Supervised Learning Approaches for Gut Cell Annotation

Margaret Y. Li
Paul G. Allen School of Computer Science and Engineering
University of Washington
Seattle, WA
marg33@cs.washington.edu

Abstract

Cell type annotation based on single-cell RNA-seq data is a critical step for further research on disease and medicine (Yang et al), and supervised learning methods have shown great promise in recent years towards this goal. Evaluating the performance of different methods on new datasets is important to see their abilities to generalize. In this paper, I apply various supervised learning methods for single cell annotation to gut cell data. I focus on gut cells because they are traditionally understudied – finding better ways to identify them would help lay the groundwork for future gut health research, which is more important than ever as the prevalence of gut diseases rises (Hills et al). I first review the previous literature on single cell annotation and on gut health research, to explain the motivation for my paper. Then, I discuss the details of my experimental setup and methodology, and review the results.

1 Background

1.1 Supervised cell typing

To investigate the roles and interactions of different cells, it is necessary to first identify the types of cells in scRNA-seq data by cell type annotation (Zhao et al, Ma et al). In recent years, computational approaches for cell type annotation have received much attention. One method is unsupervised cell clustering – grouping cells with similar profiles, and then labeling the clusters by matching them to canonical cell type markers (Zhao et al).

A relatively newer, but extremely promising method involves supervised learning to perform single cell label prediction. A systematic evaluation of different classifiers showed that certain ones such as random forest performed better than others (Zhao et al). As mentioned in Ma et al, supervised cell typing methods have become a popular research topic in recent years, as the amount of high-quality, annotated scRNA-seq data has grown rapidly. Supervised cell typing also has additional advantages over unsupervised methods: better performance on imbalanced datasets, better computational scaling, and less dependency on the size of the dataset. However, despite these advantages, the performance of supervised approaches can vary widely, subject to key factors such as model construction and dataset choice (Ma et al). This gives the first motivation for my paper: to evaluate how supervised cell typing performs across different models and different datasets.

1.2 Gut cell research

As discussed in Hills et al, gut diseases like metabolic diseases, gastrointestinal disorders, and cancers are increasing in incidence among Western societies, and it is more crucial than ever to advance gut health research to better address individual and societal health. Recent advances in single cell technologies have given rise to a better understanding of human biology at a cellular level, including the human gut. For instance, finding interactions between gastrointestinal microbiota and immune cells in the human colon gave insight into colonic immune niches (James et al). But as discussed in Elmentaite et al, intestinal tract physiology is highly complex across different regions and ages of development, and understanding the cellular biology is no easy task. In addition, gut cells are typically understudied in cell type annotation. As reported in Svensson et al, no gut tissues are among the top 10 most studied tissue types. The first gut tissue to appear is colonic tissue at 11th. Finding better ways to identify gut cells would help pave the way for future advancements in gut health research, diagnostic methods, and treatments.

To build a comprehensive map of the human gut, Elmentaite et al created the Space-Time Gut Cell Atlas, a catalogue of gut cells across space and time, "encompassing around 428,000 cells from the small and the large intestines as well as associated lymph nodes during in utero development, childhood and adulthood." I selected Space-Time Gut Cell Atlas datasets for my project because of their high quality, comprehensiveness, and cell-level detail.

2 Methods

2.1 Model selection

I train different supervised learning models for the single cell annotation task:

1. Random forest
2. SVM with linear kernel
3. SVM with radial basis function kernel
4. Naive Bayes
5. Logistic regression

As mentioned in Zhao et al, random forest (1) had good performance for single cell annotation. In addition, 1-3 are supervised learning models analyzed in Ma et al. I set out to find whether I would obtain comparable results as Ma et al on different datasets.

Naive Bayes (4) can be a surprisingly powerful classifier, despite the fact that its conditional independence assumption rarely holds on real world data (Zhang). According to Zhang, Naive Bayes works well even when there are strong dependencies if those dependencies cancel each other out. But that can be difficult to discern just by looking at a dataset, and so it still requires testing to see if a Naive Bayes classifier will perform well. Because of its efficiency and potential high effectiveness, I include Naive Bayes as another model to investigate for this task.

Logistic regression (5), like Naive Bayes, can be surprisingly effective sometimes, especially for classifying binary states. It remains a popular choice for modeling medical data (Boateng et al). While more modern ML techniques can perform better, just as for Naive Bayes, it can be highly effective and even outperform in the right situations.

2.2 Dataset selection

I use the following datasets of raw counts from the Space-Time Gut Cell Atlas:

1. Myeloid
2. B Cells

For each dataset, I set apart 20% of the data as test data, stratifying on the cell type labels, and train on the rest. I extract gene expression data (adata.X) for features, and predict on cell type annotation labels (adata.obs['annotation']). I use a random state of 0 for all train-test splits (and also for relevant model initializations) for consistency with all experiments.

For the random forest classifier, I use `sklearn.ensemble.RandomForestClassifier()` with 50 estimators, following the setup from Ma et al. For the SVM with linear kernel, I use `sklearn.svm.LinearSVC()`, and for the SVM with radial basis function kernel, I use `sklearn.svm.SVC()` with "rbf" kernel.

For the Naive Bayes classifier, I use `sklearn.naive_bayes.MultinomialNB()`, and for the logistic regression model, I use `sklearn.linear_model.LogisticRegression()` with 1000 max iterations, to ensure the model has enough iterations to converge.

3 Results

3.1 General accuracy

Accuracy of the models on the two datasets is shown below in Figure 1. The performance varied depending on the dataset chosen, and the model used. Logistic regression had the best performance on both datasets, achieving an accuracy of 0.87 on the myeloid cells and 0.9 on the B cells. SVM with linear kernel was a close second, with an accuracy of 0.87 on the myeloid cells and 0.88 on the B cells. For the classifiers that didn't perform as well, a divergence in performance began to occur across the two datasets: the SVM with radial basis function kernel had an accuracy of 0.79 on the myeloid cells but 0.69 for the B cells. Similarly, the Naive Bayes classifier achieved 0.79 accuracy on the myeloid cells but only 0.65 on the B cells.

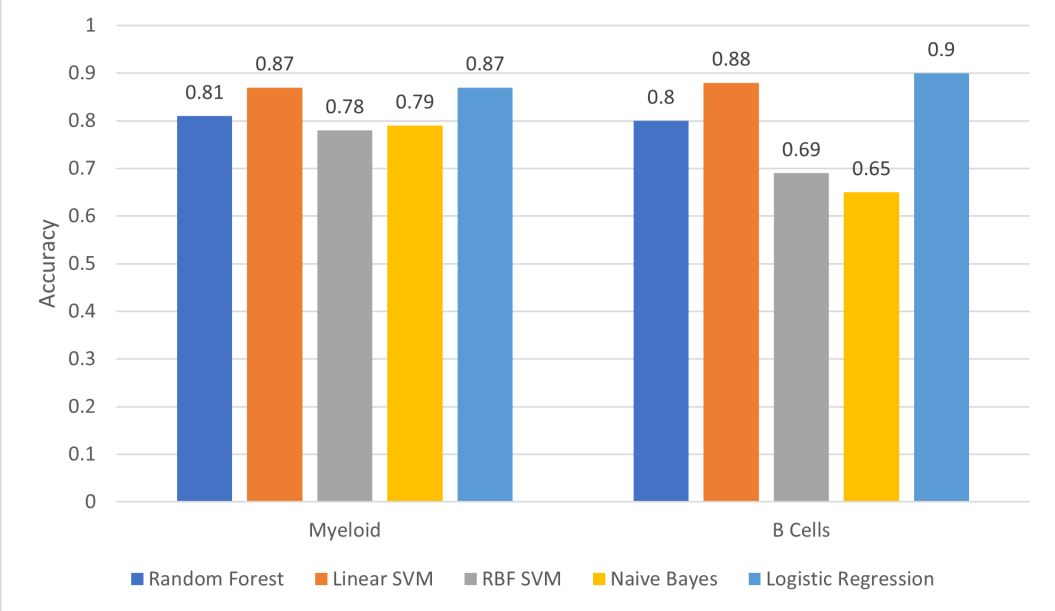

Figure 1: Accuracy of classifiers on myeloid and B Cells data

The high performance of the logistic regression model and SVM with linear kernel indicates that they fit well to the data, and that a linearity assumption worked well. Ma et al found that SVM with linear kernel generally had high performance for single cell annotation, and these results are consistent with that. Also consistent with their results, the random forest classifier did not have the strongest performance compared to others. However, while Ma et al found that the SVM with radial basis function kernel had mostly comparable performance to the SVM with linear kernel, it performed significantly worse in my experiments. But further tuning and parameter selection for models like

SVM with radial basis function kernel could help it achieve better results. As Ma et al also found, some models like the random forest classifier also had vastly different performance depending on dataset size or feature selection. So these results should not be taken to conclude that these models are always worse, but that more experimentation is needed to find the optimal strategies on different datasets.

3.2 Detailed performance results

Detailed classification reports for each dataset are shown in figure 2 and figure 3. Each classifier has a separate report recording its precision, recall, and f1-score on each cell type in the dataset. The number of examples (cells) for each cell type in the dataset's train set is also included as the first image, to help with comparison with the classification reports.

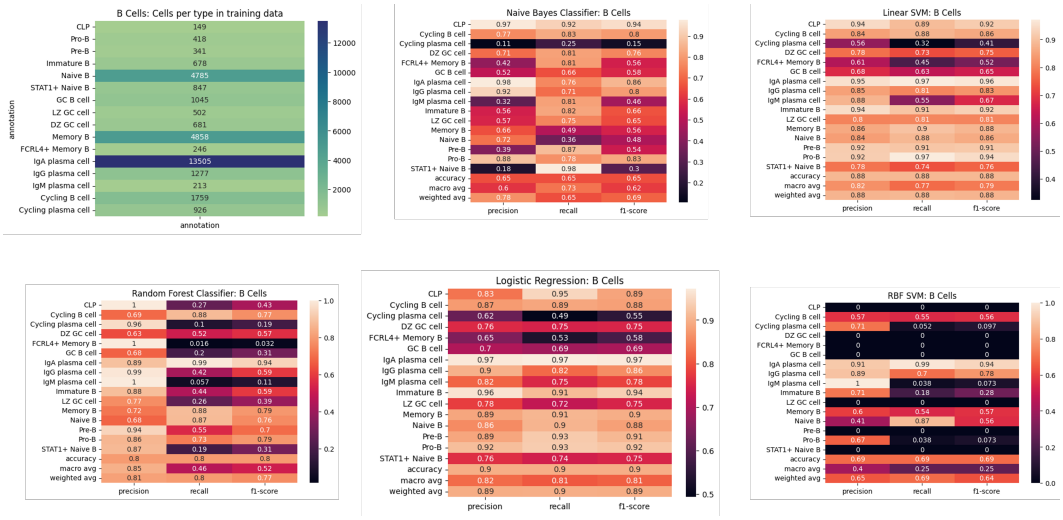

Figure 2: Detailed classification reports for B Cells

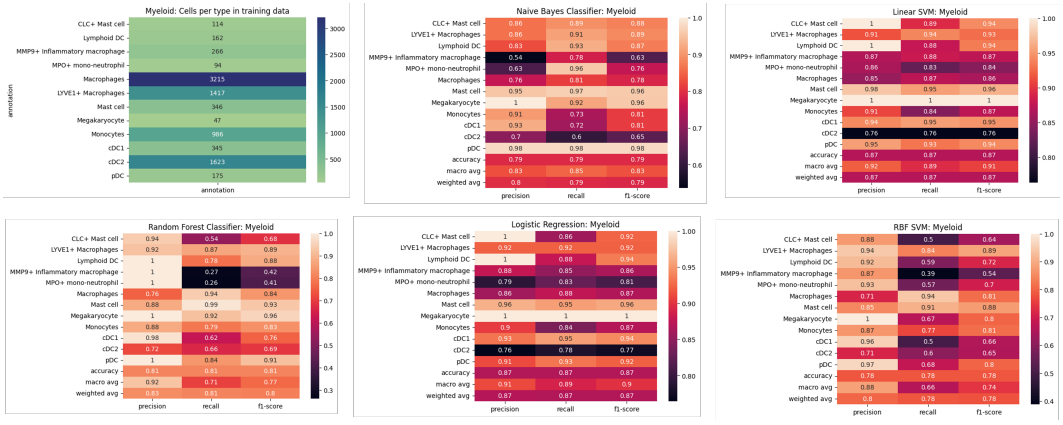

Figure 3: Detailed classification reports for myeloid cells

Some of the models had poor recall (and sometimes poor precision) on underrepresented cell types – such as the SVM with radial basis function kernel on the B cells dataset. This is an important thing to track when picking a model, to see if it is poorly predicting on certain cell types, even if it does well on other types. As seen by comparing the classification results and the number of examples per cell type in the training data, the cell types that the model got poor f1-scores on are correlated with the cell types with generally fewer examples. For example, with the SVM with radial basis function kernel, it had many low f1-scores across the board, but it had the highest f1-score of 0.94 on the Iga plasma cells – the most ubiquitous cell in the B cells training data by far, at 13505 counts.

The model still had decent performance on the next most ubiquitous cell types: Memory B (4858 counts, f1-score: 0.57), Naive B (4785, f1-score: 0.56), Cycling B cells (1759, f1-score: 0.56) and IgG plasma cells (1277, f1-score: 0.78). The performance drops significantly on the less populous cells after this, with an f1-score close to 0 on many of the extremely underrepresented cell types. Therefore, it's still important to consider how balanced datasets are as a factor in this process. Further testing is needed to find if the good performance on imbalanced datasets that certain models like logistic regression showed in these experiments will hold when applied to more datasets.

Another insight is that the models had poor performance on some specific cell types of a dataset, even the models that had generally great performance and achieved good f1-scores on most of the cell types. For example, none of the models got very high f1-scores for the cycling plasma cells (926 counts in train set), with the highest f1-score being 0.55 from the logistic regression model. For the myeloid dataset, while there was less variation in general, the highest f1 score that any model achieved on the cDC2 cells (1623 counts) was the logistic regression model at 0.77. While cycling plasma cells were relatively quite rare in the B cells data, cDC2 cells were not that rare in the myeloid data, being the second most ubiquitous type. So it's possible that having too few examples can make it difficult for models to learn how to identify that type. But models also might not learn to predict well on some overrepresented cell types. This could have many causes, such as that the features selected weren't ideal.

4 Conclusion

Supervised learning approaches for cell type annotation have deservedly received much attention in recent years. The abilities of supervised learning approaches to learn patterns in annotated single-cell RNA-seq data and then predict the types of cells in new data is extremely promising. I found that performance can vary greatly depending on the dataset and on the model choice, so picking a good model for the data being evaluated is important. As the field of gut health research advances, it is critical to view and understand the human gut at the most basic level: cells. Across the field of single cell transcriptomics, it's not just gut cells that have received less attention: many tissue and cell types are greatly understudied. Further work to evaluate the abilities of supervised learning methods on diverse tissue types is important to advance medicine across the whole of human biology.

References

[1] Hills RD Jr, Pontefract BA, Mishcon HR, Black CA, Sutton SC, Theberge CR. Gut Microbiome: Profound Implications for Diet and Disease. Nutrients. 2019 Jul 16;11(7):1613. doi: 10.3390/nu11071613. PMID: 31315227; PMCID: PMC6682904.

[2] James, K.R., Gomes, T., Elmentaite, R. et al. Distinct microbial and immune niches of the human colon. Nat Immunol 21, 343–353 (2020). https://doi.org/10.1038/s41590-020-0602-z

[3] Xinlei Zhao, Shuang Wu, Nan Fang, Xiao Sun, Jue Fan, Evaluation of single-cell classifiers for single-cell RNA sequencing data sets, Briefings in Bioinformatics, Volume 21, Issue 5, September 2020, Pages 1581–1595, https://doi.org/10.1093/bib/bbz096

[4] Ma, W., Su, K. and Wu, H. Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction. Genome Biol 22, 264 (2021). https://doi.org/10.1186/s13059-021-02480-2

[5] Yang, F., Wang, W., Wang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell 4, 852–866 (2022). https://doi.org/10.1038/s42256-022-00534-z

[6] Svensson V, da Veiga Beltrame E, Pachter L. A curated database reveals trends in single-cell transcriptomics. Database (Oxford). 2020 Nov 28;2020:baaa073. doi: 10.1093/database/baaa073. PMID: 33247933; PMCID: PMC7698659.

[7] Elmentaite, R., Kumasaka, N., Roberts, K. et al. Cells of the human intestinal tract mapped across space and time. Nature 597, 250–255 (2021). https://doi.org/10.1038/s41586-021-03852-1