів полягають у тому, що вони дають стандартний спосіб рішення багатьох нестандартних завдань (неважливо, що спеціалізована машина краще вирішить один клас задач, важливіше, що один нейрокомп'ютер вирішить і цю задачу, і другу, і третю – і не треба кожен раз проектувати спеціалізовану ЕОМ – нейрокомп'ютер зробить все сам і майже не гірше); замість програмування використовується навчання (нейрокомп'ютер вчиться – потрібно лише формувати навчальні задачники, праця програміста, розпорядчого машині всі деталі роботи, заміщається працею вчителя, що створює «освітнє середовище», до якого пристосовується нейрокомп'ютер); нейрокомп'ютери особливо ефективні там, де необхідно подобу людської інтуїції і важко створити явний метод.

Основними ***властивостями нейромереж*** є:

– *однорідність* нейроелементів і базових операцій, а також технологічна простота різних способів їх фізичної реалізації: НМ за аналогією з мозком будуються з множини простих уніфікованих типових елементів (нейронів), що виконують елементарні дії (множення, додавання, обчислення найпростішої нелінійної функції) і з'єднаних між собою різними зв'язками;

– *можливість реалізації нелінійних відображень* шляхом використання нелінійних функцій активації нейронів (це важливо для вирішення завдань управління з істотними нелінійностями, для яких традиційні підходи поки не дають практично реалізованих рішень);

– *пластичність* – обумовлює складність поведінки НМ, яка розглядається як результат взаємодії багатьох елементів, кожен з яких обмежує дію інших і сам обмежується іншими на шляху до формування глобальної спостережуваною поведінки. Розрізняють *нейронну пластичність* (як пластичні елементи розглядаються нейрони), а також *синаптичну пластичність* (модифікація сили синаптичного зв'язку між нейронами). Оскільки кількість синапсів, як правило, на кілька порядків перевищує кількість нейронів, то мережу з пластичними синапсами буде володіти більшими можливостями, ніж мережа еквівалентного розміру з пластичними нейронами;