Machine Learning Final Project

1. Description

Task 1. Mapping clusters between days & mapping clusters to known lineages

Do unsupervised learning on E5 and E7 cells respectively to identify clusters in each day, mapping the clusters between days, and mapping the clusters with known lineages.

Task 2. Classifying cells into lineages

Use E5 clusters (lineages) to train a lineage classifier (with feature selection), use CV to study the performance. And then apply the classifier to different days and analyze the results.

Group Member

Shengjian Chen (陈声健), Hsien Hwa Cheong (张显华), Jianqiao Hao (郝健乔)

Contribution

Shengjian Chen: mainly worked on Task 1 and the related part in the report.

Hsien Hwa Cheong: mainly worked on Task 2 and the related part in the report.

Jianqiao Hao: worked on both two tasks, code and report integration.

2. Experiment Design

Data Preparation

First of all, we took the steps similar to the original paper to select the most variable genes from the dataset. This was done by assuming that the expression distribution of a gene follows a negative binomial for which the variance depends on the mean, v = m + m2/r, where r is the overdispersion, implying that cv2 = v/m2 = 1/m + 1/r. I fitted each gene to such a model to estimate r. Then we selected genes with cv2 lower than 20 as the most variable genes. 248 genes were finally selected as the new dataset. To stabilize the estimation, we performed winsorization of the expression distribution of each gene, setting the most extreme value to the expression of the second most extreme cell.

For task 1, we first did log transformation on the whole dataset. To avoid any data going extreme, I did it like this

$$y(x) = \log{(1 + x)}.$$

Clustering in Days

We used PCA to reduce dimensionality to 100 and then t-SNE to 2. We plotted the t-SNE result and gave cells from different days different colors. This is a simple way to recognize clustering between days in the data. We also used unsupervised learning algorithms (k-means and PAM clustering) on the PCA subspace to separate the whole dataset into 5 clusters to see whether it could capture the developmental trend in days.

Segregation of ICM and TE

To recognize the clustering among lineages, we downloaded the RPKM data, which indicates the gene expression level, from the paper website. We extracted the rpkm of the marker genes of each lineage and assigned color to each cell according to the weighted mean of the expression of each lineage using weights -1 and 1 for ICM and TE genes respectively. This is also a very simple way to find the segregation of lineages.

In the original paper, they fitted a principal curve in the t-SNE subspace to indicate the developmental time and extracted pre-lineage cell using this time. However, we skipped this step. In order to recognized the pre-lineage cell, we did k-means on the t-SNE subspace with k=3 and got the pre-lineage cells. Then we did PCA on both original genes after log transformation and ICM-TE's 20 marker genes respectively and did k-means with k=2. We mapped the k-means result back to previous t-SNE space together with the pre-lineage cells.

To get rid of the effect of days, we repeated the procedure above on the E5 data. According to the previously known lineages, we also did k-means on E7 data with k = 2.

Segregation of ICM into EPI and PE

According to the literature, ICM will further separate into EPI and PE. To explore this segregation, we again plotted the E5 non-pre-lineage genes with color according to the weighted mean of the expression of PE and EPI marker genes (1 and -1 respectively). Then we extracted non-pre-lineage cells in E5 with high level of expression of ICM marker as the ICM cell, did PCA on these ICM cells and did k-means clustering on them. Color associated with the PE-EPI expression level in the ICM cells was also used to indicated this kind of segregation of ICM cell. Similar operation was carried out on E7 data to recognize the segregation of ICM into EPI and PE.

Classification of E5 cells into lineages

We then proceeded to the classification of E5 cells into lineages using labels from earlier results. According to the literature, E5 can be further separated into sub-stages called E5 early, E5 mid and E5 late, and the lineage assignment at each sub-stage might be different. For instance, E5 early still predominantly consists of pre-lineage cells and at E5 late, there will be more cells with EPI and PE because the segregation of ICM into EPI and PE occurs at this stage. Thus, E5 cells as a whole will be classified into four different categories. Before training the classifier, we performed feature selection on the genes to reduce complexity and to ensure their relevance in the model. To achieve that, we fitted a support vector machine (SVM) model on the data and examined the score of each feature, features with score below a certain threshold were then eliminated. This was done recursively by considering smaller and smaller set of features, each round the performance of the classifier was assessed using cross-validation. The process stops and returns an optimal number of features based on the cross-validation scores. The dataset with reduced features was then divided into train and test sets, the train set was used to train an SVM classifier, and predictions were made on the test set.

Predicting lineages in cells from other embryonic days

The SVM classifier was then applied on cells with other embryonic days in order to identify the lineages. First, the classifier was applied on E6 and E7 cells, we expected no cell will be classified as pre-lineage as E6 and E7 cells are well-differentiated. Next, we applied the same classifier on E3 cells and we expected the it to classify E3 cells as pre-lineage as these cells are undifferentiated. Lastly, we did the same thing on E4 cells, and we still expected the cells to be predominantly pre-lineage.

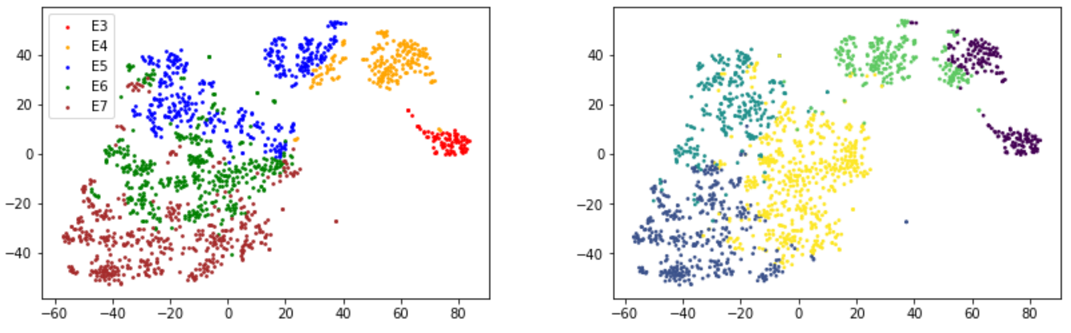
3. Method

All the experiments were carried out on Anaconda Jupyter Notebook with Python 3.6. Required libraries included Numpy, sklearn, pandas, matplotlib and pyclustering. Main machine learning algorithms used in this project include PCA, t-SNE, feature selection, k-means clustering, PAM clustering, SVM, Logistic Regression, Decision Tree, Random Forest, Adaboost, and cross validation. The data was provided by the course. We also used extra RPKM data from the original paper website (https://www.ebi.ac.uk/arrayexpress/experiments/e-mtab-3610/). To select the most variable genes, we used code on Github to fit data to a negative binomial model (https://github.com/gokceneraslan/fit_nbinom).

4. Result

Task 1 result

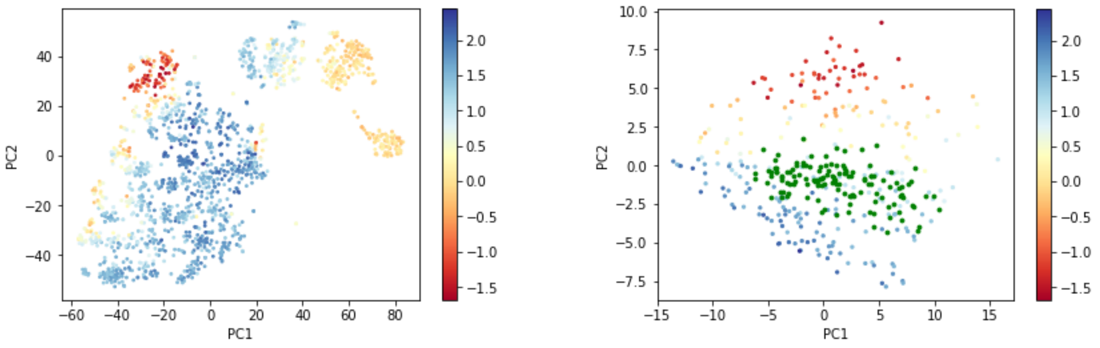
After data preprocessing, PCA and t-SNE, we got **Figure 1A**, clearly showing the clustering and segregation of the data by days and also the developmental continuity and trend. This was similar to the result on the original paper. However, the unsupervised learning methods failed to cluster the dataset by developmental days (**Figure 1B**). The unbalance among the size of cells from each day may be a very critical reason. Another reason is that day is not the only factor that affects the clustering in cells.



(A) (B)

Figure 1 Clustering in day. (A) result of colored clustering. (B) result of unsupervised learning

When we plotted the ICM-TE weighted mean expression across the whole dataset in **Figure S2A**, we could find roughly 3 clusters. The blue region with high positive value indicated TE expression while the red region indicated ICM expression. Those regions with value close to 0 are mainly pre-lineage cells. We have also noticed the large blue region in the plot. This seems to indicate the way we were using to measure expression level was biased, because different marker genes are expressed on different levels. Simply sum them up is not a very reasonable way. Doing unsupervised learning on the whole dataset was not a good idea. When doing clustering on the t-SNE subspace, cells were separated by days (**Figure S2B**). Since the term of "pseudo-time" in the original paper is defined in the t-SNE subspace, we found a way to avoid fitting such a pseudo time. We just simply assigned the yellow region in Figure S2B as the pre-lineage cells. When doing clustering on the PCA subspace of the non-pre-lineage cells, the result was in the middle, concerning both days and lineages (**Figure S2C**). However, when doing clustering on the PCA subspace of RPKM of the 20 marker genes (**Figure S2D**), we got result quiet familiar with Figure S2A.

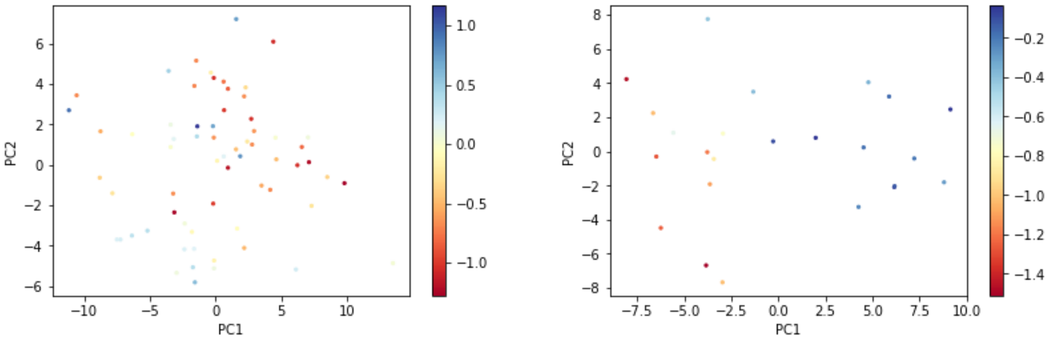


(A) (B)

Figure 2 Segregation of ICM and TE. (A) ICM-TE expression across all cells; (B) ICM-TE expression in E5 non-pre-lineage cells, green dots are pre-lineage cells;

In E5 data, we used the similar way to find the pre-lineage cells (purple region in **Figure S3A**) and did PCA on the non-pre-lineage cell. Clustering in the PCA subspace perfectly separated the E5 cells along the PC1 (**Figure S3B**) but using the weighted mean expression of the ICM-TE marker genes shows clear separation along PC2. And the pre-lineage cells lay in the middle of these two lineages (**Figure S3C**). Doing clustering on PCA subspace of RPKM of the marker genes of E5 non-pre-lineage also shows segregation on PC2 (**Figure S3D**). According to the weights we used previously, the bull region should be ICM cells and the red region should be TE cell. This is different from the original paper because they found the separation appeared in PC1. We think we still can't get rid of the effect of the developmental process of cells in time/day, or other unknown strong effect, even though we only work on the single day cells.

We repeated the above operation on E7 data. Again, unsupervised clustering in PCA subspace appeared along PC1 but related marker genes expression indicated that the segregation of lineages appeared along PC2 (**Figure S4**).



(A) (B)

Figure 3 Segregation of ICM into EPI and PE. (A) EPI-PE marker genes expression on E5 ICM cells; (B) EPI-PE marker genes expression on E7 ICM cells.

We further explored the segregation of ICM into EPI and PE on E5 and E7 data. We plotted the weighted means expression of EPI-PE on E5 and E7 cells respectively (**Figure S5A, B**). We found clear difference expression levels on ICM cells but not on TE cells, indicating that EPI-PE segregation only happened on ICM cells. We extracted ICM cells, did clustering on the PCA subspace and assigned color according to the marker expression level (**Figure S5C, D**). Segregation appears on PC1 but not so sharp. Such segregation also appears in clustering result in EPI-PE maker genes RPKM PCA subspace (**Figure S5E, F**). In the experience, the ICM cells we found with our method are much less that TE cells. Lack of enough samples might hurt the result of our clustering. As mentioned above, using weighted mean expression method to recognize different lineages has its disadvantage (different marker genes may express on different levels). Better methods are needed to improve the result.

Task 2 result

After the lineages have been found for E5 cells, we use the lineages as labels to train our classifier. Several classifiers are tried by us to find the best classifier which could fit well for our data. For each classifier, we first select the optimal number of features with RFECV and then use the optimal features to train our model. Table 1 shows the performance of each classifier. It is clear that both SVM and logistic perform well, and their performances are similar. The optimal number of features is about 200, where the accuracy rate reaches about 91%. Figure 5 gives the feature selection process of SVM and its confusion matrix on our test set.

Table 1 Comparison between Different Classifiers

Classifier	Optimal # of features	accuracy	precision	recall	F1-score
SVM	209	90.40%	88%	86%	87%
Logistic	188	91.51%	86%	88%	87%
Decision Tree	475	76.21%	70%	74%	71%
Random Forest	22	82.52%	78%	77%	76%
AdaBoost	2	67.41%	47%	52%	41%

* accuracy is reported based on CV; precision, recall, F1-score are averaged with macro method.

We apply both SVM classifier and Logistic Regression classifier to E3, E6, E7 and E4 cells. Table 2 shows the results, which mean the number of cells assigned to each lineage. The number given in the table is the result given by SVM classifier and the number shown in the parenthesis is predicted by Logistic Regression. Both the classifiers give the similar results. As we could see, when both classifiers are applied to E3, the result is as expected: all the cells are categorized into pre-lineages, since the cells haven't been differentiated. When applying the classifiers to E6 and E7 cells, we could find that there are nearly no undifferentiated cells, and TE cells take a quite large proportion. As for the ICM cells, there will be more EPI cells than PE cells, which is also coincided with the paper's experiment results (Figure2F in paper). However, when we further examine the E4 cells, according to the results given by both classifiers, we tend to believe that E4 cells are also (mostly) undifferentiated.

Table 2 Results when Applying the Classifiers to Other Cells

Cells	Pre-lineage	TE	PE	EPI	Total
E3	81(81)	0(0)	0(0)	0(0)	81
E4	186(190)	4(0)	0(0)	0(0)	190
E6	2(4)	371(361)	7(6)	35(44)	415
E7	0(1)	416(395)	22(17)	28(53)	466

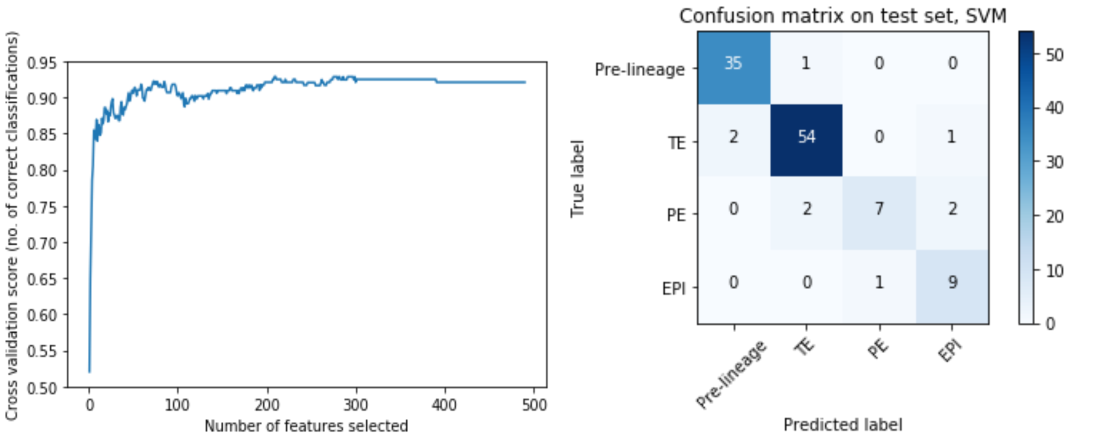


Figure 5. (a): Learning Curve when Selecting Features. (b): Confusion Matrix of SVM on test set

5. Conclusion and Discussion

In task 1, we explored the development of human preimplantation embryos and the segregation of lineages. With all the know information, we successfully found the clustering of cells from different days and lineages. However, after trying many algorithms and parameters setting, we have to conclude that purely using unsupervised learning methods that depend heavily on the distance between data points in not a very good practices in our particular situation. For one thing, there are many influential factors in our dataset, making it hard to study the effect of one factor and control the other. For another thing, the unbalance of cells size from different day and expression level of different genes make such unsupervised methods hard to capture the underlying clustering. As for how to better describe ICM-TE or EPI-PE segregation, our way of weighted mean expression is far from enough and we are still thinking about better methods.

When we use classifiers to train the lineages of the cells, we could find the both SVM and Logistic Regression could provide a satisfying result. About 90% of cells could be classified accurately with our model. And if we further apply the classifiers to other cells of different days, the expected result could be reached: all the E3 cells are classified into pre-lineage and nearly no E6/E7 cells are classified into pre-lineage. This might indicate that our lineages obtained in task 1 is quite precise. Moreover, according to our results, we tend to support the conclusion that E4 cells are more likely to be still undifferentiated.

6. References

Brennecke, P., Anders, S., Kim, J.K., Ko1odziejczyk, A.A., Zhang, X., Proser- pio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experi- ments. Nat. Methods 10, 1093–1095.

Petropoulos S , Edsg?Rd D , Reinius B , et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos[J]. Cell, 2016:S009286741630280X.

Acknowledgments

We want to express our great appreciation for Professor Zhang's instruction and TA's works in the past semester. We also want to thank TA Sijie Chen's helpful discussion and advice for our final project.