2022

ΨΗΦΙΑΚΕΣ ΥΠΗΡΕΣΙΕΣ ΥΓΕΙΑΣ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

DIGITAL HEALTH AND ANALYTICS

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ ΧΑΡΟΚΟΠΕΙΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΙΟΝΙΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΕΙΡΑΙΩΣ

SINGLE CELL RNA-SEQ TRANSCRIPTOMIC ANALYSES OF HUMAN PANCREATIC ISLETS

CHRISTINA MORAITI

Introduction

The goal of this project is the exploration of proper data preprocessing practices, so as to gain the best results via the dimensionality reduction algorithms during the classification and the clustering process. Three different algorithms for dimensionality reduction were applied. Furthermore, four classification algorithms and four clustering algorithms were used in the machine learning process. Two different versions of the main dataset were used, the first given the preprocessing steps and the second where a subset of two thousand highly variable genes were selected, mainly for the clustering process. However, these two versions were used in both learning methods in order to compare and learn from the given results. The results of the experiments are displayed in tables. In the notebook provided, there is the possibility of plotting boxplots of the metrics.

Transcription of DNA

Transcription is the first stage of gene expression and describes the process of producing molecule of RNA using a DNA chain as a template. Furthermore, the term transcriptome refers to the collection of all RNA transcripts and can be used to refer to all RNA or simply mRNA, depending on the experiment.

All known organisms' basic structural, functional, and biological unit is the cell. Because RNA functions as a messenger and a buffer molecule, transcription is an essential component of cell identity. Transcription elements found throughout the genome are appropriate for revealing the state of a cell in a specific environment because they are ideal for describing all encodings and non-cellular transcripts.

The dataset

In the exercise we use the dataset *GSE86469*, which contains single cell RNA sequencing data with gene expressions from 26.616 genes for 638 cell samples of two types of cells, diabetic and non-diabetic. These 638 human pancreatic islet cells were obtained from five non-diabetic and three type 2 diabetic cadaveric organ donors. Given these samples, cells that can cause diabetes belong to class 1 (positive) and non-diabetic ones belong to class 0 (negative). The dataset has a slightly uneven distribution given that the normal cells take up 59.56% and the diabetic ones 40.44%.

The dataset is a matrix of raw counts. The rows names represent genes, while the column names represent unique cell identifiers that were assigned by the authors of the dataset. The higher the count, the higher the expression. Genes play the role of encoders for proteins while proteins control how the cell functions. DNA is the storage of the genetic information which represents the genotype, while, the phenotype is the result of the 'interpretation' of the genotype. The most fundamental level at which the phenotype (observable trait) is given rise by the genotype, is the gene expression. The phenotype may be expressed by the synthesis of a protein which will control the organism's structure and development or a protein that will act as enzymes catalyzing specific metabolic pathways.

Therefore, a cell's capabilities are determined by the hundreds of genes that are expressed in that cell. Additionally, the cell can potentially regulate each stage of the informational chain from DNA to RNA to protein by altering the quantity and kind of proteins it produces. This allows the cell to self-regulate its operations.

Type 2 Diabetes

The first common symptom in type 2 diabetes is insulin resistance, a condition in which insulin is used by the body's cells with less efficient compared to normal. As this condition remains, higher levels of insulin are required in order to keep the blood sugar levels in the normal range. Meanwhile, the β cells in the pancreas, the cells in charge of producing the insulin must keep up with the increased needs of the body and produce higher amounts of insulin. The problem occurs when as a result of this procedure the cells become more and more tired of producing these large amounts and thus to a shortage of insulin and uncontrolled values of blood sugar levels in the body. In most cases, the body has a level of insulin resistance but as it ages up or by

the gain of weigh and inadequate exercise, the likelihood of developing type 2 diabetes is increased. [3]

To date, there have been found numerous mutations that can lead to a higher risk of type 2 diabetes, however, with a small contribution of each gene. The higher the additional mutations, the higher the risk. Generally, the genes included in the list of gene that can affect the appearance of type 2 diabetes include those controlling the production of glucose, the regulation of glucose and those controlling the way the glucose levels are being sensed in the body.

Data Preprocessing

The data frame was converted into an annotated data matrix in order to use the toolkit of Scanpy. Because of dealing with a dataset of RNA, a key problem is dealing with the large appearance of zero values, thus necessary to distinguish systematic, semi-systematic, and stochastic zeroes. Genes that are constitutively silent across the dataset's cells are referred to as systematic zeroes, with counts for each cell's equal to zero. Since they carry no information, they can be dropped. Stochastic zeroes are found in genes that are actively expressed, yet counts of zero are obtained for some cells due to sampling stochasticity. By eliminating these genes before normalization, biases may be introduced since they might contain information about the relative differences between cells. Semi-systematic zeroes are defined in genes that are silent in a cell subpopulation but are expressed in other cells and thus these genes provide information about the differences between subpopulations. [2]

➢ Quality Control

Removal of genes whose expression level is deemed "undetectable" is generally a good idea. A gene is considered detectable if it is present in at least two cells and has more than five readings. The threshold, however, is highly dependent on the depth of the sequencing. Since some genes may only be found in cells of low quality, it is crucial to keep in mind that genes must be filtered after cell filtering. Genes that are found in few cells are outliers and should thus be removed.

As a first step, the quality control metrices across cells and genes were calculated. After the quality control there were found 5,807 genes which were found in fewer than 2 cells and 6,009 genes which had less than 10 as a sum of their counts. As a result, 6,462 genes we're dropped off the dataset.

➢ Normalize expression values

The first bias that is necessary to be accounted is the gene length. For example, we have gene A and gene B, where gene A is longer than the other, meaning that a larger number of fragments will be mapped to gene A compared to gene B. While quantifying these fragments mapping to the genes, there will be more values for gene A and by looking at the counts it would seem like gene A is more expressed, which actually is not true. When comparing gene expression, it is necessary to correct for the bias caused by the differing lengths of the two genes.

The second type of bias that we need to account for is the sequencing depth. For example, when quantifying the fragments into counts for two samples and the counts for the genes in sample one is higher than the counts for genes in the second sample, we can be misled into the conclusion that the genes in the first sample are more expressed compared to the second sample. That is not the case because there is a difference due to the sequencing depth.

Also, a key problem of the RNA-seq data is a phenomenon called 'dropout'. Dropout is a phenomenon, where a gene is observed at low or moderate expression levels in one cell but not detected in another cell of the same cell type. This excessive amount of zero counts in the dataset can lead to the data to be zero-inflated, only capturing a small fraction of each cell’s transcriptome.[1]

➢ Identify highly variable genes

Highly variable genes are the genes whose expression varies across the cells. By finding these features it will provide a good separation of the cell clusters in the process of clustering. A subset of two thousand genes was selected, mainly to result in a better clustering of the data.

- **SCALE DATA**

The majority of machine learning models are based on Euclidean distances. Therefore, the square difference with the lower value in comparison to the far greater value will almost be ignored. In order to prevent that, it is necessary to transform all the variables into the same scale.

ALGORITHMS FOR DIMENSIONALITY REDUCTION

Dimensionality reduction algorithms are applied before clustering to extract principal components and reduce computational complexity. Faster calculations require the minimization of noise. This can be achieved by selecting traits that aim to identify the most informed genes, such as those with the highest variation. The data are projected onto a lower dimension space, preserving the essential information while eliminating noise. In the case of life sciences, the goal is to segregate samples based on gene expression patterns in the data. The following dimensionality reduction algorithms were used for that purpose. [6]

- **T-SNE**

T-distributed Stochastic Neighbor Embedding is a nonlinear dimensionality reduction approach in two- or three-dimensional spaces. If the number of features is too large, it is strongly advised to utilize another dimensionality reduction technique to lower the number of dimensions to a manageable quantity (such as 50). By following this approach, some noise will be reduced and the computation of pairwise distances between samples will be sped up. PCA was first applied in the data, by reducing the components to fifty.

- PCA

Principal component analysis is a linear dimensionality reduction algorithm that eliminates dependency or redundancy in the data by removing those genes that contain the same information as given by other attributes. Therefore, there is no dependence between the resulting components. Fifty components were chosen for the algorithm.

- FA

Factor analysis is also a linear dimensionality reduction algorithm, which explicitly assumes the existence of hidden components underlying in the observed data. Factor analysis takes into account only common variance, meaning only the variance shared with other variables will be considered, by excluding that way the specific and error variance. PCA, on the other hand, looks for variables that are combinations of the variables that were observed. Fifty components were chosen for the algorithm.

SUPERVISED LEARNING MODELS

In the following models k-fold cross validation was used, by dividing the data into ten even parts. For every experiment, there were also ten separate iterations.

In the selecting of the values of the models' parameters, Grid Search was used to automatically find the best parameters for the classifiers.

- **K-NEAREST NEIGHBORS CLASSIFIER**
- **DECISION TREE CLASSIFIER**
- **SUPPORT VECTOR MACHINE**
- **BAGGING CLASSIFIER**

CLASSIFICATION METRICS

For the classifiers the metrics used were accuracy, precision, recall and f1 score. [4]

- **ACCURACY**

  The fraction of predictions the model labeled truly (TP + TN). However, when working with an imbalanced dataset, accuracy alone as a metric isn't reliable.

- **PRECISION**

  The ratio of correctly predicted positive observations to the total predicted positive observations. High

precision relates to the low false positive rate. The question that we get the answer to is of all the cells that were predicted as diabetic (TP+FP) how many were actually indeed diabetic?

- **RECALL**

  Recall is the ratio of correctly predicted positive observations to all the observations in actual observations that were diabetic (TP+FN). The question recalls answer is: Of all the cells labeled as diabetic how many were predicted as diabetic?

- **F1 SCORE**

  F1 score is the weighted average of precision and recall. A very useful metric given that the dataset has an uneven class distribution, F1 score takes into account both false positives and false negatives.

Metrics	Models	PCA(adata_copy)	PCA(adata_hv)	PCA(Original data)	T-SNE	FA
Accuracy	K-NN	0.78-0.80	0.68-0.71	0.66-0.68	0.75-0.77	0.77-0.79
	DesicionTree	0.70-0.73	0.60-0.65	0.65-0.67	0.67-0.71	0.74-0.78
	SupportVector	0.82-0.83	0.71-0.74	0.68-0.69	0.62-0.64	0.83-0.85
	Bagging	0.76-0.80	0.65-0.69	0.63-0.67	0.71-0.74	0.78-0.81
Precision	K-NN	0.74-0.76	0.59-0.63	0.58-0.63	0.7-0.74	0.82-0.85
	DesicionTree	0.66-0.74	0.53-0.62	0.55-0.60	0.6-0.65	0.72-0.77
	SupportVector	0.86-0.90	0.63-0.68	0.62-0.64	0.5-0.55	0.82-0.85
	Bagging	0.75-0.80	0.59-0.66	0.55-0.61	0.67-0.74	0.76-0.85
F1 score	K-NN	0.72-0.75	0.60-0.65	0.55-0.58	0.68-0.71	0.66-0.69
	DesicionTree	0.57-0.64	0.36-0.50	0.51-0.58	0.55-0.64	0.64-0.72
	SupportVector	0.74-0.76	0.64-0.68	0.58-0.60	0.58-0.6	0.78-0.80
	Bagging	0.66-0.73	0.49-0.58	0.49-0.57	0.59-0.64	0.69-0.74
Recall	K-NN	0.72-0.75	0.63-0.70	0.54-0.56	0.66-0.7	0.56-0.58
	DesicionTree	0.52-0.60	0.29-0.45	0.49-0.59	0.5-0.63	0.57-0.67
	SupportVector	0.65-0.66	0.65-0.68	0.55-0.57	0.65-0.68	0.74-0.76
	Bagging	0.60-0.69	0.42-0.53	0.45-0.54	0.51-0.6	0.62-0.67

Table 1: Classification results.

RESULTS

Considering that this is a medical problem it is important to focus on the true positive predictions, as it is of greater seriousness to predict correctly a cell as "diabetic" and not misclassify it as "normal". The bigger problem in medicine is to misdiagnose an ill patient as healthy rather than misdiagnose a healthy patient as ill. Given the above, the metric that we need to be more focused upon is precision.

UNSUPERVISED LEARNING MODELS

The use of unsupervised clustering algorithms based on such noisy single-cell gene expression data has become the main computational strategy for identifying cell types, which is usually the first step for the subsequent analysis of RNA-seq data. During the learning process of the following models, the parameter for the number of clusters varied from one to ten in order to get ten different results.

- **K-MEANS**
- **HIERARCHICAL AGGLOMERATIVE SINGLE-LINKAGE**
- **HIERARCHICAL AGGLOMERATIVE MAX-LINKAGE("COMPLETE")**
- **SPECTRAL**

Clustering Metrics

For the clustering algorithms, the metrics used were silhouette coefficients, Dunn index, purity and rand measure. There are measurements for both internal and external goodness. Internal metrics evaluate the goodness of clusters based solely on the initial data, while external metrics evaluate the goodness of clusters using the knowledge about the known true split.

- *Internal Scoring Schemes*
  - **Davies–Bouldin index**

    The score is described as the average of the similarity measures of each cluster with a cluster most similar to it. Clusters which are farther apart and with a lower intra-cluster dispersion will result to a better score, with the lower scores indicating a better clustering. The minimum score is zero.
  - **Silhouette coefficients**

    This metric tells us how well-assigned each individual point is. A result close to one represents that the cluster's samples are away from the neighboring cluster's samples. A value equal to zero or close to zero indicates that the sample is on or very close to the decision boundary between the two clusters and negative results indicate that those samples might have been assigned to the wrong cluster. A higher Silhouette Coefficient score relates to a model with better defined clusters.
- *External Scoring schemes*
  - **Purity**

    A label is assigned in each cluster based on the most frequent class in it. The purity score is the number of correctly matched class and cluster labels divided by the number of total data points. Each cluster is assigned with the most frequent class label. In order to calculate it is necessary to create a confusion matrix. The purity result comes from all the true labeled cells divided by all the cells and is helpful to give us details on the composition of each cluster.
  - **Rand Measure**

    This metric can be considered as a measure of the percentage of correct decisions made by the algorithm. The Rand Index is a measure of similarity between the predicted and the ground truth clusters by examining at whether pairs of data points are in the same or different clusters. The results lying between 0 and 1, with values closer to 1 as better.

Metrics	Models	PCA(adata_hv)	PCA(adata_copy)	PCA(Oiginal data)	T-SNE	FA
Silhouette co.	K-means	0.57	0.16	0.28	0.23	0.78
	HAsingle-l	0.79	0.72	0.78	0.85	0.78
	HAmax-l	0.79	0.71	0.78	0.85	0.78
	Spectral	0.56	-	-	-	0.78
Davies-Bouldin	K-means	1.39-2.37	1.68-2.59	1.17-3.00	0.71-1.69	0.13-0.94
	HAsingle-l	0.13-0.18	0.18-0.22	0.13-0.15	0.09-0.41	0.13-0.14
	HAmax-l	0.13-1.25	0.63-1.27	0.14-1.06	0.10-0.76	0.14-1.80
	Spectral	0.31-0.41	-	-	-	0.80-2.05
Purity	K-means	0.60-0.62	0.59-0.61	0.59-0.68	0.59-0.61	0.59-0.60
	HAsingle-l	0.59-0.60	0.59	0.59	0.59-0.60	0.59
	HAmax-l	0.59-0.60	0.59	0.58-0.64	0.59-0.60	0.59
	Spectral	0.59	-	-	-	0.59
Rand measure	K-means	0.49-0.52	0.49-0.50	0.51-0.52	0.49-0.50	0.49-0.51
	HAsingle-l	0.51	0.51	0.51	0.51-0.52	0.51
	HAmax-l	0.51	0.51	0.51	0.48-0.52	0.51
	Spectral	0.51	-	-	-	0.51

Table 2: Clustering results.

Results

Indeed, when providing a model with the data which obtained a subset of the highly expressed genes, we get better results as seen above in the table. The original data had good results in the silhouette scores for the Hierarchical clustering but had higher scores in the Davies-Bouldin scores.

Conclusion

The first step, when given a new unseen before dataset is understanding the data and the existing variance underlying in them. In this case, I believe there is a limit in the information provided for the cells. For example, what type of cell is each cell from the pancreas or maybe characteristics of the donor as their age, ethnicity or BMI in order for a more efficient preprocessing approach of the data. Furthermore, each dataset has its specificities and requires a different handling on towards obtaining the most information possible and creating a reliable and efficient model.

References

1. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat Commun* **11**, 1169 (2020). https://doi.org/10.1038/s41467-020-14976-9.
2. L. Lun, A.T., Bach, K. & Marioni, J.C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 17, 75 (2016). https://doi.org/10.1186/s13059-016-0947-7.
3. Emanuele Bosi, Lorella Marselli, Carmela De Luca, Mara Suleiman, Marta Tesi, Mark Ibberson, Decio L Eizirik, Miriam Cnop, Piero Marchetti, Integration of single-cell datasets reveals novel transcriptomic signatures of β-cells in human type 2 diabetes, *NAR Genomics and Bioinformatics*, Volume 2, Issue 4, December 2020, lqaa097, https://doi.org/10.1093/nargab/lqaa097.
4. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, Parasa S. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022 Apr 8;12(1):5979. doi: 10.1038/s41598-022-09954-8. PMID: 35395867; PMCID: PMC8993826.
5. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LDF, Rodrigues FA. Clustering algorithms: A comparative approach. PLoS One. 2019 Jan 15;14(1):e0210236. doi: 10.1371/journal.pone.0210236. PMID: 30645617; PMCID: PMC6333366.
6. Xiang R, Wang W, Yang L, Wang S, Xu C, Chen X. A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. Front Genet. 2021 Mar 23;12:646936. doi: 10.3389/fgene.2021.646936. PMID: 33833778; PMCID: PMC8021860.