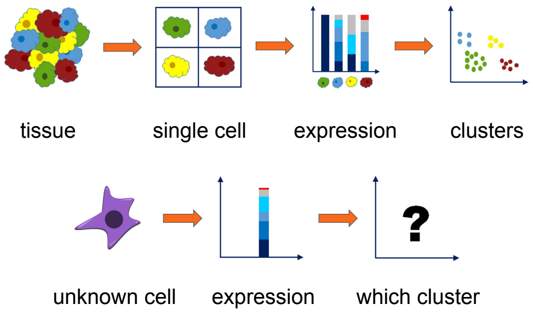
Single Cell Analysis Using Dimensional Reduction and SVM

Derun Gu, Zhengyu Chen, Zihao He
10701 final project

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Introduction



tissue → single cell → expression → clusters

unknown cell → expression → which cluster

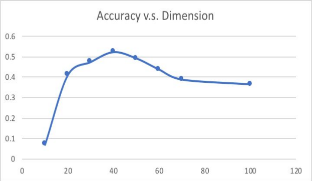
Genes contain information of instructing how to build a molecule, and different cells in our body express different subsets of these genes at different levels. A technique named single-cell RNA sequencing (***scRNA-seq***) can be used to measure the activity levels of some annotated genes in a single cell, the result of which is termed gene expression profile. Same type of cells (e.g. lung cell, brain cell, skin epidermis) are closely related with its gene expression profile.

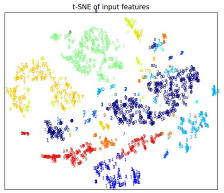
The goal of the project is to classify the type of an unknown cell solely based on its gene expression profile. Specifically, some experiments are used for training and a cell from another set of experiments never seen before is tested. The model itself must possess balanced abilities of learning and generalization.

Methodology

Data Preprocessing

We apply Principal Component Analysis (PCA) on the training data first and record the first 40 principal component. Then we project both training data and the testing data on those 40 principle component and reduce the dimensionality from originally 20499 to 40. We also tried other reduced dimension. However, in experiments we found that when the dimension is small, we can lose some important information, and when the dimension is large, some unrelated information encoded by minor principal components can mislead the classification. By tuning carefully, we choose 40 as our final choice. The trend is shown in the left bottom graph. We have also tried LLE, LDA, auto-encoder for dimensionality reduction, did not observe a significant improvement.





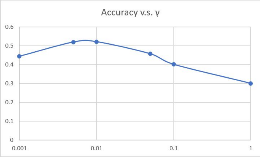
T-SNE 2D representing of the first 10 classes after PCA are plotted on the right, indicating the PCA maintains enough information for further classification.

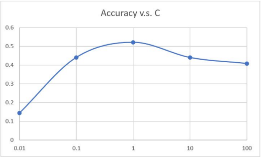
Random Forest

We have trained a random forest model and achieved 46.02% accuracy. Due to high fitting ability of decision tree, we applied early stop and path pruning techniques to improve generalization ability.

Support Vector Machine

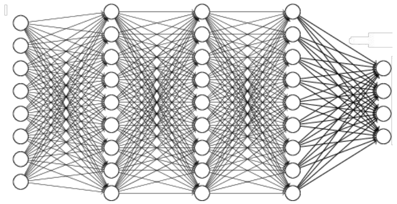
The input dimension is now 40 and 20,000 data samples are still relatively sparse in the space. SVM is suitable for this kind of problems as it assumes that there exists a hyperplane separating the input data into different classes, which is easy to be done in high-dimensional space. Here we employ one-v.s.-rest scheme and RBF kernel for multi-classification.





Neural Network

We have tried to train a classifier using neural network. A 5 layer fully connected network is used to train the model. Due to the constraint of computational power, we're unable to use wide-deep networks. Each layer has no more than 100 nodes right now. The performance is pretty poor till now with less than 20% accuracy. We would try to use AWS and GPU acceleration to train deeper & wider model to test if it works in the future.



Results

	SVM		Decision Tree
Pure	0.3345	Pure	0.3363
With RBF	0.3752	Random Forest	0.4084
With PCA	**0.5215**	With PCA	0.4602

The best accuracy we achieved is **52.15%**, with d^{PCA}=40, γ=0.01 and C=1. These parameters are determined by cross validation. The training on 21389 samples takes around 30s, and the testing on 2855 samples takes around 3s on Xeon E7 server.

Conclusion

We have presented a combined method to classify the type of a single cell based on its scRNA-seq expression with an accuracy of 52.15%.

First, PCA is used to extract the most important information and to accelerate training, and the lower dimensional data is normalized using Z-scores. The RBF kernel SVM with the one-v.s.-rest scheme is imposed for multi-class classification. RBF SVM is powerful enough and is suitable for sparse inputs in the high dimensional space. The parameters (e.g. γ, C) are also selected through cross validation.

In the future, ensemble methods such as adaboost will be examined to improve the accuracy.

References

[1] 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, Pages e156, https://doi.org/10.1093/nar/gkx681

[2] Amir Alavi, Matthew Ruffalo, Aiyappa Parvangada, Zhilin Huang, Ziv Bar-Joseph. scQuery: a web server for comparative analysis of single-cell RNA-seq data. bioRxiv 323238; https://doi.org/10.1101/323238

[3] Hankyu Jang, Samer Al-Saffar. Classifying Single-Cell Types from Mouse Brain RNA-Seq Data using Machine Learning Algorithms. http://hankyujang.com/Papers