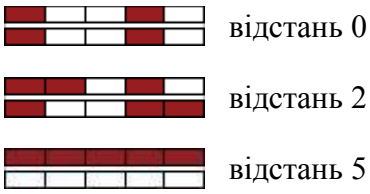
Застосування мереж Хемінга для нечіткого пошуку

Принцип роботи пам'яті в комп'ютері з архітектурою Фон-Неймана та людини принципово відрізняються один від одного. Комп'ютер використовує для пошуку інформації адресу, а людина асоціацію. Тому, якщо знати, де шукати інформацію, комп'ютер знайде її швидко, але якщо не відомо, то доведеться все перебирати, і добре, якщо дані є не спотвореними. Ймовірно, "якісніша" пам'ять людини дозволяє при набагато меншій обчислювальній потужності краще аналізувати.

Принципову обмеженість сучасних комп'ютерів можна обійти за допомогою різного роду систем асоціативної пам'яті, наприклад, мереж Хеммінга.

Принципи роботи мереж Хемінга

Алгоритм роботи базується на визначенні відстані Хемінга. Відстань Хемінга – це кількість позицій, що відрізняються, в бінарних векторах. Результатом роботи мережі є знаходження образу з найменшою відстанню.



Відсутність сигналу кодується як (-1), наявність (1). Мережа містить лише два прошарки.

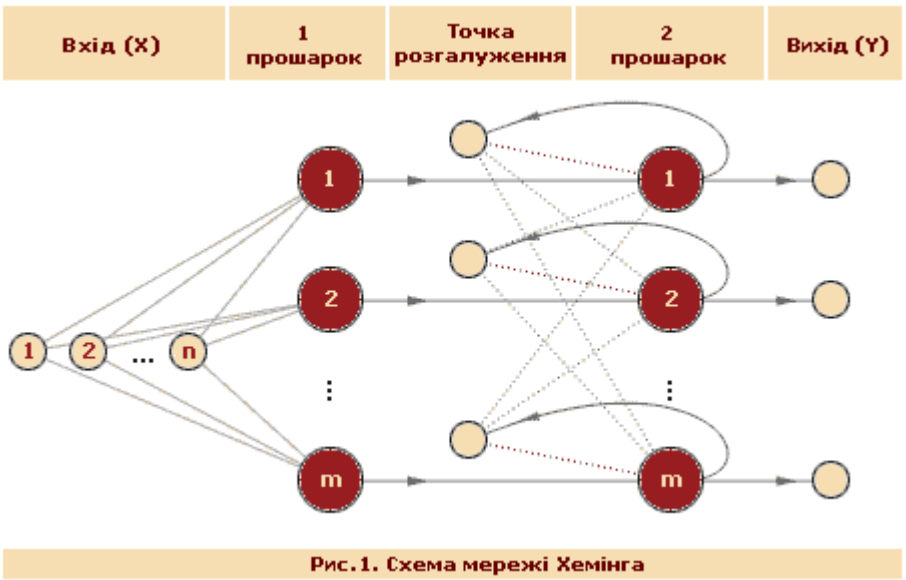


Рис.1. Схема мережі Хемінга

Алгоритм роботи

Ініціалізація вагових коефіцієнтів першого прошарку.

$$w_{ij} = \frac{x_i^j}{2}, i = 0...n\text{-}1, j = 0...m\text{-}1$$,

де X – образи, що запам'ятовуються, i – відповідний компонент вектора X, j – номер образу, n – розмірність вектора X, m – кількість образів, що запам'ятовуються.

Розрахунок стану нейронів першого прошарку.

$$y_j^{(1)} = s_j^{(1)} = \sum_{i=0}^{n-1} w_{ij} x_i + T_j$$,

де X – невідомий образ, T = n/2 – поріг передатної функції

Розрахунок стану нейронів другого прошарку.

$$s_j^{(2)}(p+1) = y_j(p) - \varepsilon \sum_{k=0}^{m-1} y_k^{(2)}(p), k \neq j, j = 0...m-1$$,

де p – номер ітерації, 0 < < 1/m

$$y_j^{(2)}(p+1) = f\left[s_j^{(2)}(p+1)\right], j = 0...m-1$$,

де f – порогова передатна функція.

$$f(x) = \begin{cases} 0, x < 0 \\ x, 0 \le x < F \\ F, x \ge F \end{cases}$$,

де F – поріг передатної функції. Зазвичай F вибирається найменшим, так, щоб при будь-якому допустимому значенні входу не наступало насичення. На практиці F зазвичай береться рівним до кількості прикладів.

Перевірка умови виходу

Якщо виходи не стабілізувалися, тобто змінювалися за останню ітерацію, то перехід на крок 3, інакше кінець.

Область застосування

На думку відразу приходить завдання оптичного розпізнавання символів (OCR). Дійсно, так воно і є. Мережі Хемінга активно використовуються при рішенні таких задач. Але це доволі просте застосування. На практиці вони використовуються для відновлення зашумленого початкового сигналу, завданнях оптимізації і в багатьох інших випадках. Розглянемо один з таких окремих випадків – нечіткий пошук.

Отже, у нас на вході словник, необхідно знайти шукане слово в цьому словнику, навіть якщо воно було набране з помилкою. Для цього потрібно спочатку придумати систему кодування символьної інформації вектора. Задамо для кожного символу його бітову маску.

А	00001
Б	10001
В	10010
...	

Хотілося б звернути увагу, що при кодуванні бажано враховувати джерело отримання інформації. Наприклад, якщо для введення інформації використовується клавіатура, то краще за все задавати коди символів так, щоб у символів, розташованих поряд на клавіатурі, були б і близькі за Хемінгом коди. Якщо ж джерелом є OCR програма, то близькі коди повинні бути у схожих за написанням символів. Після кодування таким чином подаємо отримані вектора на вхід нейромережі.

Тут необхідно враховувати одну особливість мереж Хемінга. Якщо при написанні була друкарська помилка або навіть дві, то алгоритм працює добре, але якщо був пропущений символ або доданий зайвий, то відстань Хемінга може виявитися дуже великою. Для того, щоб згладити цей недолік, подамо на вхід як саме шукане слово, так і це ж слово, де по черзі буде виключено по одному символу в кожній позиції та додано по одній букві в кожну позицію. Такий підхід дозволить знайти практично всі випадки помилок – друкарська помилка, пропуск символу, зайвий символ.

Реалізовану на Perl-е мережу Хеммінга можна випробувати на сайті http://www.basegroup.ru. В якості словника використовується перелік великих міст СНД.

Програма нечіткого пошуку на основі мереж Хемінга

Ця програма призначена для демонстрації можливостей мереж Хемінга у розпізнаванні образів. Завдання нечіткого пошуку вибрано як простий і зрозумілий приклад. Це далеко не єдина область застосування цих мереж. Наприклад, вони використовуються для відновлення образів з неповною і/або спотвореною інформацією.

Як працювати з програмою

Для роботи системи необхідно мати файл з образами (словник). Для цього потрібно відкрити будь-якій текстовій файл. На основі цього файлу система сама складе словник. Після цього потрібно ввести слово для пошуку, програма виявить слово найбільш близьке до нього і зафіксує на ньому вказівник.

Математичний апарат

Мережі Хемінга є одним з різновидів нейронних мереж. Принцип роботи мереж Хемінга базується на визначенні відстані Хемінга між об'єктами і знаходженні найбільш близького. Відстанню Хемінга називається число бітів, що відрізняються, в двох бінарних векторах. Для кодування букв в цифри в нашому випадку використовується ASCII код, хоча можна використовувати і інші методи кодування. Більш того, добре підібравши систему кодування можна значно поліпшити якість розпізнавання.

Наприклад, є сенс для виправленні друкарських помилок, приймати до уваги розташування букв на клавіатурі. Кодування має бути розроблене так, щоб букви, які розташовані на клавіатурі поряд, мали близькі (за Хемінгом) коди.