КЛАСТЕРИЗАЦІЯ ДАННИХ ВИСОКОЇ РОЗМІРНОСТІ З ВИКОРИСТАННЯМ МОЖЛИВІСНОГО ПІДХОДУ

Жернова П.Є., Лобинцев А.А.
Харківський національний університет радіоелектроніки
(61166, Харьков, пр. Науки,14, каф. системотехніки, тел. (057) 702-10-06)
e-mail: polina.zhernova@gmail.com, 0661394730;
kyks1997@gmail.com, 0500507369

The neural network's approach for data stream clustering task, that in online mode are fed to processing in assumption of uncertainty about amount and shapes of clusters, is proposed in the paper. The main idea of this approach is based on the kernel clustering and idea of neural networks ensembles that consists of the T. Kohonen's self-organizing maps. Each of the clustering neural networks consists of different number of neurons, where number of clusters is connected with the quality of these neurons. All ensemble members process information that sequentially is fed to the system in the parallel mode. Experimental results have proven the fact that the system under consideration could be used to solve a wide range of Data Stream Mining tasks.

Завдання кластеризації багатовимірних спостережень, які послідовно надходять на обробку, є важливим напрямком в рамках Data Stream Mining, а для її вирішення розроблено досить велику кількість різних методів. Необхідно відзначити нейронні мережі Т. Кохонена [1], які найкращим чином пристосовані для обробки інформації в online режимі. При цьому апріорно передбачається, що кількість кластерів, на яку розбивається аналізований масив даних, відома заздалегідь.

Таким чином, якщо задана вибірка даних (можливо зростаюча) $X = \{x(1),...,x(2),...,x(k),...,x(N),...\} \subset R^n$, яка подається на входи ансамблю, який працює в припущенні, що число можливих кластерів $2 \le m^* \le M$ в шарі Кохонена, де m^* визначає істинну кількість класів в оброблюваній вибірці.

Ситуація істотно ускладнюється, якщо кластери які формуються, перетинаються у просторі ознак. Такі завдання вирішуються за допомогою методів нечіткої кластеризації [2, 3], найбільш популярним з яких є алгоритм нечітких С-середніх (FCM). Для роботи в online режимі з успіхом можуть бути використані нечіткі кластерувальні мапи Кохонена [4].

Тому є доцільною розробка ансамблю нечітких карт Кохонена з використанням тих або інших цільових функцій для вирішення задач кластеризації за умов невідомої кількості класів, що довільним чином перетинаються у просторі ознак.

Для можливісного підходу до кластеризації критерій, що мінімізується, має вигляд

$$E_k\left(u_j(k), w_j\right) = \sum_{j=1}^{m} u_j^{\beta}(k)\left\|x(k) - w_j\right\|^2 + \sum_{j=1}^{m} \mu_j\left(1 - u_j(k)\right)^{\beta}, \qquad (1)$$

де скалярний параметр $\mu_j > 0$ визначає відстань, на якій рівень належності приймає значення 0.5, тобто якщо $\left\|x(k) - w_j\right\|^2 = \mu_j$, то $u_j(k) = 0.5$.

Робота процедури кластеризації розпочинається із завдання початкової (зазвичай випадкової) матриці розбиття W^0. На основі її значень обчислюється початковий набір прототипів w_j^0, які потім використовуються для уточнення нової матриці W^1. Наступним кроком в пакетному режимі є обчислення $W^2,\ldots,W^t,W^{t+1}$ і так далі, доки різниця $\left\|W^{t+1} - W^t\right\|$ не стане меншою за деяке наперед задане порогове значення ε. Таким чином, вся вибірка даних оброблюється багатократно.

В межах можливісного підходу результат оптимізації при $\beta = 2$ записується як:

$$u_j^{pos}(k) = \frac{\mu_j(k)}{\mu_j(k) + \left\|x(k) - w_j(k)\right\|^2}, \qquad (2)$$

$$w_j^{pos}(k+1) = w_j^{pos}(k) + \eta(k)u_j^2(k)\left(x(k+1) - w_j^{pos}(k)\right), \qquad (3)$$

$$\mu_j(k+1) = \frac{\sum_{p=1}^{k} u_j^2(p)\left\|x(p) - w_j(k+1)\right\|^2}{\sum_{p=1}^{k} u_j^2(p)}. \qquad (4)$$

Запропонований підхід є узагальненням низки відомих процедур нечіткої можливісної кластеризації та може бути використаний для вирішення задач аналізу потоків даних великої розмірності.

Література

[1]. Kohonen, T. Self-Organizing Maps. Springer-Verlag, Berlin, 1995; 362 p.
[2]. Bezdek, J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. N.Y., Plenum Press, 1981; 272 p
[3]. Bezdek, J.; Keller, J.; Krisnapuram, R.; Pal, N. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer, 2005; 776 p.
[4]. Gorshkov, Ye.; Kolodyazhniy, V.; Bodyanskiy, Ye. New recursive learning algoritms for fuzzy Kohonen clustering network. In Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems. Rapperwil, Switzerland, 2009; pp. 58-61.