12-1-2021

How to Select Typical Objects

Mariana Benitez
The University of Texas at El Paso, mbenitez3@miners.utep.edu

Jeffrey Weidner
The University of Texas at El Paso, jweidner@utep.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-21-104

How to Select Typical Objects

Mariana Benitez, Jeffrey Weidner, and Vladik Kreinovich

Abstract In many practical situations, we have a large number of objects, too many to be able to thoroughly analyze each of them. To get a general understanding, we need to select a representative sample. For us, this problem was motivated to analyze the possible effect of an earthquake on buildings in El Paso, Texas. In this paper, we provide a reasonable formalization of this problem, and provide a feasible algorithm for solving thus formalized problem.

1 Formulation of the problem

General problem. We have a large number of objects N. Each object is characterized by the values of q quantities. Let us denote the value of the j-th quantity for the i-th object by v_{ij}. Then, the object i is characterized by a tuple

$$v_i = (v_{i,1}, \ldots, v_{i,q}).$$

We can only thoroughly process $n \ll N$ objects. We therefore want to select n out of N objects so that the resulting sample of n objects be the most representative; see, e.g., [2].

Case study. We are interested in possible effect of an earthquake on buildings in El Paso, Texas – a potentially seismic area in which, however, earthquakes have been very rare. There are many thousands of buildings in El Paso, it is not realistic to thoroughly analyze each of them. So, we need to select a feasible-to-analyze sample. For this problem, each buiding is characterized by 4 parameters: occupancy, age (i.e., equivalently, year of construction), number of stories, and height.

Mariana Benitez, Jeffrey Weidner, and Vladik Kreinovich
University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA
e-mail: mbenitez3@miners.utep.edu, jweidner@utep.edu, vladik@utep.edu

2 Main idea and how we can implement it

In the general case, we want to make sure that each object is similar to one of the selected objects. How can we describe this similarity? In general, the q quantities have different effect on the properties that we want to analyze: a difference of one unit in one quantity may affects this property much more than a difference in 1 unit in some other quantity.

For example, in our case study, the 1 year difference in the building's age will have practically no effect on the building's stability against a strong earthquake, but a difference in 1 story can drastically change this stability – e.g., if we consider the difference between 1-story and 2-story buildings.

To take this into account, it makes sense to "equalize" these quantities. For example, if the effect of adding 1 story is roughly equivalent to the effect of adding w years to the age, this means that adding s stories is equivalent to adding $w \cdot s$ years. We can estimate similar "weights" for other quantities, so that for the correspondingly equalized quantities

$$e_{i,j} \stackrel{\text{def}}{=} w_j \cdot q_{i,j} \tag{1}$$

the unit change in each of these quantities has approximately the same effect on the property of interest. In the following text, we will assume that the values of the weights have been found, and that the values of the quantities have already been equalized. In these terms, each object i is characterized by the tuple

$$e_i = (e_{i,1}, \ldots, e_{i,q}).$$

In geometric terms, each tuple e_i can be represented as a point in a q-dimensional space. So, to describe the degree of dissimilarity between the two objects i and i' characterized by the tuples $e_i = (e_{i,1}, \ldots, e_{i,q})$ and $e_{i'} = (e_{i',1}, \ldots, e_{i',q})$, it is reasonable to take the distance between these two q-dimensional points, i.e.. the value

$$d(e_i, e_{i'}) \stackrel{\text{def}}{=} \sqrt{\sum_{j=1}^{q} (e_{i,j} - e_{i',j})^2}.$$

Our goal is to select, among N given objects $1, \ldots, N$, n typical objects $t(1), \ldots, t(n)$. Once we have selected them, then, for each object i, as its approximate representation, we will take the typical object $t(n(i))$ which is the closest to the object i, i.e., for which the distance to the i-th object is the smallest:

$$d(e_i, e_{t(n(i))}) = \min_{k=1,\ldots,n} d(e_i, e_{t(k)}).$$

In general, the distance is the smallest if and only if the square of the distance is the smallest, so

$$d^2(e_i, e_{t(n(i))}) = \min_{k=1,\ldots,n} d^2(e_i, e_{t(k)}). \tag{1}$$

We want to make sure that for each object i and for each (equalized) quantity j, the values of this quantity for the original object i and for the approximating typical object $t(n(i))$ be close, i.e., that we should have $e_{i,j} \approx e_{t(n(i)),j}$. In other words, we want to make sure that following approximate equalities hold:

$$e_{1,1} \approx e_{t(n(1)),1}, \ldots, e_{1,q} \approx e_{t(n(1)),q},$$

$$\ldots$$

$$e_{N,1} \approx e_{t(n(N)),1}, \ldots, e_{N,q} \approx e_{t(n(N)),q}.$$

We want these approximate equalities to be as accurate as possible. This means that the distance between the tuple

$$\ell = (e_{1,1}, \ldots, e_{1,q}, \ldots, e_{N,1}, \ldots, e_{N,q})$$

formed by all the left-hand sides and the tuple

$$r = (e_{t(n(1)),1}, \ldots, e_{t(n(1)),q}, \ldots, e_{t(n(N)),1}, \ldots, e_{t(n(N)),q})$$

formed by all the right-hand sides should be as small as possible. As we have mentioned, the distance is the smallest if and only if the square of the distance is the smallest. Thus, we must select the typical values $t_1, \ldots, t_n$ for which the value

$$(e_{1,1} - e_{t(n(1)),1})^2 + \ldots + (e_{1,q} - e_{t(n(1)),q})^2 +$$

$$\ldots +$$

$$(e_{N,1} - e_{t(n(N)),1})^2 + \ldots + (e_{N,q} - e_{t(n(N)),q})^2$$

is the smallest possible. The sum

$$(e_{1,1} - e_{t(n(1)),1})^2 + \ldots + (e_{1,q} - e_{t(n(1)),q})^2$$

of the first q terms in this expression is simply the square $d^2(e_1, e_{t(n(1))})$ of the distance between the tuples e_1 and $e_{t(n(1))}$. Similarly, the sum of the next q terms is the square $d^2(e_2, e_{t(n(2))})$ of the distance between the tuples e_2 and $e_{t(n(2))}$, etc. So, the overall expression that we want to minimize has the form

$$\sum_{i=1}^{N} d^2(e_i, e_{t(n(i))}).$$

In view of the formula (1), this expression takes the form

$$\sum_{i=1}^{N} \min_k d^2(e_i, c_k), \tag{2}$$

where we denoted $c_k \stackrel{\text{def}}{=} e_{t(k)}$.

Minimizing this expression is exactly the problem solved by k-means clustering (see, e.g., [1]), where each c_k is called the center of the k-th cluster. The only difference between the k-means and our problem is that:

- in the k-means clustering, we can take any point c_k, while
- in our problem, c_k must be one of the original points e_i.

Thus, after we apply the k-means clustering algorithm and get the resulting values c_k, then, for each k, we must find the point $t(k)$ which is the closest to c_k:

$$d(e_{t(k)}, c_k) = \min_i d(e_i, c_k).$$

So, we arrive at the following algorithm.

3 Resulting Algorithm

We start with N objects $i = 1, \ldots, N$ characterized by tuples $v_i = (v_{i,1}, \ldots, v_{i,q})$. Among these objects, for some pre-defined value n, we want to select n most representative ones. To do this, we use the following algorithm:

- first, for each of q quantities $j = 1, \ldots, q$, we find the "equalizing" weight w_j, i.e., the weight such that the effect of adding 1 unit to quantity j is equivalent to the effect of adding w_j units to the quantity 1;
- then, we use the weights w_j to equalize all the values $v_{i,j}$ into the values $e_{i,j} = w_j \cdot v_{i,j}$; this way, we get N tuples $e_i = (e_{i,1}, \ldots, e_{1,q})$;
- next, we apply the k-means algorithm to these N tuples and find the centers $c_1, \ldots, c_n$ of the corresponding clusters;
- finally, for each k from 1 to n, we find the original tuple closest to this c_k, i.e., the tuple $e_{t(k)}$ for which the distance $d(e_{t(k)}, c_k)$ is the smallest possible.

As the resulting "most representative" set of n objects, we select the objects

$$t(1), \ldots, t(n).$$

Comment. In addition to "typical" objects, we may also want to select one or more extreme objects – to make sure that we do not miss the objects for which the effect is expected to be the largest.

For example, in the earthquake-analysis case, in which the effect increases with an increase in each of the values $v_{i,j}$, we may want to consider the building with the largest possible value of the corresponding weighted sum $\sum_j w_j \cdot v_{i,j}$.

Acknowledgments

This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to Michael Beer from Leibniz University, Hannover, for valuable discussions.

References

1. J. C. Bezdek, *Elementary Cluster Analysis: Four Basic Methods that (Usually) Work*, River Publishers, Gistrup, Denmark, 2021.
2. J. Salomon, M. Broggi, S. Kruse, S. Weber, and M. Beer, "Resilience decision-making for complex systems", *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 2020, Vol. 6, Paper 020901-1.