Enhancing the Efficiency of Cell Classification in ScRNA-seq Data by Weakly Supervised Learning

Xinyi Zhou

Faculty of Mathematics, University of Waterloo, Waterloo, N2L 3G1, Canada
c32zhou@uwaterloo.ca

Abstract. Single-cell RNA sequencing (scRNA-seq) is a genomics technology that enables research of cellular diversity and operation through assessing gene expression. However, no widely available automated classification technology currently exists, and distinguishing cells within scRNA-seq datasets still relies on expert manual annotation. To address that, this study explores weakly supervised learning for automated cell type classification in scRNA-seq. The study applies lightweight machine learning techniques which have been trained with 10% labeled cells using the PBMC 3k dataset. The random forest models and logistic regression were trained on 30 major components and evaluated using their accuracy. Both models were able to classify major immune cell populations, with B cells achieving near-perfect classification. Random Forest performed better in separation of similar subtypes, such as CD4 and CD8 T cells, and exhibited alignment with the clusters in UMAP visualizations. These results indicate that weak supervision may provide efficient results of accurate and meaningful annotations and, therefore, lightweight nonlinear models can be an effective tool in scRNA-seq analysis.

Keywords: Machine Learning, ScRNA-seq, Cell Classification

1 Introduction

Since the single-cell RNA sequencing (scRNA-seq) enables the transcriptomic profiling of individual cells, the method has transformed the investigation of the heterogeneity of the cells entirely. ScRNA-seq can record differences between individual cells, whereas bulk RNA sequencing can only measure the average gene expression of a population. This technique has recently become a valuable one in modern biology, both to find out the complexity of tissue microenvironments, to trace the ancestry of cells, and to discover new cell types [1, 2].

The identification of the cell type is one of the most significant procedures of scRNA-seq data analysis. This process largely depends on manual annotation by the experts. This raises two problems. First, manual annotation has high labor costs and significant time costs. Second, manual annotation lacks standardized rules which leads to inconsistencies between each annotator. For example, one annotator labels only the broad category “T cell”, while another subdivides it into “CD4 T cell” and “CD8 T

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_62

cell". In the real-world single-cell data analysis, this costs researchers extra effort to ensure consistency of the annotations.

Weakly supervised learning provides a creative idea to address these challenges. It trains models by a small subset of annotated cells and uses other unlabeled data to enhance the learning process. This approach makes analyzing large datasets more efficient and effectively reduces the need for manual annotations [3].

This study investigates whether machine learning models such as random forest and logistic regression can be used to classify single-cell RNA-seq data when it only has weak supervision. By comparing the performance of the two models, this study evaluates whether simple and easy algorithms can provide a robust result on classification tasks with limited labeled samples. The contributions of this study are as follows:

1. The presented work suggests a cell classification framework achieved through machine learning techniques, which allows an accurate annotation of single-cell RNA-seq data with small amount of labeled cells and then significantly lowers the manual annotation effort and expenses.

2. This paper is a systematic assessment of lightweight machine-learning models (Logistic Regression and Random Forest) when there is a limited number of labeled samples, proving that they are useful and showing their strengths and weaknesses.

3. This paper makes comparisons between model predictions and the Louvain unsupervised clustering to understand consistency and describe the capacity of the models to separate the cells with similar traits.

2 Data Preparation

The single-cell RNA sequencing (scRNA-seq) dataset used in this study is from 10x Genomics. 10x Genomics is an industry standard for single-cell transcriptomics due to its ability to capture thousands of individual cells and accurately quantify gene expression at single-cell resolution. It has created a data format, utilized in storing scRNA-seq statistics, that is usable in the database. It also offers high quality and strong single cell profiles of gene expression. The data on this site is widely used in the field of single-cell research.

2.1 Dataset Overview

The PBMC 3k dataset has around 3,000 peripheral blood mononuclear cells (PBMCs) of healthy human donors. There are eight types of immune cells in PBMC 3k dataset, including T cells, B cells, natural killer (NK) cells, monocytes, dendritic cells, and megakaryocytes. This dataset is of high quality with balanced cellular composition. Being one of the most popular sample datasets, it is widely used as a unifying benchmark in comparison with clustering, annotation, and machine learning approaches in single cells analysis. Therefore, this study selects the PBMC 3k dataset for training and evaluation.

2.2 Quality Control

To guarantee data reliability, normal quality control (QC) steps were applied to filter out low-quality or damaged cells. Cells expressing less than 200 genes were filtered because these are usually empty droplets or anomalous data. To make sure that the dataset does not contain dying or stressed cells, cells with a mitochondrial gene expression proportion higher than 10% were also removed. These thresholds are consistent with the single cell best practices which is the optimally accepted pipeline in single-cell analysis pipelines [4].

2.3 Normalization and Feature Selection

The expression matrix was normalized so that the total counts per cell summed to 10^4 to mitigate sequencing depth differences between cells. To stabilize variance and get close to a normal distribution, the normalized counts were then log-transformed (log1p) [5]. Based on the extent of gene expression across different cell types, the top 2,000 highly variable genes were selected to highlight the most informative features. Principal Component Analysis (PCA) reduces computational complexity and captures the majority of expression variance and was used to reduce dimension while keeping the top 30 principal components.

2.4 Weakly Supervised Partitioning

The annotated dataset was randomly split into two groups at a ratio of 1:9. The annotations for the 90% group were removed and used as the prediction and analysis set. The annotations for the 10% group were retained and used as targets to train the model. This setting simulates the real-world scenario where only a portion of cells are annotated, enabling the study to investigate the real-world training conditions with limit labeling.

3 Methods

3.1 Modeling Approaches

Two lightweight machine-learning models, Random Forest (RF) and Logistic Regression (LR), were selected to study weakly supervised cell-type classification in scRNA-seq dataset. Logistic Regression provides a baseline as a linear model that estimates decision boundaries by a logistic function [6]. In contrast, Random Forest employs an ensemble of decision trees to learn non-linear patterns [7]. Therefore, it can eliminate the impact of noise in high-dimensional biological datasets. Under the assumptions in this study, the Random Forest should demonstrate superior performance on scRNA-seq datasets.

The dimension of gene expression matrices was reduced by Principal Component Analysis (PCA). The top 30 principal components were selected as input features. PCA is widely used in single-cell workflows to capture dominant variance while reducing noise and overfitting risks [5].

A performance matrix is generated by scikit-learn, including Macro-F1, Precision, Recall and F1 score. A cross-validation is also applied [8]. In this study, comparisons and evaluations between the two models use precision as the baseline.

3.2 Weakly Supervised Learning Procedure

This study simulated real-world scRNA-seq annotation environments by using only a small fraction of manually annotated cells. Specifically, the model was trained using only 10% of annotated cells and then applied to predict cell identities in the rest of the dataset. The PBMC 3K dataset is pre-annotated. The annotations are assumed to represent the true identities of the cells. The model's predicted identities will be compared against the existing annotations to evaluate its accuracy. To validate the predictive results in biological research, model outputs were cross-checked with Louvain clustering [9] and visualized using UMAP to examine spatial coherence and cluster separation [10]. This design reflects real-world scenarios of scRNA-seq annotation, while also providing graphical interpretations as reference for secondary verification by experts.

4 Experiments

This study used the PBMC 3k (Peripheral Blood Mononuclear Cells) single-cell RNA-seq dataset throughout the experiment. After the quality control process, the dataset contains approximately 2,638 cells and 1,838 expressed genes.

According to the data preprocessing workflow of Scanpy [11], the expression matrix was normalized, log-transformed, and reduced to 30 principal components by PCA. These components served as features for downstream classification.

The training set and dataset were randomly assigned. All random processes in this paper used random seed 42.

Two machine-learning models were implemented and evaluated using the scikit-learn Python package. The choice of hyperparameters for both methods is shown in Table 1 and Table 2. The model outputs both predicted cell types and predicted probabilities, which can be understood as the model's confidence level in its predictions.

Model performance is reflected in the performance report, which includes the precision of predictions for each cell type. Additionally, confusion matrices for both machine learning methods were generated using scikit-learn built-in functions to compare their results.

Finally, two sets of UMAP plots were used to visually display the prediction results and prediction probabilities for each method. These plots facilitate easier comparison of the graphical behavior of the prediction results.

Table 1. Hyperparameters for The Linear Regression Model.

Hyperparameter	Value
Max_iter	500
Class_weight	“balanced”
Multi_class	“multinomial”

solver	"lbfgs"
Random_state	42

Table 2. Hyperparameters for The Random Forest Model.

Hyperparameter	Value
N_estimators	300
Class_weight	"balanced_subsample"
Random_state	42
N_jobs	1

4.1 Experimental Results and Analysis

On the experimental results, the logistic regression model and the random forest model have been found to work well under weakly supervised conditions. The random forest, however, works better than the logistic regression in some cell populations where there are slight disparities in the transcription as well as in cell types that may be easily confounded.

In Table 3, both the models are observable to be highly accurate on the classification of key immune cell types, such as B cells, CD4 T cells, and CD14+ monocytes (0.95 for logistic regression and 0.93 for random forest). This is because of their different gene expression profiles. This fact is also confirmed by the UMAP plot in Figure 1, which indicates that the cluster of B cells is distinctly distinguished among other clusters. This observation is consistent with findings that B cells exhibit clear transcriptional profiles in PBMC datasets [12].

Table 3. The model precision of logistic regression and the random forest on different types of cells.

	Logistic Regression	Random Forest
B cells	0.99	0.99
CD14+ Monocytes	0.96	0.93
CD4 T cells	0.95	0.94
CD8 T cells	0.79	0.92
Dendritic cells	0.91	1.00
FCGR3A+ Monocytes	0.91	0.96
NK cells	0.91	0.98
Average accuracy	0.92	0.96

In contrast, for cells with minimal genetic differences such as CD4 and CD8 T cells or CD14+ and FCGR3A+ monocytes, the two models performed less effectively. Logistic regression had a relatively low accuracy (0.79) in predicting CD8 T cells, whereas random forests maintained a comparatively high level of accuracy (0.92). The significance of modeling nonlinear relationships in scRNA-seq data is highlighted by this comparison. The random forest improves the ability to distinguish between immune subsets that are easily confused by capturing complex feature relationships.

Megakaryocytes constitute a minority group with only 3 cells, whose predictive results are unstable and are therefore excluded from the accuracy comparison between the two models.

The UMAP visualizations also highlight the performance differences between the two models. Comparing Figure 2 and Figure 3 with Figure 1, both approaches accurately recovered major immune cell populations, and consistent with the prior result. B cells form a highly distinct cluster, making them the easiest population to classify. In contrast, CD4 and CD8 T cells exhibit partially overlapping transcriptional states, leading to less clearly defined boundaries in the embedding and consequently lower accuracy for these T-cell subsets, particularly for logistic regression. Logistic regression predictions show greater mixing between similar cell types, reflecting the limitation of linear decision boundaries in separating closely related immune populations. Interestingly, when examining prediction-probability maps in Figure 2, the logistic regression model appears more confident in its predictions, giving higher average probability scores even in regions where group boundaries are indistinct. This suggests that the linear model may overestimate its certainty when faced with ambiguous transcriptional signatures. In contrast, the random forest model can distinguish T cell subsets more clearly. As shown in Figure 3, for the monocyte portion that is difficult to differentiate, the random forest also provides lower certainty scores for researchers to reference and refine. Random forest predictions also better resolved the rare cell populations like the dendritic cell population and the NK cells.

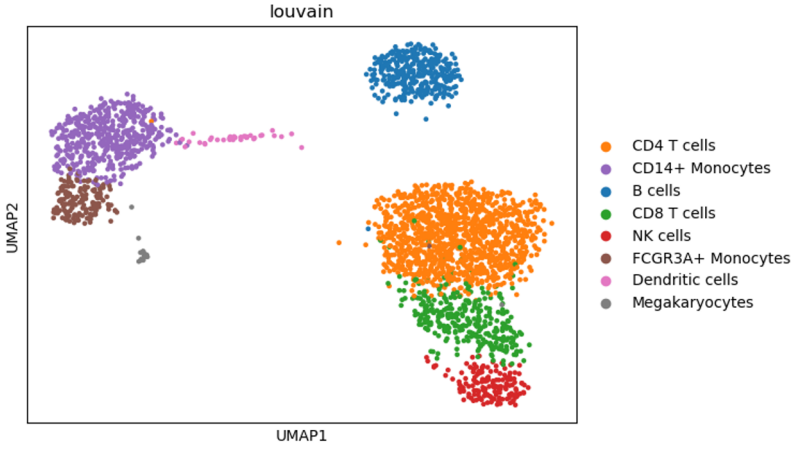

Fig. 1. Louvain cluster map with human-annotated labels (Picture credit: Original)

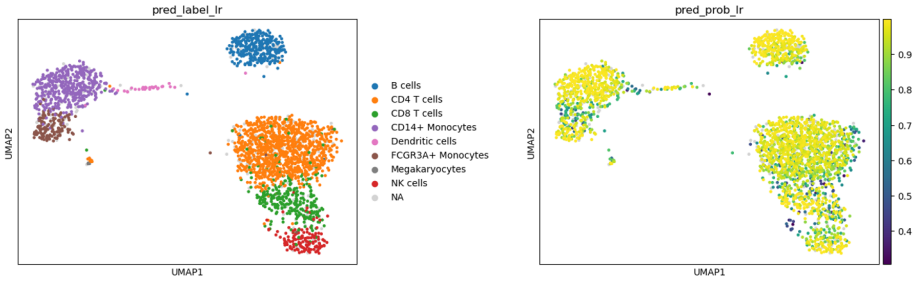

Fig. 2. UMAP plots of linear regression model predictions and prediction probabilities (Picture credit: Original)

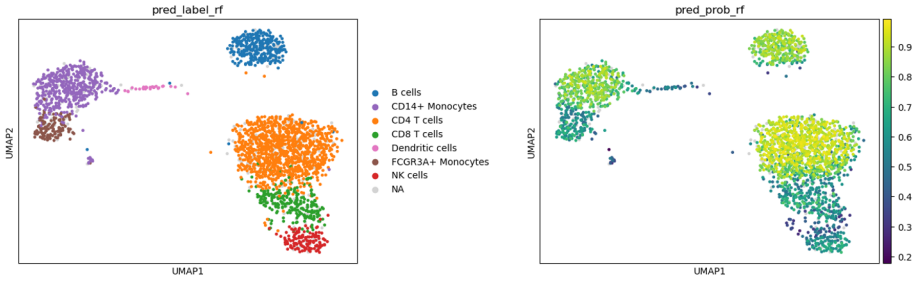

Fig. 3. UMAP plots of random forest model predictions and prediction probabilities (Picture credit: Original)

Figure 4 demonstrates that logistic regression exhibits significant confusion between CD4 and CD8 T cells. The confusion matrix also reveals that CD8 T cells and NK cells are prone to misclassification. Compared with logistic regression, the random forest model achieves relatively accurate differentiation among these cell types. These findings further validate that nonlinear methods can enhance the resolution of subtypes in immune cell classification.

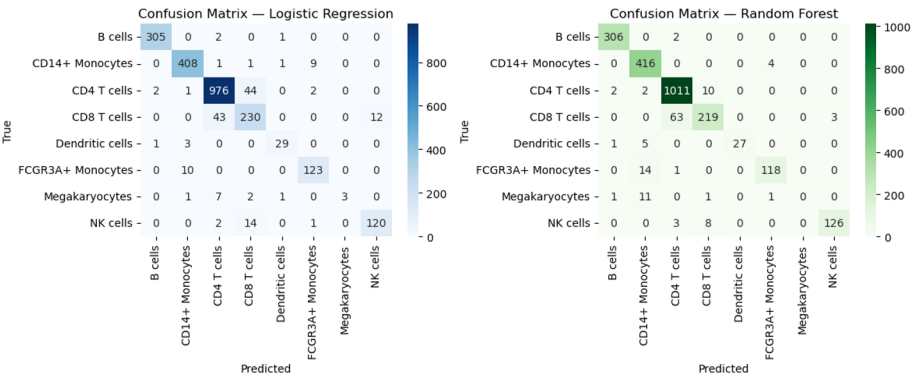

Fig. 4. Confusion Matrices for linear regression and random forest (Picture credit: Original)

Overall, the presented results indicate that weakly supervised approaches could be effectively used to predict cell types in scRNA-seq matrices given that few annotated samples are provided. Of these methods, random forest is one of those that capture higher granularity of cellular heterogeneity, and their predictive accuracy is high as well as offers sound prediction scores. These scores and results of predictions are important references that can be used by the researchers to carry out future studies.

5 Conclusion

This study addresses the challenges in cell type annotation for single-cell RNA sequencing (scRNA-seq) by developing a simple weakly supervised method of classification and demonstrating its effectiveness in the PBMC 3k dataset. Both approaches can identify seven major populations of immune cells with limited annotation set. Logistic regression was found to work well with populations that can be identified easily, whereas the random forest used nonlinear associations to improve the rate of classification especially with confusable immune subtypes.

Although cross-dataset predictions were not performed in this study, models trained on gene expression can be generalized to other datasets within the same batch for accurate predictions. Nonetheless, cell types in various batches are not necessarily uniform. In such cases, models will require additional annotated data for training and incorporation of supplementary information, such as gene markers, which contradicts this study's exploration of lightweight and weakly supervised models.

In the aim of assisting the individual researchers in eliminating repetitive tasks and improving annotation efficiency, further studies may explore semi-supervised and graph-based neural networks to enhance generalization capabilities. Simultaneously, preprocessing steps and the machine learning-based automatic annotation process can be integrated into an AI workflow to make it truly automated.

Overall, this study provides evidence on the feasibility of weakly supervised learning as a tool for scRNA-seq data annotations. The proposed methods in the paper offer effective and time-saving alternatives to manual annotations by experts. These findings demonstrate the promising prospects of developing machine learning methods in single cell analysis to assist researchers in more efficient annotating, analyzing, and contributing to clinical medicine.

References

1. Tang, F., Barbacioru, C., Wang, Y. et al.: mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6, 377–382 (2009)
2. Svensson, V., Vento-Tormo, R. & Teichmann, S.: Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 13, 599–604 (2018)
3. Zhou, Z.: A brief introduction to weakly supervised learning, National Science Review 5(1), 44–53 (2018)

4. Luecken, M. D., & Theis, F. J.: Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology, 15(6), e8746(2019)
5. Germain, PL., Sonrel, A. & Robinson, M.D.: pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. Genome Biol 21, 227 (2020)
6. Hosmer Jr, David W., Lemeshow, S., and Rodney X. Sturdivant.: Applied logistic regression. John Wiley & Sons, 2013
7. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
8. Kohavi, Ron.: A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai 14(2), (1995)
9. Combe, D., Largeron, C., Géry, M., Egyed-Zsigmond, E.: I-Louvain: An Attributed Graph Clustering Method. International symposium on intelligent data analysis. Cham: Springer International Publishing, 2015
10. Cottrell, S., Hozumi, Y., Wei, G.: K-nearest-neighbors induced topological PCA for single cell RNA-sequence data analysis. Computers in Biology and Medicine 175, (2024)
11. Wolf, F., Angerer, P. & Theis, F.: SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19, 15 (2018)
12. Zhao, Y., Cai, H., Zhang, Z. et al.: Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. Nat Commun 12, 5261 (2021)