CSIC 5011 Mini-Project 1 Rebuttal

FAN Junyi XIAN Zhuozhi

Thanks to all reviewers for their comments. The following responses will address the comments of notable weaknesses mentioned.

Comment:

- lack of mathematical statement of algorithms.
- maybe explained more on Random forest?

Response: Due to the limit size of the poster, we omitted the detail of the algorithms. But random forest is a common algorithm for classification.

Comment:

- Maybe some discussion on why MDS performs worse than the two other data reduction methods. Analyze the classification with more evaluation methods like recall.
- But it would be better if author can explain why MDS result differs a lot in full data size and small data size.

Response: We only presented the result of MDS in two dimensions for embedding result to see if the SNPs data keep the graphic structure well. MDS is for Euclidean embedding visualization. Hence its result for classification performs worse than other methods. It's a good idea that we try other evaluation like recall or confusion matrix for better illustration.

Comment:

- The SNPs data is limited in reflecting the geographical information and only a very small part can decipher the region information for the samples. The future work should be related to the selection of the SNPs and the region information which is mapped to the SNPs.
- The further analysis is somewhat weak. For classification, only give analysis to PCA.
- The underlying meaning of the results of different dimension reduction method is not well illustrated. For example, when deciding to choose one method, which is the best?
- The accuracy decreases a lot after it reduces the number of SNPs. It can also compare the accuracy in which we do not use PCA and only use Random Forest to predict 1000 randomly selected SNPs.

Response: In future work, we can try different classification models like neural network to further explore the intrinsic structure and figure out how SNPs connect the region information. Meanwhile we can try classification on different dimension reduction methods or with/without dimension reduction to further explain the effect.

Comment:

- The application and the goal for data visualization and classification were not specified. (why is it important to classify people with different geographical variation)

Response: We try to use result of classifying geographical information as evaluation of the effect of different dimensional reduction methods. It's okay to use other information like region

or population as evaluation. We just choose geographical information since it only contains 7 distinct labels and other information have plenty more labels.

Comment:

- Could analysis more about why PCA has the best performance and t-SNE has the worst.
- The underlying meaning of the results of different dimension reduction method is not well illustrated. For example, when deciding to choose one method, which is the best?
- It might be better to point out why to choose these three out of plenty of data reduction methods, perhaps, they may give some explanations, like, PCA and MDS are linear dimensionality reduction techniques, meanwhile, t-SNE is a non-linear technique?

Response: Actually we choose PCA just for its better visual result and t-SNE actually looks it can well separate 7 regions. Numerically speaking, there are no better or worse between different methods and we need try different classification on different reduced data to see the classification result then we can tell which methods perform better. And we also try ISOMAP but we didn't give the result since it was not satisfying.

Comment:

- The label in the graphic result can enlarge a bit.

Response: Since the space of the poster is limited, all the figures are zoomed out, hence the labels look small.