Reviewer’s report

Title: Cluster Ensemble Based on Random Forests for Genetic Data

Version: 0 Date: 23 Sep 2017

Reviewer: Michael Banf

Reviewer's report:

The paper presents an approach to infer population structures based on genomic datasets, such as Single Nucleotide Polymorphism data. It is based on the idea of constructing an ensemble of individual clusterings, using a final consensus function. The base clustering method proposed is an unsupervised extension of the traditional random forest method.

While reading the paper, several theoretical and practical questions and concerns came up that I would like to see addressed for a better evaluation of the approach presented.

The general idea of using random forest in an unsupervised fashion to facilitate clustering, as summarized in afanador et al., 2016, "Unsupervised random forest: a tutorial with case studies" already makes use of the property of random forests to be an ensemble of individual decision trees. Therefore, in the unsupervised setting the proximity matrix, as the direct input for clustering methods, has already been constructed as a result of an ensemble of trees, averaging across all trees. This approach can be superior to linear approaches such as PCA, given the random forests general properties. However, to me, it is not clear, why an additional round of generating multiple random forest ensembles, followed by multiple proximity matrices, which subsequently used for individual clustering with the results in turn being integrated by a consensus function (basically averaging across all individual clusters) would do better than just increases the number of trees in a single round of (ensemble based) random forest, thereby increasing the accuracy of a single proximity matrix which subsequently would increase clustering quality.
To this end, I was searching for the specific parameter settings and post-processing setups of the competing approaches, in particular the RFclust approach (a single clustering method based on calculating the average proximities of multiple random forests and then applying Ward's hierarchical clustering.), which seems to be missing. Given that the margin of performance improvements of RFclue in 9 out of 12 case studies is rather small, I am wondering whether this is caused by the approach itself, or due to an invalid comparison.

As an example, I am wondering how many trees were grown in the single random forest based clustering approach (RFclust). To be comparable to RFclue, especially since the authors show that cluster accuracy is correlated with the number of trees, for a fair comparison, I would assume that RFclust should at least have N * M trees, with N being the number of trees in each random forest in RFclue and M being the number of individual clusterings in RFclue.

Further, RFclust should be followed by exactly the exactly the same post processing (MDS, k-means, etc. with exactly same parameters) as RFclue. Therefore, TrueK, FixedK, RandomK should also evaluated in RFclust.

Finally, the ensemble clustering approaches, including RFclue, are given the true sub population number for the final consensus clustering, while it seems that this ground truth information is not provided to the single clustering methods, including RFclust. If that is the case than the comparative analysis between single and ensemble based clustering approaches is biased, as the ensemble cluster approaches have been provided with the final test data in the process of "learning" the model.

I would appreciate these aspects to be evaluated in order to verify that the performance improvements are not caused by artifacts to an invalid comparative analysis.

With respect to the writing, I would appreciate for the paper to be heavily revised. In particular, the flow of the paper is often interrupted. Introductions of the biological problem as well as the
methods used should be made in a consistent manner, not scattered throughout the text. Below, I give several examples:

67 / 68
- introduce SNPs / genetic markers at this point (not way later in the text)

81
- if necessary, this term should be explained

85
- introduced bootstrap aggregation, random subspace sampling (put into parenthesis in 88 - 90

87
- "objects" should be introduced / defined and distinguished from variables

89
- gini index, should be introduced later (with equation) as it is an essential part of the decision tree construction

98
- "independent sampling in each dimension" should be explained

127
- unclear

137 - 144
- This is one of the major motivation for the application of random forests for population structure analysis. It should be introduced much earlier, within the context of a more elaborate introduction of the biological problem. "linkage nature of genetic markers" - linkage should be introduced along with the biological problem
- stay with general mathematical introduction, the motivation should be given once in the beginning

- define genetic dataset (snps, gene marker)

- define Ci in \{d1,d2,d3\}

- the specific definition of what an object in this context refers to should have been made earlier (to general and vague at this point)

- Genetic dataset G was formerly defined as D (inconsistent)

- MDS - as it is used in RFclue, it should be explained

- algorithm, explain and show (equation) of gini index - central to random forest

- explanation into the beginningg
  - consistent writing - observations / oboecjts/ individuals

- "A genetic dataset usually contains sequences of genetic markers that represent variations among individuals. Single Nucleotide Polymorphisms (SNPs) are the most common type of genetic variation used to infer population structure. SNP occurs when a single nucleotide from
DNA sequence differs at the same position between individuals. A SNP has three categories: homozygous with the common allele (genotype AA), heterozygous (genotype AB), and homozygous with the rare allele (genotype BB)."

- This should have been introduced way in the beginning with the biological problem introduction.

271
- t not introduced
- the term objects is used again, even though it was replaced by a more specific biological problem oriented term.

397
- dramatic improvement is a rather vague term and not obvious - a statistical significance analysis might be a better way to show performance improvements

**Level of interest**

Please indicate how interesting you found the manuscript:

An article of limited interest

**Quality of written English**

Please indicate the quality of language in the manuscript:

Not suitable for publication unless extensively edited

**Declaration of competing interests**

Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
5. Do you have any other financial competing interests?
6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal

Do you want to get recognition for reviewing this manuscript?
Add a record of this review to Publons to track and showcase your reviewing expertise across the world’s journals. Signing up is quick, easy and free!

No