КЛАСТЕРІЗАЦІЯ ПОТОКІВ ДАНИХ З ВИКОРИСТАННЯМ САМООРГАНІЗОВНИХ КАРТ Т. КОХОНЕНА

Дмитрієв О.В.
Науковий керівник – ас. каф. СТ Жернова П.Є.
Харківський національний університет радіоелектроніки
(61166, Харків, пр.Науки,14, каф. системотехніки, тел. (057) 702-10-06)
e-mail: dmytriievalexzndr@gmail.com

This thesis is devoted to clustering of multidimensional data using T. Kohonen`s self-organizing maps. Today amount of data increases each day. The problem is that it becomes almost impossible too process and cluster this quantity of multidimensional data manually. That`s why we use clustering. Main idea of self-organizing maps is to produce a low-dimensional (typically two-dimensional), discretized representation of input values. This artificial neural network is trained unsupervised learning and apply competitive learning as opposed to error-correction learning. Competitive learning is achieved by using a neighborhood function to preserve the topological properties of the input space.

У наш час кластеризація даних застосовується майже всюди, бо кількість даних зростає і потрібно мати змогу їх розрізняти. Оскільки нові данні з'являються постійно, є важливим обробляти їх «на ходу», тобто одразу реагувати на нові тенданції.

Нейронні мережі Т. Кохонена [1] використовуються для кластеризації багатовимірних даних, які послідовно надходять на обробку, що є важливою частиною Data Stream Mining (обробки потокових даних). При цьому кількість кластерів має бути відомою заздалегідь.

Нейронна мережа Т. Кохонена має один рівень нейронів, кількість яких співпадає з заданою кількістю кластерів. Кожний нейрон має кількість входів, рівну розмірності вхідного вектору. У класичному підході вігові коефіцієнти задаються випадковим чином.

Алгоритм роботи самоорганізовної карти Т. Кохонена:

1. Оберіть початкові значення C для векторів w_j, $j = 1,...,C$. Це можна зробити, вибираючи випадково різні зразки даних C.

2. Виберіть один зразок для набору даних. Це можна зробити або випадковим чином, або послідовно, проходячи через весь набір даних (циклічний порядок).

3. Розрахувати відстань вибраного зразка даних до всіх нейронів-векторів. Як правило, використовується евклідова міра відстані. Нейрон з вектором, найближчим до зразка даних, називається нейроном-переможцем.

4. Оновити вектор нейрона-переможця таким чином, щоб перемістити його до вибраного зразка даних x:

$$w_{win}^{(new)} = w_{win}^{(old)} + \eta\left(x - w_{win}^{(old)} \right). \quad (1)$$

5. Якщо будь-який нейрон-вектор був значно переміщений, скажімо більше, ніж на ε, на попередньому кроці перейдіть до кроку 2, інакше зупиниться.

На кроці 4 розмір кроку (швидкість навчання) η повинен бути обраний належним чином [2]. Для більш швидкої збіжності рекомендується починати з великого розміру кроку, скажімо 0,5, який зменшується в кожній ітерації алгоритму.

Двовимірний SOM є відмінним інструментом для візуалізації розподілів даних високої розмірності. Щоб забезпечити, що сусідні нейрони являють собою подібні області в p-мірному вхідному просторі введена функція сусідства, яка визначає активність тих нейронів, які є сусідами нейрона-переможця.

$$\varphi(i_1, i_2) = \exp\left(-\frac{1}{2} \frac{\left(i_1^{(win)} - i_1 \right)^2 + \left(i_2^{(win)} - i_2 \right)^2}{\sigma^2} \right) \quad (2)$$

де $i_1^{(win)}$ та $i_2^{(win)}$ позначають індекси нейрона-переможця, а i_1 та i_2 позначають індекси будь-якого нейрона. Алгоритм навчання SOM починається з широкої функції сусідства та зменшує його в кожній ітерації. Завдяки цій стратегії в перших ітераціях мережа вивчає грубе представлення розподілу даних та уточнює її, оскільки функція сусідства стає все більш локальною. Це така ж стратегія, що й для розміру кроку η. Якщо процес усадки виконується повільно, небезпека зближення до локального мінімуму зменшується.

Особливостями цього підходу є відносно велика швидкість навчання та навчання без вчителя. Також важливим фактором є те, що вагові коефіцієнти перераховуються після кожного вхідного вектору, що дозволяє досить ефективно використовувати такий підхід при кластерізації потоків даних.

Література

[5]. Kohonen, T. Self-Organizing Maps. Springer-Verlag, Berlin, 1995; 362 p.
[6]. P. Zhernova, A. Deyneko, Y. Bodyanskiy and V. Riepin, "IEEE Second International Conference on Data Stream Mining & Processing," in Adaptive kernel data streams clustering based on neural networks ensembles in conditions of uncertainty about amount and shapes of clusters, Lviv, Ukraine, 2018.