Hyperparameter Optimization via K-fold Cross Validation in Seurat Clustering of scRNAseq data

Alex Ferrena – apf2139

Abstract

Motivation: Selection of Hyperparameters can greatly influence clustering and downstream analysis of single-cell RNAseq datasets. "Resolution" is a hyperparameter in Louvain Clustering as implemented in Seurat which controls the number of clusters. Selection of this is non-trivial and there are no methods to do so aside from general suggestions by the authors.

Results: A bootstrapping method inspired by K-fold Cross Validation and a metric for scoring clustering quality were developed to facilitate the selection of the resolution hyperparameter. This method was applied to a scRNAseq dataset generated from a lung adenocarcinoma primary mouse tumor.

Contact: apf2139@columbia.edu ; https://github.com/apf2139/clusterquality

1 Introduction

Single cell transcriptomic cDNA sequencing (scRNAseq) is a method for observing the molecular phenotype of single cells. A common goal in analysis of this type of data is the classification of cells into various cell types based on unsupervised clustering methods. One approach for this is the Seurat toolkit [1,2], which involves differentially expressed gene selection; dimensionality reduction, for example by principle component analysis (PCA); construction of a shared nearest neighbor graph (SNN); and classification of clusters based on density within the graph by Louvain clustering.

The selection of several key hyperparameters in this process has a strong impact on assignment of cells to clusters and associated downstream analysis, such as detection of differentially expressed marker genes and pathway analysis. These include the number of principle components to include in the construction of SNN graph; and critically, the selection of a tunable "Resolution" parameter in the Louvain clustering. The documentation of this parameter in the Seurat::FindClusters() function indicates that lower values produce fewer, larger clusters, while higher values produce more, smaller clusters. Testing via running a large range of PCs (1:5 – 1:50) and resolution parameters (0.1 - 3.0) [data not shown] revealed that the latter had a much greater effect on downstream analysis, such as cell type identification and marker detection. It should be noted that the selection of PCs used in graph construction is certainly non-trivial and testing over a range resulted in very distinct projections in t-distributed Stochastic Neighbor Embedding (tSNE) plots, along with different numbers of clusters, especially at the extremes. Nevertheless, the strongest impact on cluster number and cell assignment came from tuning resolution. Additionally, Seurat provides an iterative resampling method known as JackStraw to determine the optimal number of PCs to include, but no function is provided to optimize resolution parameter selection. Thus, hyperparameter optimization for the purpose of this project focused on resolution.

The default algorithmic parameters of Seurat's FindClusters() function (ie, with options *modularity.fxn* = 1, corresponding to the standard modularity function, defined below; and *algorithm* = 1, corresponding to the original Louvain algorithm introduced in [3], an algorithm based on community detection by modularity maximization). These options actually offer an extension upon the original Louvain algorithm described in [3] by including a tunable resolution parameter. This parameter is described well in Newman 2016 [4], equation 6:

$$Q(\gamma) = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \gamma\frac{k_i k_j}{2m}\right)\delta_{g_i g_j}.$$

where Q is the modularity score; m is the total number of edges in the network; Aij denotes whether there exist edges between nodes i and j; the value kikj / 2m denotes the expected number of edges between nodes i and j after graph randomization while respecting the degree, or number of incident edges, of i and j; δij denotes the Kronecker delta function with a value of 1 if i and j are in the same community and 0 if not; and γ denotes the resolution parameter. This parameter thus allows one to shift weights between the observed and expected edge values. A value of 1 is equivalent to the standard modularity function. This function has a well-characterized theoretical limitation, known as the "resolution limit", in that it cannot detect small communities in very large networks, providing motivation for introduction of the resolution parameter. Setting the value of γ lower than 1 gives more weight to observed edges and results in larger communities via recursive modularity maximization, as per [3]; setting it higher than 1 gives more weight to the observed edge term and results in fewer, smaller communities. Practically, the resolution parameter should increase as the network size / number of nodes increases to counteract the "resolution limit", but as described in [4], selection of γ is non-trivial. Therefore, this project set out to estimate an optimal parameter for γ in a dataset of single cell transcriptomes clustered in Seurat.

2 Methods

A resampling procedure inspired by the method of k-fold cross validation was devised to select the optimal resolution parameter for Seurat clustering of single cell transcriptome data. This method involves running the clustering algorithm iteratively, first on the whole dataset, and then on a

subsampled dataset after removing a subset of cells. The subset size is chosen by multiplying the number of cells in the dataset by 1/k. The dataset is then processed in the same way k times on the subset's inverse (ie after removing a distinct cell subset of this size) each pass. The value of k should ideally be as high as possible to reduce bias and variance in downstream analysis, but low enough that the computation time of the operation remains feasible. Since the Seurat authors recommend [5] a range of γ = 0.6 to 1.2 for a dataset of roughly 3k cells, and since the test dataset in this case was close to this number of cells, a value of k=10 was used for this range by increments of 0.1, while k=5 was additionally used for the range of γ=0.0 to 1.5. by increments of 0.1. (Edit – a recent update to the vignette [5] recommends 0.4 – 1.2, the lower end of this range is not reflected by 10-fold cross validation here).

Clustering of the whole dataset for a given value of γ is performed. This is referred to as the "reference" for a given value of γ. Next, the downsampling procedure as described above is performed. Processing of the dataset involves the standard steps described in the Seurat workflow [5]. This includes normalization; "scaling"; and the clustering procedure described above, including PCA, SNN graph construction, and Louvain clustering. The scaling step invoked by the function Seurat::ScaleData() is by far the longest step in this workflow, as it includes the optional but important step of performing regression to remove potential cofounders such as mitochondrial RNA content and the library size of the cell transcriptome, two key sources of noise in many scRNAseq datasets.

Once the processing pipeline has been performed for the reference and the k downsampled datasets, a scoring system is applied to choose the best value of γ. This metric involves detecting the matching pairs of clusters between the reference and each downsampled run; calculating the Jaccard index (intersection / union) of cells between each pair member; taking the mean of the Jaccard indices for each of the clusters in the downsampled run to combine the score of each cluster in a given downsampled run; then taking the "mean of means" between each downsampled run to get a score for the given value of γ. The value 1/k is added to the "mean of means", thus normalizing the score to the range 0 to 1. This metric is referred to as the "Jaccard Index Score".

It should be noted that this approach differs slightly than typical k-fold cross validation in that the paradigm of training vs test datasets is not utilized. For example, typical 5-fold cross validation would involve using four-fifths of a dataset to train a model and the final one-fifth of the dataset to test the model, using a different test subset each iteration for 5 iterations. This approach differs slightly as described above but the goal of testing models to select an optimal approach, here defined as the optimal γ value, remains the same.

3 Results

This algorithm was run on a scRNAseq dataset of a lung adenocarcinoma tumor isolated from a genetically engineered mouse model, along with associated non-cancer components, including immune cells and potentially stromal cells, endothelial cells, and surrounding epithelial cells. Running this algorithm on this dataset produced a reference dataset and 5 k-fold downsampled datasets for each value of γ from 0.0 to 1.2 by 0.1 ("5-fold"). Additionally, 10 downsampled datasets for the range 0.6 to 1.2 by a range of 0.1 were also generated ("10-fold"). The results are summarized in Figure 1. The Jaccard Index Score is plotted along with a 95% confidence interval error bar calculated using the variance of the vector of pairwise means against the t distribution. The t-distribution was chosen because the variance of all potential 1/k * cell number draws is unknown, and the number of samples is small (5 or 10).

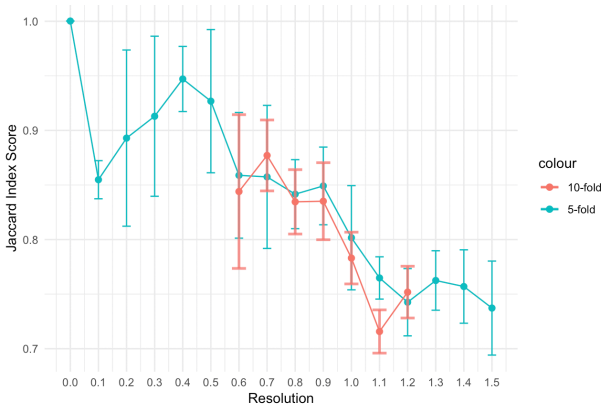

Fig. 1. Jaccard Index Scores for range of resolution (γ) parameters. 5 fold-cross validation was used for the range 0.0 to 1.2; 10-fold was used for range 0.6 – 1.2 as per Seurat Clustering Vignette.

The distribution of Jaccard index scores is generally higher for smaller values of γ and lower for greater values of γ. This is not unexpected, as the Jaccard index is a measure of similarity, and since lower γ values tend to produce fewer, larger clusters, cells are assigned to the same cluster pairing between reference and downsampled clustering more often.

This pattern holds true for the range between 0.6 – 1.2, where there is a decline from left to right corresponding with worse jaccard scores for increased resolution parameters. This reveals a key tradeoff in this approach, namely, exchanging "cluster robustness" for sensitivity in the form of identification of smaller populations with higher variance. The edge case of $\gamma = 0.0$ consistently produces a single cluster, wherein Jaccard score between reference and downsampled run is always equal to 1. The steep drop at $\gamma = 0.1$ reflects a jump from a single universally inclusive cluster directly to 4 clusters; that pass also appeared to show large differences in tSNE projection from case to case, likely as a result of the draw of the cell subset rather than any effect of γ. The maximum score aside from $\gamma = 0$ was at $\gamma = 0.4$, a pass in which the same number of clusters was detected in each downsampled run, and the tSNE projection did not appear to vary much, the first such case after $\gamma = 0.1$.

Furthermore, using 10-fold cross validation rather than 5-fold successfully tightened the error bars for some cases within the specified range, but not all. Additionally, for $\gamma = 1.1$, 10-fold cross validation produced markedly distinct results from 5-fold. This may simply be due to a particular draw of inverse-sampled cells in either run; however, the general trend of decreasing Jaccard Index Score remains apparent. In both 5 and 10-fold approaches, there is a steep drop in quality as assessed by the score between γ=0.9 and 1.0. To maximize both score and sensitivity, γ=0.9 would therefore be an appropriate choice in this circumstance.

Future directions of this method include updating the selection of the "pairing" of clusters between reference and each downsampled run. The pairing method is required because the downsampled run may produce a different number of clusters (usually fewer but occasionally more) than the reference. Currently, the pairing approach involves calculating for each reference cluster the jaccard index with each downsampled cluster and selecting the downsampled cluster with the maximum index; thus the reference cluster "chooses" its pair among the downsampled clsuters. This approach may be improved by the realization that the jaccard index can be used as an ad-hoc distance metric to produce a pairwise distancematrix, upon which a max matching algorithm may choose the optimal pairing.

Additionally, a critical second future direction for this method involves the development of a way to implement sensitivity into the quality score. While Jaccard Index Score provides insights to cluster robustness,

this alone may underestimate the "quality" of a particular clustering run; some potential methods might include a simple approach like somehow normalizing the score using the number of clusters detected; calculating the modularity score of the clustering; taking into account underlying graph qualities, such as community centroid distance; or implementing a metric of biological significance to the clustering, such as by comparing the differentially expressed markers between referenced and downsampled runs. The latter may be too computationally costly to practically implement in full, but the modularity score and centroid approaches are currently in active development as sensitivity metrics.

Acknowledgements

The author would like to thank Dr. Itshack Pe'er for very helpful conversation during this project and teaching assistant Raiyan Khan for kindly reading and grading through the semester. The author would also like to thank Dr. Jason Chan in the Tammela lab for mentorship and for generating the scRNAseq dataset and Dr. Tuomas Tammela for providing mentorship and access to resources with which to undertake this project, and both for kindly allowing usage of the dataset for this project.

References

1. Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nature Biotechnology (2018). nature.com/articles/doi:10.1038/nbt.4096

2. Comprehensive integration of single cell data
Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck III, Marlon Stoeckius, Peter Smibert, Rahul Satija
bioRxiv 460147; doi: https://doi.org/10.1101/460147

3. Fast unfolding of communities in large networks,
Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre,
Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)
doi: 10.1088/1742-5468/2008/10/P10008. ArXiv: http://arxiv.org/abs/0803.0476

4. M. Newman, Community detection in networks: Modularity optimization and maximum likelihood are equivalent, arXiv preprint
arXiv:1606.02319.

5. Seurat Clustering Documentation Vignette. https://satijalab.org/seurat/v2.4/pbmc3k_tutorial.html