Qeios

Peer Review

Review of: "Fixed-Sized Clusters k-Means"

Andrzej Młodak[1]

1. Inter-faculty Department of Mathematics and Statistics, Akademia Kaliska, Kalisz, Poland

The paper contains an interesting proposal for a clustering algorithm, aimed at obtaining optimal clusters of similar size. It is based on the cluster slots, which are as many as the objects being analyzed. The clustering is then made by minimization of weights associated with every pair of object-slot from the point of view of the bijection assigning to each object a slot to which it should belong. In the proposed solution, the weight is computed as a squared distance from an object to the centroid of the cluster for which the cumulative sum of cluster sizes is minimal but not smaller than the established thresholds.

The paper deserves to be widely known. However, it seems to be carelessly written. First of all, the proposed solution should be well motivated. Only the need to have evenly distributed cluster sizes is rather not sufficient. The added value in relation to, e.g., balanced clustering should be emphasized. Moreover, it is unclear what the “equal or more equal number of points in each cluster” (on page 2) means. The relation of equality is unique. Two numbers are equal or not, and nothing more.

The second question is the cluster slots. If their number at the start equals the number of objects, it is obvious that many slots will create empty clusters, which can extend the calculation time unnecessarily. A short discussion about the optimal number of slots would then be desirable. I recommend also giving an example of what the weight and bijection can be in practice and what importance they can have.

The application (Section 4) is very hard to understand. What does "the compatibility of persons within tables" mean? What criterion of compatibility is here applied: e.g., the same sex, similar occupation, similar political views, lack of conflicts of interest, or something else? It is worth noting that some of such compatibility criteria can be subjective and not quantifiable. How can one measure them? All computation should be exactly presented and described. The content of subsection 4.1 does not allow for an assessment of what the authors actually calculated and what value it has.

doi.org/10.32388/KYY838

That is, can you suggest specific examples or scenarios where the proposed solution provides a clear advantage over existing balanced clustering methods?

Moreover, there are some typos and technical errors in the paper. Below I present them:

1. In formula (1), instead of "$X_i \in C_j$", it should be "$i \in C_j$" (an object belongs to a cluster, not data).
2. In the formula defining $C_j^{(t+1)}$, instead of "$X_i \in P_j^{(t)}$", it should be "$i \in P_j^{(t)}$" (the same reason as above).
3. In the last paragraph of Section 1, instead of "constraits", it should be "constraints" (two times). Moreover, whose and what experiments do the authors mean by writing "their experiments"?
4. At the beginning of Section 2, A is defined as the set of cluster slots, whereas S is defined as the original data sets. The formulas and considerations in the further part of the paper suggest that it is the other way around, i.e., S denotes a set of slots and A denotes the original data.
5. In formula (2), instead of "$X_j \in C_i^{(t)}$", it should be "$j \in C_i^{(t)}$" (an object belongs to a cluster, not data).
6. It would be desirable to mention - directly after formula (3) - that $c(k) = n$.

Declarations

Potential competing interests: No potential competing interests to declare.