D. Khominich, Master student
S. Smirnov PhD, research advisor
O. Medkova PhD in Eng., As. Prof., language advisor
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

ADAPTIVE CLUSTERING IN DATA MINING

Clustering is the task of dividing a set of objects into groups called clusters.

To describe the degree of similarity, the space in which the objects are located, the scalar metric d (x, y) can be taken, this is the distance between any two objects. This metric must be symmetrical, inherent, and consistent with the triangle equation. The most popular and fastest is the K-means clustering method.

K-means is a cluster analysis method, the purpose of which is to divide m observations into k clusters, with each observation referring to the cluster to the center (centroid) of which it is closest. Euclidean distance is used as a measure of proximity $\rho(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n}(x_p - y_p)^2}, x, y \in R^n$.

The k-means method divides m observations into k groups ($k \leq m$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ to minimize the total square deviation of cluster points from the centroids μ_i of these clusters S_i:

$$\min \left[\sum_{i=1}^{k} \sum_{x^{(j)} \in S_i} \left\| x^{(j)} - \mu_i \right\|^2\right], x^{(j)} \in R^n, \mu_i \in R^n,$$

Adaptive clustering is understood as an analysis in which the parameters that determine the result are selected and adjusted in the process of performing the task, based on the specified criteria and the recommendations given by the expert, in order to achieve the best result. Each of the parameters is suitable for use in certain cases (for example, the Squared Euclidean distance is used to give more weight to objects that are more distant from each other).

The number of clusters, the threshold value for stopping the operation of the algorithm, the method for choosing the initial centers, the maximum number of iterations, the amount of simultaneously processed data, the number of previous sections, the distance coefficient, the exact value of these parameters is unknown and is selected in the range of values from 2 to | X | clusters, as well as through expert judgment.

Using standard values can lead to very poor results. Expert estimates of the algorithm launch parameters will be averaged and refined in the process of applying the algorithm. In general terms, the criterion for assessing the quality of the clustering task is a numerical indicator that is calculated based on the results of clustering at a given iteration, the essence of which is a quantitative assessment of the quality of the solution.

1. Partition clarity indicators:
   - Partition factor:

$$PC = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K} u_{ij}^2}{Q}, \quad PC \in \left[\frac{1}{K}, 1\right].$$

   - Clarity index:

$$CI = \frac{K \cdot PC - 1}{K - 1}, \quad CI \in [0,1].$$

2. Entropy criteria (the lower the entropy value, the better the clustering is done)
   - Modified entropy

$$PE_M = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K} u_{qk} \ln(u_{qk})}{Q \ln K} = \frac{PE}{\ln K}, \quad PE_M \in [0,1].$$

3. Compactness indicator:

$$CS = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K} u_{qk}^2 \cdot d^2(x_q, c_k)}{Q \cdot \min\left\{d^2(c_i, c_j) \middle| i, j \in \overline{1,K}, i \neq j\right\}}.$$

The low value of this criterion says that all our clusters are well separable from each other, that is, they differ that 1 cluster retains 1 class.

4. Index of efficiency:

$$PI = \sum_{k=1}^{K}\sum_{q=1}^{Q} u_{qk}^2 \left(d^2(c_k, \overline{x}) - d^2(x_q, c_k)\right).$$

With the larger criterion, the larger the optimal number of clusters for our tasks. Iterative criterion.

In practice, these actions give us great efficiency and the ability to use our model several times for different tasks. In practice, the dependence of the quality of clustering on the number of clusters is greater, because this is very much. In practice, the choice of the number of clusters is one of the most important and difficult tasks. Choosing the right number of clusters will give the best result.