Consensus clustering for Bayesian mixture models

Point-by-point response to round 2 reviewers' comments

Stephen Coleman, Paul D.W. Kirk, Chris Wallace

Reviewer 1

I have a major concern on the research problem or research motivation. In the introduction part, the authors mentioned several concepts, consensus clustering, Bayesian clustering. However, there is no clear connection between these two terms. In another word, the motivation of this paper is not clear. Moreover, the related work on consensus clustering is not extensive. Some recent work is not included. I did not see strong competitive methods in the experimental section.

In our opinion, the combination of consensus clustering and model-based Bayesian clustering is not controversial: in the abstract of the seminal Monti *et al.* 2003 paper ("Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data"), it was suggested that "[Consensus clustering] can also be used to represent the consensus over multiple runs of a clustering algorithm with random restart (such as K-means, model-based Bayesian clustering, SOM, etc.)". Thus, they are already established as connected concepts. The novelty of our current work is to explore the degree to which we can successfully use consensus clustering to combine a large number of *short* runs of model-based Bayesian clustering implementations with different initialisations, which (thanks to parallelisation) may be less computationally costly than running a small number of long runs.

Our motivation is stated at the end of the Introduction:
"Motivated by the lack of scalability of existing implementations of sampling-based Bayesian clustering (due to prohibitive computational runtimes, as well as poor exploration, as described above), here we aim to develop a general and straightforward procedure that exploits the flexibility of these methods, but extends their applicability. Specifically, we make use of existing sampling-based Bayesian clustering implementations, but only run them for a fixed (and relatively small) number of iterations, stopping before they have converged to their target stationary distribution. Doing this repeatedly, we obtain an ensemble of clustering partitions, which we use to perform consensus clustering."

As the focus of our paper is the application of consensus clustering to model-based Bayesian clustering, we have cited the literature that is relevant to this focus, including the original consensus clustering paper ([10]), a number of consensus clustering implementations ([11-13]) and applications ([14-16]) and then some references to other work to highlight some developments over the years in the area of consensus clustering ([17, 18, 21]). These references span a range of publication dates from 2003-2020.

Since a principal aim of our manuscript is to demonstrate that the proposed application of consensus clustering can extend the applicability of existing implementations of sampling-based Bayesian mixture models (and their elaborations), our direct competitor is "full" Bayesian inference. In the experimental section of our manuscript, we therefore compare to Bayesian inference via MCMC, as well as to Mclust, a commonly used model-based clustering approach that performs fitting via maximum likelihood estimation.

[10] Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning. 2003;52(1-2):91–118.
[11] Wilkerson, D M, Hayes, Neil D. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572–1573.

[12] John CR, Watson D, Russ D, Goldmann K, Ehrenstein M, Pitzalis C, et al. M3C: Monte Carlo reference-based consensus clustering. Scientific reports. 2020;10(1):1–14.
[13] Gu Z, Schlesner M, Hübschmann D. cola: an R/Bioconductor package for consensus partitioning through a general framework. Nucleic Acids Research. 2020. Available from: https://doi.org/10.1093/nar/gkaa114
[14] Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. The Journal of clinical investigation. 2011;121(7):2750–2767.
[15] Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1,EGFR, and NF1. Cancer cell. 2010;17(1):98–110.
[16] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nature methods. 2017;14(5):483–486.
[17] Li T, Ding C. Weighted Consensus Clustering. In: Proceedings of the 2008 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics; 2008. p. 798–809.
[18] Carpineto C, Romano G. Consensus Clustering Based on a New Probabilistic Rand Index with Application to Subtopic Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012 Dec;34(12):2315–2326.
[21] Ünlü R, Xanthopoulos P. Estimating the Number of Clusters in a Dataset via Consensus Clustering. Expert Systems with Applications. 2019 Jul;125:33–39

Reviewer 2

I have greatly appreciated the efforts made by the authors to clarify the aim of the manuscript. This is clearer now, as well as clearer are the original contributions of the manuscript.

I still find that there are a couple of points where some of the authors' sentences/conclusions are too strong and misleading. Please, correct them.

- p 15 "However, three different modes emerge across the chains showing that the chains are failing to explore the full support of the posterior distribution of the clustering and are each unrepresentative of the uncertainty in the final clustering. This shows that consensus clustering is exploring more possible clusterings than any individual chain and, as it explores a similar space to the pooled samples which might be considered more representative of the posterior distribution than any one chain, it suggests it better describes the true uncertainty present than any single chain. It also shows that pooling chains offers robustness to multi-modality (as expected for an ensemble) and ..."

- p 21 "However, we have shown that if a finite Markov chain fails to describe the full posterior distribution, our method frequently has better ability to represent several modes in the data than individual chains and thus offers a more consistent and reproducible analysis. We also showed that the ensemble of short chains is more robust to irrelevant features than Mclust."

I believe that you should also add a restatement of the points above along the following lines.
Any finite MCMC chain might suffer in representing the full support of the whole posterior distribution since it is finite. Convergence theorems hold as the number of iterations goes to infinity. The mixing of the chain can be poor as well. Your sentences state that, in your experience (no theorems here), many short runs are computationally less expensive that one long chain, but give similar point and interval estimates than using the limit behaviour, and that they are more convenient for your application.
Using these words, I agree with it.

Specific comments:
1) citation 24 in your list, Ferguson TS. A Bayesian analysis of some nonparametric problems. The annals of statistics. 1973, 209–230" is the paper where Ferguson introduces the Dirichlet process prior. The paper does not mention Dirichlet process mixtures that were introduced in Lo, A. Y. (1984). "On a Class of Bayesian Nonparametric Estimates: I, Density Estimates," The Annals of Statistics, 12, 351–357 or in Ferguson, T. S. (1983). "Bayesian density estimation by mixtures of normal distributions." In Rizvi, H. and Rustagi, J. (eds.), Recent Advances in Statistics, 287-303. New York: Academic Press.

2) As far as citation 38 is concerned, please, do not cite the arxiv version of the paper. I found this version: Neiswanger,W.,Wang, C., and Xing, E. (2014). "Asymptotically Exact, Embarrassingly Parallel MCMC." In Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence, 623–632.

Thank you for the specific comments and suggested restatement of the points. We have reworded these sections as follows:

p15:

“For the PSMs from the individual chains, all entries are 0 or 1. This means only a single clustering is sampled within each chain, implying very little uncertainty in the partition. However, three different clustering solutions emerge across the chains, indicating that each individual chain is failing to explore the full support of the posterior distribution of the clustering. In general, while MCMC convergence theorems hold as the number of iterations tend to infinity, any finite chain might suffer in representing the full support of the posterior distribution, as we observe here. Moreover, the mixing of each chain can be poor as well (i.e. it may take a long time to reach the stationary distribution from an arbitrary initialisation). In our empirical study, we find that using many short runs provide similar point and interval estimates to running a small number of long chains (figure 3), while being computationally less expensive (figure 4), and hence more convenient for our applications.”

p21:
“We have shown cases where many short runs are computationally less expensive than one long chain and give meaningful point and interval estimates; estimates that are very similar to those from the limiting case of a Markov chain. Thus if individual chains are suffering from mixing problems or are too computationally expensive to run, consensus clustering may provide a viable option.”

Thank you, we have updated our references accordingly.