Reply to reviewers concerning submission PONE-D-17-23730: "Efficient similarity-based data clustering by optimal object to cluster reallocation"

October 17, 2017

As a preamble, we would like to thank the editor and the reviewer for their comments and suggestions. Following these comments, we made several changes to the article, which are summarized here. The next sections list our answers to each of the reviewers comments, with references to the revised manuscript (page, column, and paragraph) where appropriate.

1 Answers to Academic Editor

1. *The manuscript requires to address all the points raised in the reviewers' reports, mainly technical issues about the proposed technique, but also English writing mistakes which make it difficult to understand the paper.*

   → We made our best to fix all the issues raised by the reviewers and to improve the readability of the paper.

2. *We are quite skeptical about the real significance of the contribution, since most of the ideas have already been accounted for in the (old) literature.*

   → Our contribution lies in the uses of two concepts which have been long standing. We completely agree on that point. Though, we are unaware of previous contributions explicitly presenting the algorithm we propose together with detailed complexity analysis, memory requirements, experimentations, etc.

   Many people from the machine learning community uses the kernel kmeans on arbitrary similarity matrices, eventhough the convergence is in this case not guaranteed. We believe, that our proposal is, in the case of arbitrary similarity matrices considered as input, a good alternative that provides guaranteed convergence, similar clustering performance, this for lower cpu and memory requirements.

2 Answers to Reviewer I

1. *Some points are not clear, and there are many English mistakes. For example, (1) an maximization of : a maximization of (abstract) (2) relaxing the conditions : what are "the conditions" here? (abstract) (3) our simple non-quadratic updates makes : make (Section 4) (4) times series : time series (Section 7)*

   → The typos have been corrected and the whole manuscript have been carefully scrutinized. We clarified (2) by providing an example.

2. *The contribution of this manuscript is not clear (or not sufficient). It is a well-known fact that minimizing the sum of squared distances between a data point and its cluster center can be converted into a form of maximizing the within-cluster similarity as noted in Duda et al. (Chapter 10.7). Also, considering the actual change of the objective function leads to a natural extension of the batch k-means as described in Duda et al. (Chapter 10.8). These two well-known ideas seem to be the main ideas of the proposed algorithm.*

   → Those properties are indeed well known and clearly stated early in the manuscript. The main contribution of the paper is to demonstrate that the combination of those 2 properties lead an algorithm, that is to the best of our knowledge, new or at least whose properties in terms of convergence for arbitrary similarity measures, convergence and efficiency do not have been studied extensively in the literature.

   We added a paragraph at the end of the introduction to more clearly state the contributions of the paper.

3. *In Section 7, the proposed algorithm has been compared with only one baseline method (kernel k-means). Since there are a number of clustering algorithms, other state-of-the-art clustering algorithms also should be compared with the proposed algorithm to show the usefulness of the method. Also, in Section 7.3, the run time of the methods should be presented.*

   → Following this comment and the comment of Reviewer II, we added the results for the spectral clustering technique, which is also widely used for dealing with proximity measures.

4. *in Section 7.3, the run time of the methods should be presented.*

   → We made a new section (section 7.4) dedicated to a runtime comparison of the different algorithms.

3 Answers to Reviewer II

1. *The flexibility to use arbitrary similarity measures is a novel contribution. Exploiting only points that change cluster membership is a clever idea and leads to lower complexity.*

→ Those are indeed the main contributions of the paper. Hopefully the manuscript now stresses more those contributions.

2. *Convergence Issues: Permitting arbitrary similarity measures raises questions about convergence. In case of kernel k-means, the similarity measure is obtained using a positive definite kernel. Such a kernel implicitly assumes a Hilbert space, where Euclidean distances can be computed by virtue of inner products. Such a mapping through kernel functions is the premise for convergence of the basic iterative MSE procedure described in Duda & Hart Ch 10.8. As arbitrary similarity measures are permitted, it is not clear how the convergence properties are impacted. Intuitively and empirically (based on Fig. 1) it seems that convergence should happen, [...]. Do arbitrary similarity measures always converge?*

   → The k-average algorithm converge by design, see Section 3.2: "To ensure global convergence, we need to compute the impact on the global objective function of moving one object from one class to another. Using such formulation and performing only reallocation that have a positive impact, the convergence of such an iterative algorithm is guaranteed."

3. *however, unfortunately the paper does not discuss the similarity measure that is used for experiments in Fig. 1*

   → The similarity measure is the Dynamic Time Warping (DTW) has it was found to be the best performing measure for those datasets as described in the beginning of Section 7. The caption of Fig. 1 have been extended to provide this information.

4. *Sensitivity to clustering performance: Duda & Hart (in Ch. 10.8) point out that the iterative optimization scheme is more susceptible to local minimum. I expect this effect to be exacerbated when arbitrary similarity measures are used. However, this has not been thoroughly investigated.*

   → This is indeed a concern. We found experimentally that the k-average algorithm is, at least experimentally, not more susceptible to local minima than the other studied approaches, see comment below and discussion in Section 7.3.

5. *Initialization of the algorithm: In Alg. 3, the initialization 'L', i.e., the initial point-cluster assignment is done randomly. This is not the case in kernel k-means. It is surprising that despite the random initialization, the performance of k-average clustering matches the kernel k-means in many cases. As in the case of k-means and its iterative version, I would expect that the final solution depends on the initialization. Randomly initializing cluster labels ignores any notion of similarity between points, and are likely to lead to local poor minima in the overall optimization. This fact has been completely ignored and not investigated at all in the paper.*

   → As stated in Section 7.2, all the algorithms are given the same initialization: "clustering is done by requesting a number of clusters equal to

the actual number of classes in the dataset, and repeated 200 times with varying initial conditions, each algorithm being run with the exact same initial assignments."

That is, for each run, each object is given a label from 1 to C, C being the number of classes. This labeling is used to initialize the k-average, kernel k-means and kmeans algorithm used in the spectral clustering technique.

There indeed exist strategies to better initialize the iterative scheme, but we believe that this matter is out of the scope of the paper which is about comparing the behaviors of the studied clustering methods.

In order to evaluate the influence of the initialization, we no longer consider the nmi of the best performing run in terms of objective function, but the mean and variance of the nmi obtained for each dataset. As now discussed in the paper in Section ??, the variance is comparable for each methods meaning that the k-average method

6. *Writing and Grammar: The paper is very poorly written. There are several grammatical mistakes (e.g., 'an maximization' in the abstract. There are several more throughout the paper).*

   → Several typos have been corrected and the whole manuscript have been carefully improve for better readability.

7. *in (1), the notation used for vector multiplication (inner and outer products) is incorrect.*

   → Equation 1 have been rewritten. Vector orientation and notation are also better introduced in the preceding paragraph.

8. *On page 7, the statement that $\mathcal{Q}$ is strictly equivalent to the average point to centroid similarity' is not true in general, but only in case when the similarity can be interpreted as an inner product. If this is not the case, I suppose a proof to back this statement up is necessary.*

   → We removed this sentence as it would indeed require more detailed explanations that we believe are out of the scope of this section.

9. *δ_o in (9) is not introduced.*

   → Introduction is now provided.

10. *I would expect that comparisons with spectral clustering in terms of quality of clustering should also be included. Spectral clustering also permits arbitrary similarity measures and thus is directly relevant to the proposed work.*

    → We added the spectral clustering technique to the reference set and discussed the matter. The main issue with the spectral clustering techniques is that it requires a diagonalization of a N^2 matrix with N being the number of objects, making it difficult to scale to very large datasets.