Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220532

Single Cell Type Prediction from Gene Profiles - An Overview of Different Computational Methods

Jiaxun Li[1*], Xinyu Liu[2] and Kaijie Xu[3]

[1]Department of Electrical Engineering & Computer Sciences (EECS), University of California, Berkeley, Berkeley 94720, California, United States
[2]Steinhardt School of Culture, Education and Development, New York University, New York, 10003, United States
[3]College of Letters & Science(L&S), University of California, Berkeley, Berkeley 94703, California, United States

jiaxun1218@berkeley.edu

Abstract. Proven to be useful in quantitative analysis of mRNA, scRNA-seq measures the individual gene expression profile and helps with rare cell population identification. Successful scRNA-seq analysis would be useful to boost knowledge of cancer cells and tumorigenesis, thus improving the ability to identify biomarkers and detect individuals' disease susceptibility. This work conducted dimensionality reduction using naive principal component analysis. Then, several classification algorithms, including support vector machine, random forest, boosting, and neural networks, were examined with best hyperparameters determined by grid search. With the comparison of data dimensionality N=83 and N=127, each method generated the prediction accuracy of the dataset, with the support vector machine achieving the highest testing accuracy of 53.52%. The relatively high prediction accuracy enables better characterization of single gene expression profiles due to support vector machine's ability to regularize high-dimensional data. Deeper architectures and usage of Bayesian optimization may further encourage efficient analysis of larger datasets with better classification accuracy.

Keywords: scRNA-seq, Dimensionality Reduction, Neural Network, Support-vector Machine, Ensemble Learning

1. Introduction

Single cell RNA sequencing (scRNA-Seq) has been developed in recent years, marking great maturation of single-cell transcriptomics. Proven to be useful in detection and quantitative analysis of mRNA, scRNA-seq measures the individual gene expression profile and helps with rare cell population identification, for example, the identification of a malignant tumor cell [1]. scRNA-seq is being used in tracing heterogeneous cellular states to find cell linkage and possible developmental relationships in cell differentiation [2]. Another application of scRNA-seq is the ability to recognize fundamental characteristics of gene expression profile, thus studying gene-regulatory networks that imply similarities and heterogeneity between cells [3].

ScRNA-seq boosts the development of bioinformatics data analysis. After the first scRNA-seq method was developed [4], further refinement regarding scRNA-seq quality control [5], scRNA-seq data

DOI: 10.54254/2755-2721/2/20220532

normalization [6], imputation algorithm of scRNA seq data [7] and techniques of dimensionality reduction were proven useful when analyzing scRNA-seq data. The successful analysis of scRNA-seq data helps with accurate cell type identification, characterization of hidden cell population, and identification of possible malignant cells that affect individuals' disease susceptibility. As cancer is the most heterogeneous of all diseases, scRNA-seq analysis provides a way to dissect human tumor tissue at single-cell resolution and determine cell composition [8], thus providing effective tumor diagnosis.

Some previous studies have developed architectures that lead to successful analysis of scRNA-seq data and encouraged better classification of cells. Lin, in the paper "Using neural networks for reducing the dimensions of single-cell RNA-Seq data" [9], chose prior biological data, like protein-protein interactions and protein-DNA interactions as neural network architectures and used denoising autoencoders to do data pre-training. The study presented several neural network analyses that were useful to gene profile annotation and cell type-specific identification [9]. Another study conducted by Xu proposed a multi-scale clustering-based feature selection method for gene expression data [10]. The feature selection method captured informative genes among tumor populations and visualized the scRNA-seq data [10]. The algorithm was applied to lung adenocarcinoma data in the study, but further study may apply the framework to other genomic data to identify robust biomarkers and increase the understanding of tumorigenesis.

For better classification of multi-scale scRNA-seq data, this study first conducts dimensionality reduction using principal component analysis (PCA), then examine several common machine learning algorithms, including Random Forest, Support Vector Machine, Boosting, and Neural Networks to build the classifier of highest accuracy of predicting and grouping cell type when analyzing a database of great amounts of scRNA-seq samples.

2. Methods

The datasets used are from 104 separate scRNA-seq experiments that collect single cell profiles from different individuals, with various focus on cell types. A training set and a test set are provided, and there's no overlap in the set of experiments used.

2.1. PCA

Several dimension reduction techniques are explored, such as Principal Component Analysis (PCA), kernel PCA and Linear Discriminant Analysis (LDA). Specifically, PCA is used to identify the features with the highest variance, and LDA aims to maximize the separation of existing features. PCA is the most common dimension reduction method, but it needs the principal components to have linear structures and to be orthogonal. As a side note, this study also makes the utilization of kernel PCA to capture the nonlinear structure, but it leads to worse results. In other words, naive PCA is enough to capture the hidden data structure. To reach a variance of 0.99, the study chooses 127 as the number of the components (83 components to reach a variance of 0.95).

2.2. SVM

After the dimension reduction by PCA, the data is used to train the classifiers. Since the data is not linearly separable, we used Support Vector Machine for classification and the Gaussian radial basis function as the kernel function. And the penalty factors and scale factors are optimized by Grid Search with 10-fold cross-validation.

2.3. Random forest

As an alternative strategy, we combined the data after dimensionality reduction with a Random Forest classifier, given its acknowledged good performance on high dimensional data. To achieve the highest accuracy, this study used grid search for 50 points on multiple hyperparameters. The optimal parameters are max_depth = 7 and random_state = 3.

DOI: 10.54254/2755-2721/2/20220532

2.4. Boosting

We also built a multi-class Adaboost decision tree to see if boosting can improve prediction accuracy on this multi-class gene expression profile data. This study used the data after dimensionality reduction with a decision-tree based adaptive boosting classifier. To achieve the highest accuracy, we used k-fold cross validation to determine the max depth of 23 as the decision tree classifier. As the PCA analysis inferred, the study tried both data dimensionality N=83 and N=127. In addition, when examining adaptive boosting, the study used both discrete and real boosting algorithms to compare the results. Generally, discrete SAMME Adaboost adapts based on predicted class labels, and real SAMME.R Adaboost adapts based on predicted class probabilities.

2.5. Neural network

In addition to those traditional methods, guided by the previous work by Lin et al. [9], we also tried another technique which uses neural networks as the classifier. As illustrated by **Figure 1**, our neural network is fully connected and consists of only one hidden layer with 796 nodes, which is based on the architecture explored by previous research. The input layer takes in the features of a cell after dimensionality reduction (we found out that N=83 works better than more components), while the output layer outputs a vector of length 46 (the number of classes / labels in the dataset), which is an unnormalized probability distribution of the correct cell type the study wants to predict, and then use a SoftMax function to choose the cell type with highest probability.

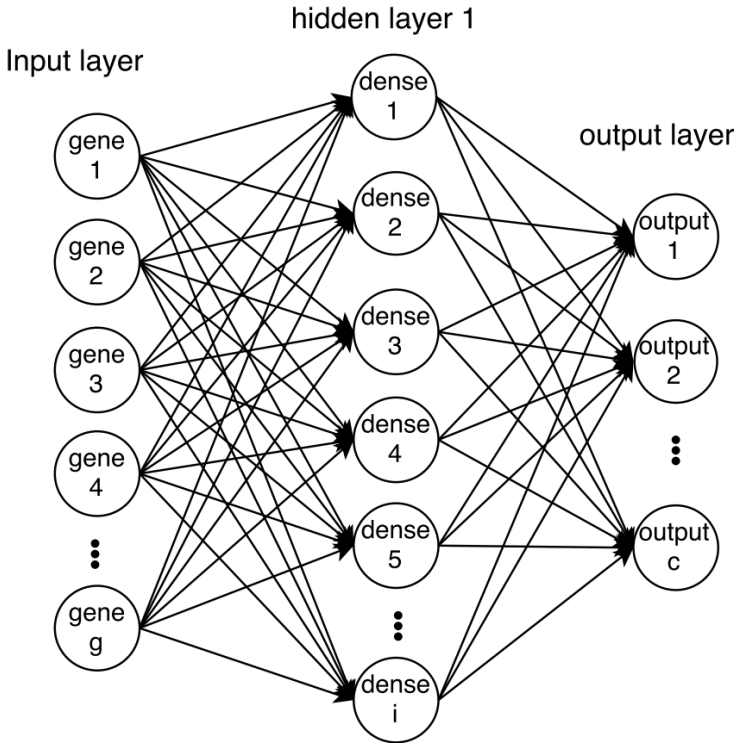

Figure 1. NN Architecture

To obtain the best performance, this study manually tuned the hyperparameters with references to recommendations mentioned in previous work. Without further optimizations, the best parameters found are listed here:

learning_rate=0.02, momentum=0.1, decay=0.0001, batch_size = 10, epochs = 70, validation_split=0.3, activation function for hidden layer: tanh(x)

Packages used: Sklearn, Pandas, Tensorflow.Keras, NumPy, Matplotlib, etc.

DOI: 10.54254/2755-2721/2/20220532

3. Results

Table 1. Testing accuracy summary of different methods (unoptimized)

Dimensionality Reduction	Number of Components	Classifier	Accuracy
None	N/A	SVM	49.22%
None	N/A	GNB	42.66%
PCA	127	SVM	53.52%
PCA	83	SVM	50.30%
SVD	127	SVM	22.45%
LDA	127	SVM	28.20%
PCA	127	GNB	26.72%
PCA	127	SGD	22.38%
PCA	83	RF	45.57%
PCA	150	Naive MLP	36.67%
PCA	83	MLP	38.95%
PCA	127	KNN	45.45%
PCA	127	NN	42.63%
PCA	127	AdaBoost	43.57%

The results of important methods are discussed below.

3.1. SVM

The parameters of the RBF kernel SVM are chosen from: $\log10(C) \in [-3,3]$ $\log10(\gamma) \in [-3,3]$.

By Grid Search Cross Validation, parameters of $C = 0.1$ and $\gamma = 0.01$ are successful. The training accuracy quickly converges to 90.57%. And the final test accuracy of RBF SVM is 53.52% for $N = 127$(50.3% for $N = 83$).

3.2. Random forest

The results found is that using N=83 after dimensionality reduction worked better than other numbers of components, with a testing accuracy of 45.57%.

DOI: 10.54254/2755-2721/2/20220532

3.3. Boosting

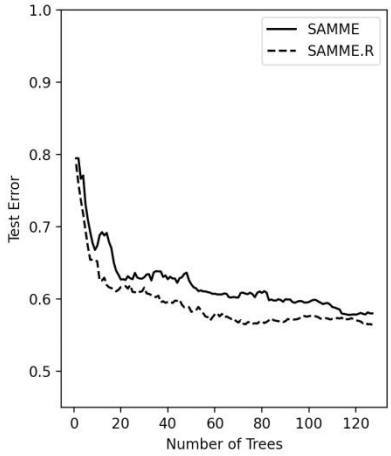

Figure 2. Test error development (N=127)

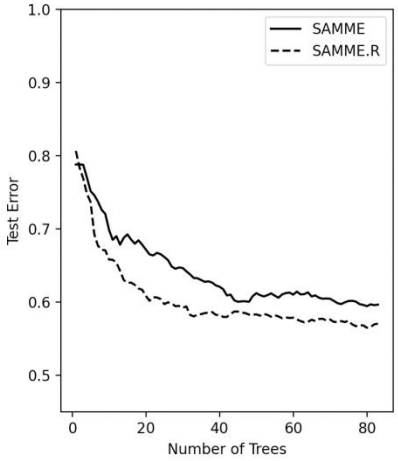

Figure 3. Test error development (N=83)

For N=127, After all boosting iteration, the test accuracy of the real boosting algorithm was 43.57%, and the test accuracy of the discrete boosting algorithm was 41.99% (**Figure 2**). For N=83, After all boosting iteration, the test accuracy of the real boosting algorithm was 42.94%, and the test accuracy of the discrete boosting algorithm was 40.35% (**Figure 3**). For both N=127 and N=83, the real boosting algorithm achieved a better prediction accuracy than the discrete boosting algorithm. N=127 worked better than N=83 for the adaptive boosting.

3.4. Neural network

When choosing the optimal hyperparameters, the training accuracy can reach above 85% within 70 epochs (**Figure 4**). And when the study uses the trained model to predict the cell types in the test dataset, the accuracy is 42.63%. For the data dimension, N=83 gives a 5% higher accuracy than N=127.

DOI: 10.54254/2755-2721/2/20220532

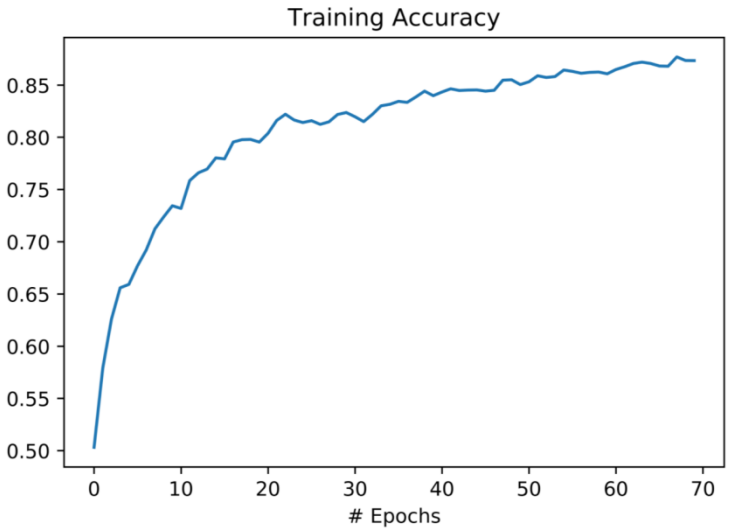

Figure 4. Training accuracy of the NN method.

4. Conclusion

While scRNA-seq analysis was seen as useful in single gene expression profile, there's more to work on. As biological data generally contains many cells with high dimensionality, new methods need to be developed in order to predict and group the single cell data for further study efficiently and accurately. In this paper, the set of possible cell types in test data is a subset of the training data to ensure the classifier predicts on gene expression profiles. After conducting principal component analysis and determining the suitable data dimensionality, we developed and tested solutions based on several base-line methods and combining methods, finding that the support vector machine worked best of all algorithms being tested, having the test accuracy of 53.52%.

Possible reasons for support vector machine's outstanding performance is due to its ability to regularize high-dimensional data, while our dataset is of high dimensionality with tens of thousands of gene profiles. Using SVM for data after dimensionality reduction wouldn't lead to overfitting. Random Forest is also a promising method for high dimensional data, and it's the second best method after tuning hyperparameters with grid search. Both SVM and Random Forests have the potential for further improvement (e.g., use Bayesian Optimization to learn the best parameters). The Neural Network method turned out to be a bit less effective compared to the previous work which used similar architecture, and such difference in accuracies might be caused by unoptimized parameters and different ways of calculating testing accuracy. The performances of other methods we attempted also matched our expectations.

In addition to the testing accuracy, it's also important to see which labels are most commonly misclassified by each method. To this end, we created the following confusion matrices, where the squares with brighter colors indicate larger numbers of hits (ideally the diagonal squares should be the only ones with bright colors). As illustrated by Figure 5-8, the study found that the most common errors are relatively consistent across different methods (in each figure, the y-axis is the true label, and the x-axis is the predicted label). Several commonly misclassified cell types across all methods include hematopoietic stem cell, liver, medullary thymic epithelial cell, and embryonic stem cell; and it turned out that these cells are not closely related biologically, and the reasons behind remains to be discovered.

DOI: 10.54254/2755-2721/2/20220532

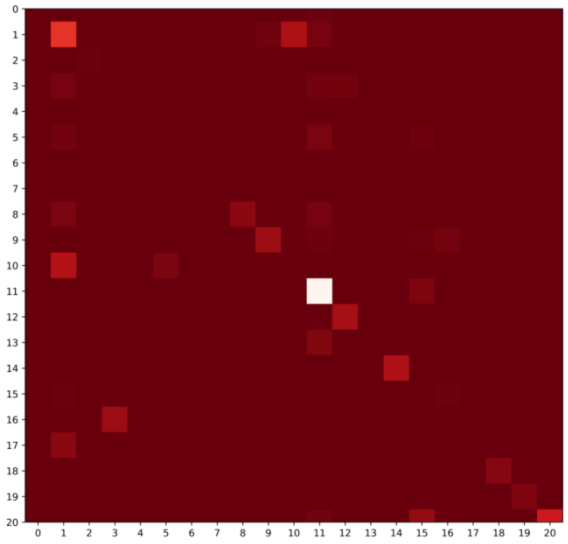

Figure 5. Confusion Matrix (SVM).

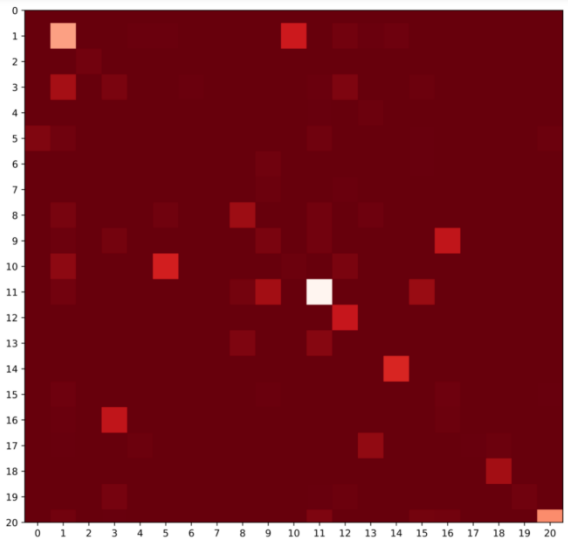

Figure 6. Confusion Matrix (NN).

DOI: 10.54254/2755-2721/2/20220532

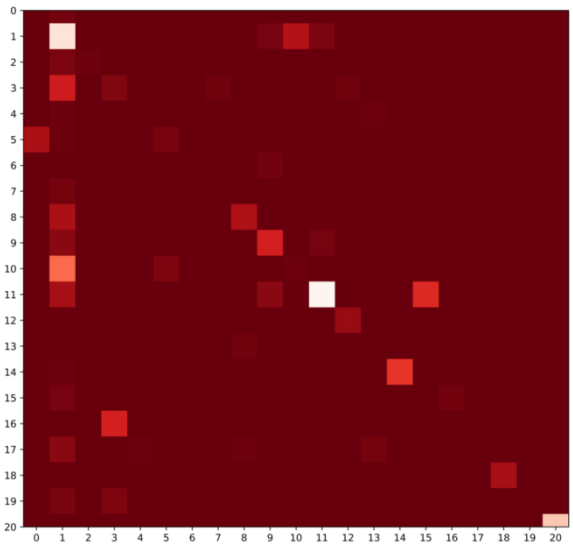

Figure 7. Confusion Matrix (Boosting).

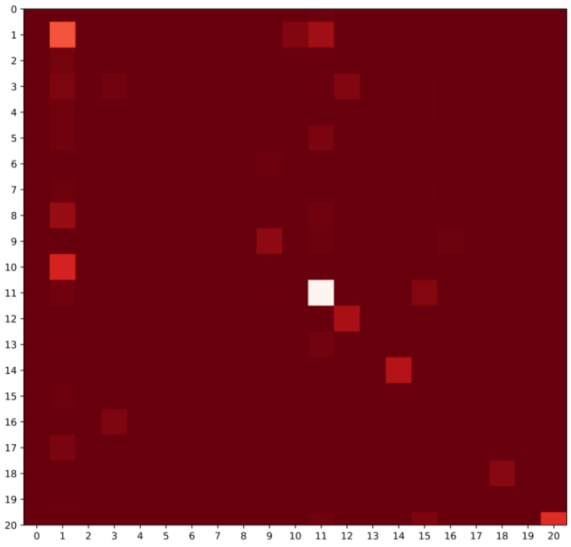

Figure 8. Confusion Matrix (Random Forest).

Although the resulting accuracy is encouraging, there are more that can be explored. New methods, including the extreme gradient boosting for feature selection and the usage of Bayesian optimization on hyperparameters, may be implemented to find the balance between bias and variance. The study experienced misclassified labels that didn't share sufficient similarities, so future research can train a separate method to identify commonalities between cell types, and thus consider type similarities into the definition of "accuracy". The study may test deeper architectures (neural works with more layers) and combine support vector machines with adaptive boosting for efficient implementation of algorithms. Future research with improved algorithms would allow the efficient analysis of larger datasets.

References

[1] Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352:189–96.

DOI: 10.54254/2755-2721/2/20220532

[2] Haque, A., Engel, J., Teichmann, S.A. et al. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med 9, 75 (2017). https://doi.org/10.1186/s13073-017-0467-4

[3] Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016;34:1145–60.

[4] Tang, F., Barbacioru, C., Wang, Y. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6, 377–382 (2009). https://doi.org/10.1038/nmeth.1315

[5] Jiang, Peng et al. "Quality control of single-cell RNA-seq by SinQC." Bioinformatics (Oxford, England) vol. 32,16 (2016): 2514-6. doi:10.1093/bioinformatics/btw176

[6] Bacher, R., Chu, LF., Leng, N. et al. SCnorm: robust normalization of single-cell RNA-seq data. Nat Methods 14, 584–586 (2017). https://doi.org/10.1038/nmeth.4263

[7] Huang, M., Wang, J., Torre, E. et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods 15, 539–542 (2018). https://doi.org/10.1038/s41592-018-0033-z

[8] Sun, Guangshun et al. "Single-cell RNA sequencing in cancer: Applications, advances, and emerging challenges." Molecular therapy oncolytics vol. 21 183-206. 8 May. 2021, doi:10.1016/j.omto.2021.04.001

[9] Chieh Lin, Siddhartha Jain, Hannah Kim, Ziv Bar-Joseph, Using neural networks for reducing the dimensions of single-cell RNA-Seq data, Nucleic Acids Research, Volume 45, Issue 17, 29 September 2017, Page e156, https://doi.org/10.1093/nar/gkx681

[10] Xu, D., Zhang, J., Xu, H. et al. Multi-scale supervised clustering-based feature selection for tumor classification and identification of biomarkers and targets on genomic data. BMC Genomics 21, 650 (2020). https://doi.org/10.1186/s12864-020-07038-3