

**Machine Learning
AAT Report
19EC6CE1ML
On
“Single cell RNA sequencing”**

Submitted in partial fulfilment of requirements for the degree
Of
Bachelor of Engineering in Electronics and Communication



Visvesvaraya Technological University, Belgaum

Submitted By:

<i>Sudarshana S Rao</i>	<i>1BM18EI053</i>
<i>Pranav Manjunath Bhat</i>	<i>1BM18EC100</i>
<i>Sudamshu S Rao</i>	<i>1BM18EI052</i>

Academic Year
2020-2021



Department of Electronics and Communication Engineering, Bangalore
B.M.S. College of Engineering
Bull Temple Road, Basavanagudi, Bangalore-560019
(Autonomous college affiliated to Visvesvaraya Technological University, Belgaum)

ACKNOWLEDGEMENT

The completion of this undertaking could not have been possible without the assistance and participation of so many people. Their contributions are sincerely appreciated and greatly acknowledged. The team would like to express their deep appreciation particularly to the following people.

We extend our heartfelt thanks to the principal Dr. Ravishankar B. V. and the management and staff of BMSCE for providing us with necessary infrastructure that enables us to implement anything our engineering mind's desire.

Words cannot describe our gratitude to the Department of ECE, BMSCE for offering this course as an elective.

We wish to express our sincere acknowledgement to our guide Prof. Latha H N, Assistant Professor, Department of ECE, BMSCE for her valuable suggestions and guidance throughout the course of our project. We are grateful for her belief in us for the implementation of a project of this magnitude. We are truly grateful for her constant support and for allowing us to work flexibly which enabled us to set our own deadlines and objectives.

A large amount of material has been obtained from online journals, textbooks and numerous other scholarly and voluminous sources. We thank all those unseen faces for their invaluable contribution to our venture.

We thank our beloved parents on whose blessings we live and thrive. It is their prayers that have helped us translate our efforts into fruitful achievements.

TABLE OF CONTENTS

Serial number	Title	Page Number
1	Area of application	5
2	Introduction	6
3	Literature survey	7-8
4	Base paper details	9-10
5	Methodology	11-14
6	Results & conclusions	15-17
7	Future trends	17
8	References	18
9	Plagiarism report	19

LIST OF IMAGES

Serial number	Name	Page number
1	Principle block diagram of the program	9
2	Hierarchical heat map	10
3	Flow chart	11
4	Shape of the dataset	12
5	Graph plot of a column	12
6	The data after the log transformation	13
7	3-D plot of the PCA data	13
8	Elbow method graph	14
9	Clusters of the gene	15
10	Confusion matrix	15
11	Classification metrics	16
12	Graph plot of predicted values vs actual values of the testing dataset	16
13	Accuracy of the model	17

SINGLE CELL RNA SEQUENCING

Area of application:

Domain- Biotechnology & gene therapy.

RNA sequencing is a sequencing technique that utilises next generation sequencing (NGS) to determine the amount of RNA in a given sample. RNA sequencing can be used to identify diseases, create biomarkers for clinical indications, find biological pathways for medicinal drug delivery and make genetic diagnoses. This information can be used for personalised disease prevention, diagnostics and therapy for individuals or subgroups.

Given that cells store huge numbers of genes, each concerned with the release of various biochemicals within the organism, sifting through and analysing thousands, if not millions of genes is impossible to accomplish through manual means. This is one of the myriad places where machine learning can be used to visualise important data and gleam useful information from. By establishing the hierarchy of the genes, researchers can identify the more important ones, which will enable them to develop better drugs for diseases and can even aid in vaccine development.

INTRODUCTION

Single cell RNA-sequencing provides transcriptional (expression) profiling which is a measure of activity (expression) of thousands of genes at once to create a global picture of cellular function. The goal of our program is to analyse the single cell RNA sequencing dataset, which is in the form of a count matrix giving the number of reads in a cell, and come up with a hierarchical structuring for the genes and identify the more important genes.

The datasets are all different subsets of a larger single-cell RNA-sequencing dataset, compiled by the Allen Institute. The dataset used contains cells from the mouse neocortex, a region in the brain which governs higher-level functions such as perception and cognition. The single-cell RNA-sequencing data comes in the form of a counts matrix, where each row corresponds to a cell and each column corresponds to the normalized transcript compatibility count (TCC) of an equivalence class of short RNA sequences, rescaled to units of counts per million. TCC entry at location (i, j) of the data matrix can be thought of as the level of expression of the j-th gene in the i-th cell. The unsupervised npy file contains only the counts matrix and the supervised npy contains a labelled dataset with the ground truths obtained by the scientists.

This report covers the research, theory, concepts and results of our python program on single cell RNA sequencing.

LITERATURE SURVEY

- *Likai Wang, Yanpeng Xi, Sibum Sung & Hong Qiao:” A survey of dimension reduction and classification methods for RNA-Seq data on malaria vector”, BMC Genomics, 20th July 2018.*

ML-based differential network analysis has been applied to predict stress-responsive genes through learning the patterns of 32 expression characteristics of known stress-related genes. In addition, the epigenetic regulation plays critical roles in gene expression, therefore, DNA and histone methylation data has been shown to be powerful for ML-based models for prediction of gene expression in many systems, including lung cancer cells. Therefore, it is promising that ML-based methods could help to identify the DEGs that are not identified by traditional RNA-seq methods. Here they determined that the model based on InfoGain feature selection and Logistic Regression classification is powerful and robust for DEGs prediction. From this paper we can conclude that Logistic Regression showed the best performances, with an AUC of 0.839 and accuracy of 78.6% for training data in this case.

- *Ren Qi, Anjun Ma, Qin Ma and Quan Zou: “Clustering and classification methods for single-cell RNA-sequencing data”, Briefings in Bioinformatics China 2019.*

In this paper they used a small dataset as one with RNA-seq data of less than 100 cells, a large dataset had more than 1000 cells and a medium-size dataset was in the middle. Using 12 published sets of single cell RNA data, we showed the classification and clustering performance of each algorithm. The dimensionality reduction methods used were Principal component analysis, T distributed stochastic neighbour embedding and zero inflated factor analysis. There were several clustering methods used and the ones with the best clustering effects were differentiated from the worst. The main clustering algorithms used were K-means, X-means, Spectral and EM. The tools used for clustering were SC3, Single-cell regulatory network inference and clustering, SEURAT, Single-cell interpretation via multi-kernel learning, SINCERA, Shared nearest neighbour (SNN-Cliq), Nonnegative matrix factorization, Monocle, BiSNN-Walk, BackSPIN and Single-cell representation learning. It was concluded that SC3 and SIN had a better clustering effect for the given data sets than the rest

- *Stephane Wenric, Ruhollah Shemirani:” Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies”, Frontiers in genetics 2018.*

Using the supervised learning-based gene selection method after the differential expression one (i.e., using only the genes with a significant differential expression adjusted p -value as input features of the random forests classifier or the EPS method) also produced worse results than when using the random forests gene ranking or the EPS gene ranking alone. Using survival analysis as a way to validate gene lists coming from cancer data sets whose average survival differs greatly, however there does not seem to be a link between the overall survival (OS) of these cancers and the performance of the proposed methods. The samples are split into a training set and a validation set. Differential expression analysis is performed on the training set with the DESeq2 software package, using default parameters and options. A random forests classifier is built on the training set with the ranger R package. The EPS method is applied on the training set(s) to extract a ranking of genes. For each data set, 60 log-rank tests have been performed on the validation set, using gene signatures $sigDEi$, $sigRFi$, and $sigEPSi$ with $i = \{1, 2, \dots, 20\}$ which contain from 1 to 20 genes out of the gene ranking derived from differential expression analysis, the gene ranking derived from the random forest's classifier, and the gene ranking derived from the EPS method respectively. The p -values of these tests have been compared two by two.

BASE PAPER DETAILS

- Ren Qi, Anjun Ma, Qin Ma and Quan Zou: “Clustering and classification methods for single-cell RNA-sequencing data”, Briefings in Bioinformatics China 2019.

In this paper, we focus on clustering and classification methods for scRNA-seq data, in particular, integrated methods, and provide a comprehensive description of scRNA-seq data. They used a small dataset as one with RNA-seq data of less than 100 cells, a large dataset had more than 1000 cells and a medium-size dataset was in the middle.

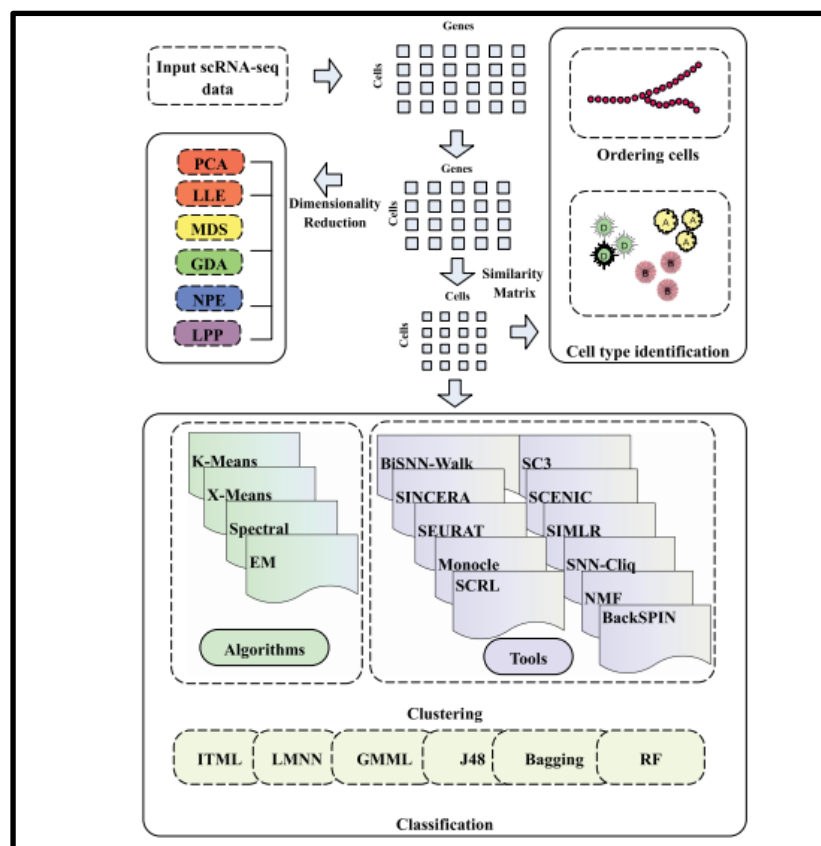


Figure 1: Principal block diagram of the program.

The dimensionality reduction methods used were Principal component analysis, T distributed stochastic neighbour embedding and zero inflated factor analysis.

Clustering is another effective method to detect cell types. The tools used for clustering were SC3, Single-cell regulatory network inference and clustering, SEURAT, Single-cell interpretation via multi-kernel learning, SINCERA, Shared nearest neighbour (SNN-Cliq),

Nonnegative matrix factorization, Monocle, BiSNN-Walk, BackSPIN and Single-cell representation learning.

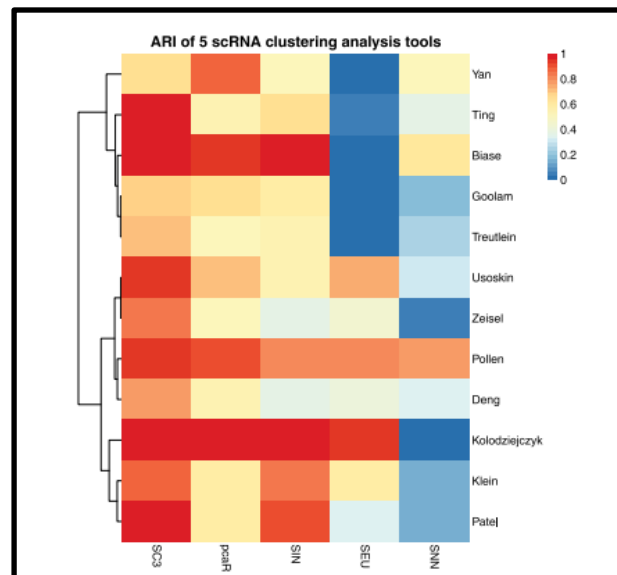


Figure 2: Hierarchical heat map.

ARI of five scRNA clustering analysis tools. Abbreviations like SNN, SIN, SEU and pcaR are used to represent the clustering tool SNN-Cliq, SINCERA, SEURAT and pcaReduce, respectively.

The final conclusion was the closer to red means the better clustering effect, and the closer to blue means the worse clustering effect which meant that SC3 and SIN were the best clustering tools used.

METHODOLOGY

Flow chart:

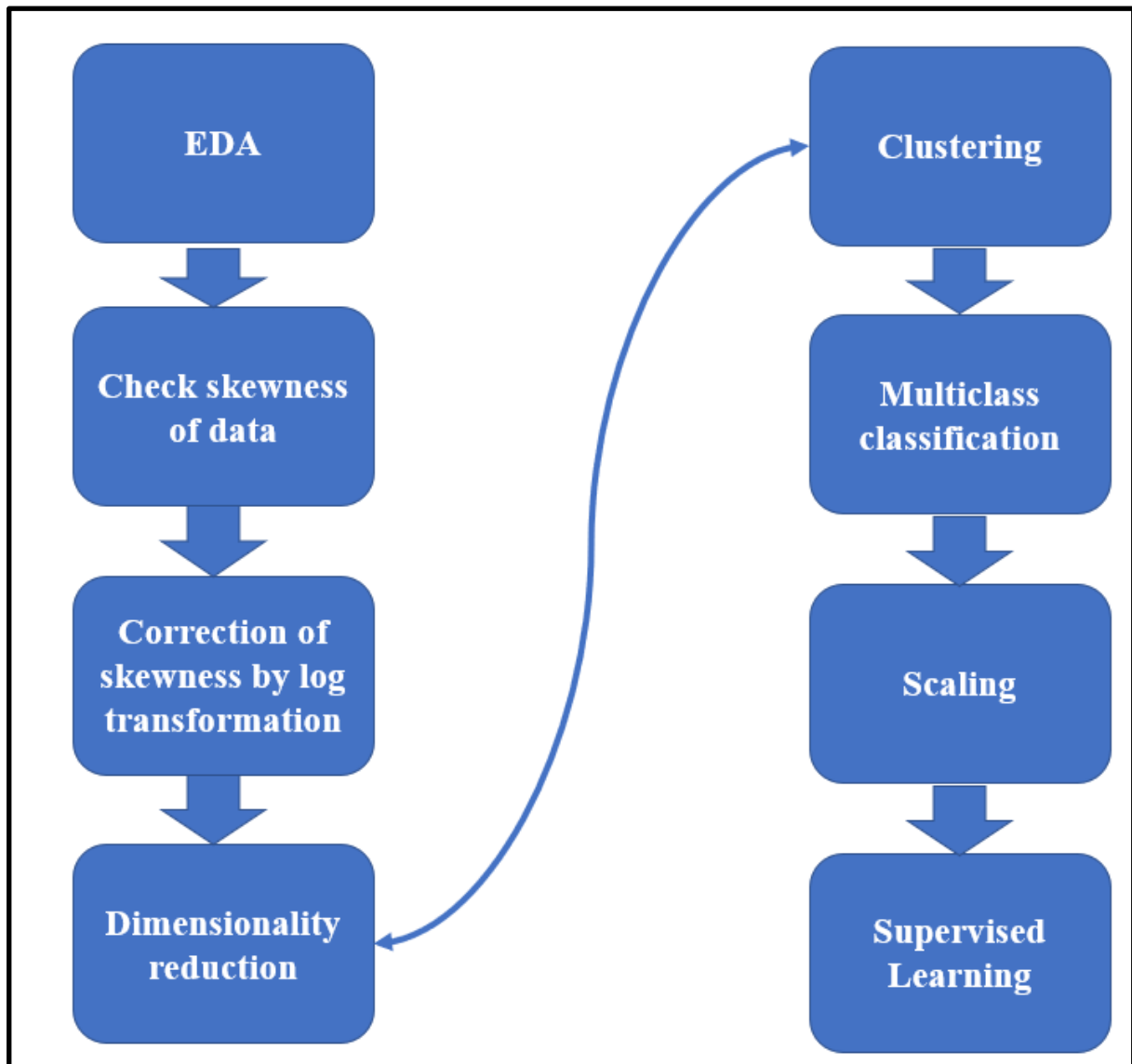


Figure 3: The above flow chart depicts the sequential flow of the program.

```
x_unsup = np.load(path+'X_unsup.npy')
x_unsup = np.array(x_unsup)
df_unsup = pd.DataFrame(x_unsup)

df_unsup.shape

(2169, 20000)
```

Figure 4: Shape of dataset

The exploratory data analysis revealed that the dataset has positive skew for the first set entries, followed by negative skew for the last set of entries.

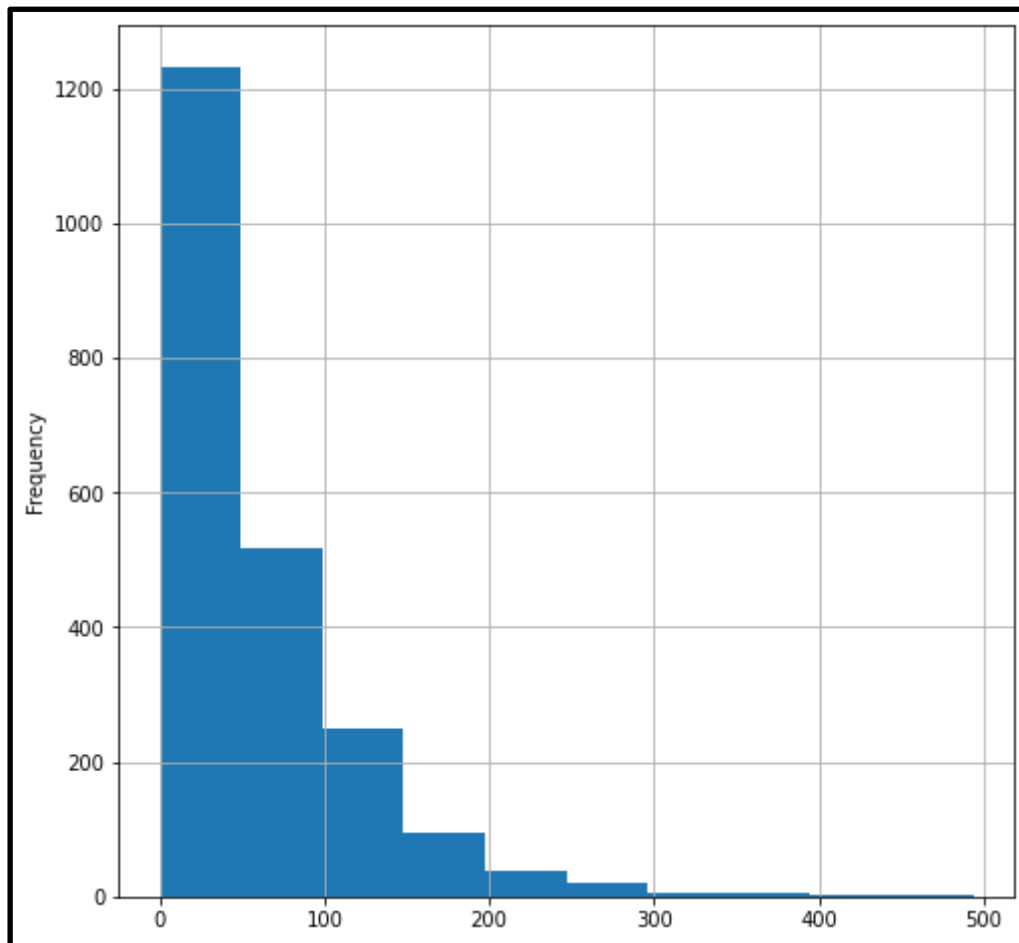


Figure 5: Data plot obeys the power law (skewed).

This skewness was corrected by $\log_2(x + 1)$ transformation. After this step, principal component analysis was used to reduce the number of dimensions of the dataset from 20,000 to 3. The new 3-dimensional data was plotted in a 3D graph.

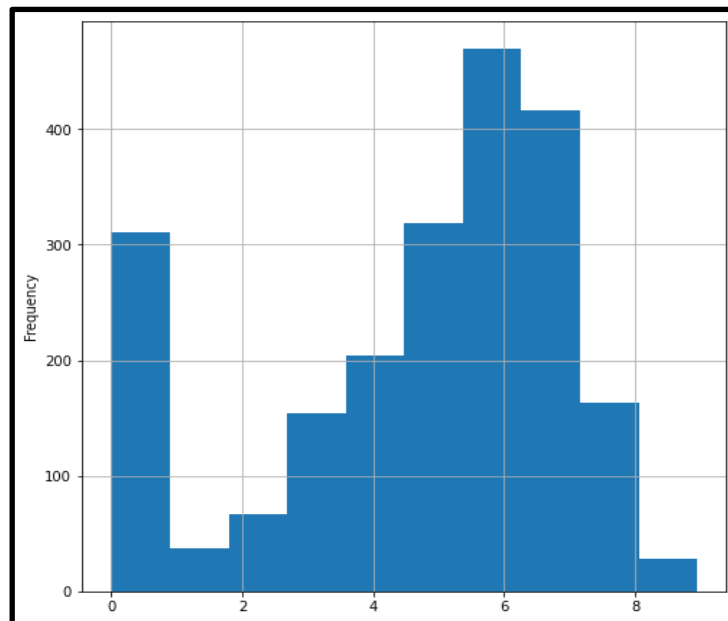


Figure 6: The data after the log transformation.

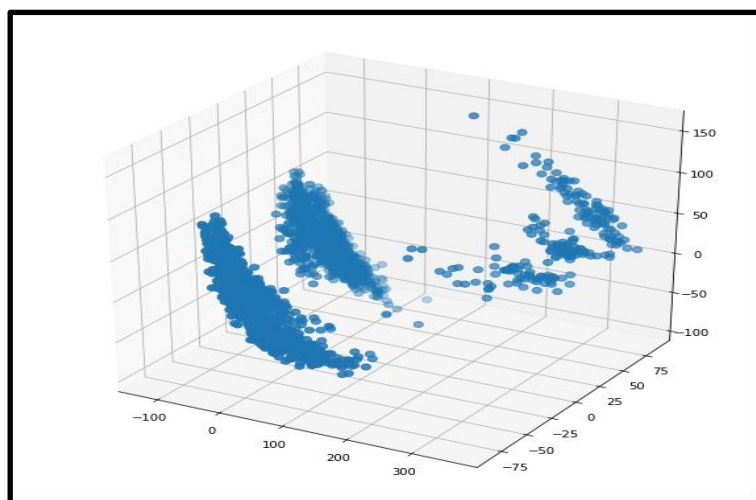


Figure 7: 3-D plot of the PCA data.

A visual analysis of the above plot indicates that 3 or 4 clusters exist. To narrow down on the optimal number of clusters for KMeans, the elbow method was used.

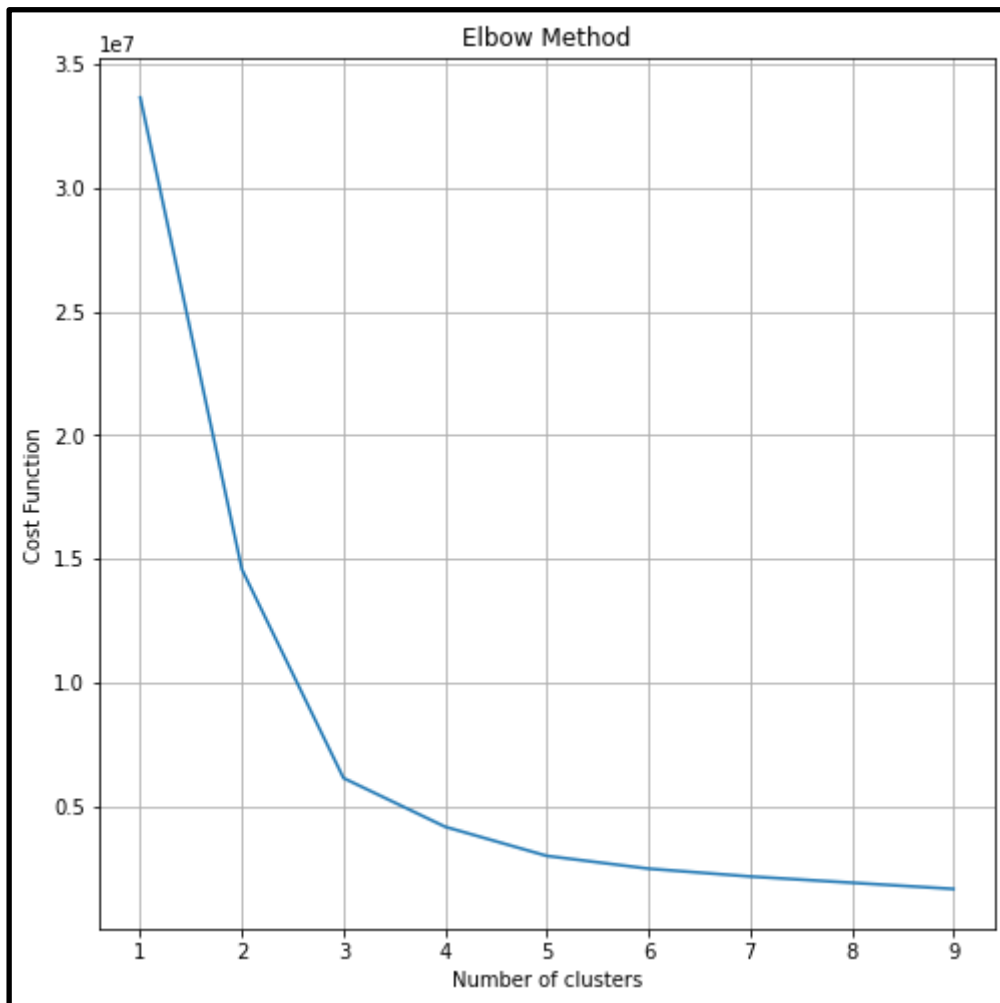


Figure 8: Elbow method graph.

RESULTS & CONCLUSION

Results:

The preceding analyses of the dataset and the process of finding the optimal number of clusters made it clear that 3 is the total number of relevant clusters. As such, they were plotted on a 3-D dynamic plot.

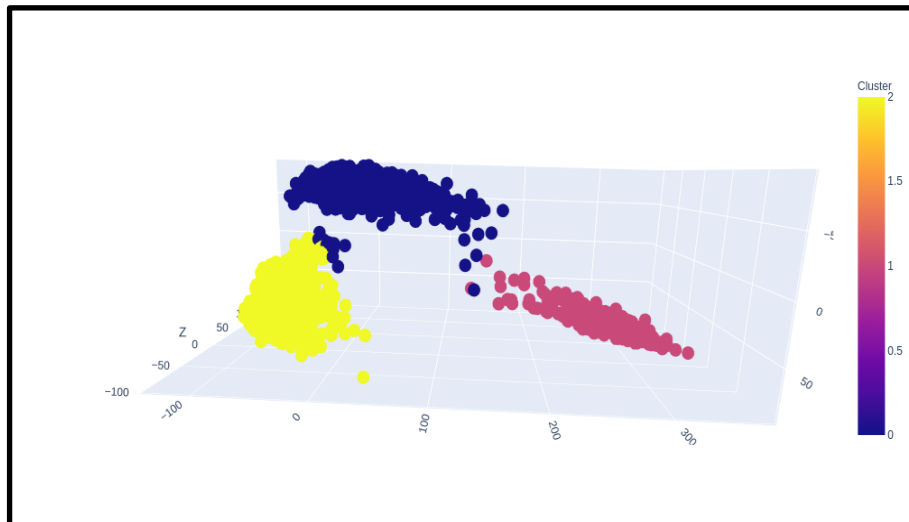


Figure 9: Clusters of the gene.

The cluster labels were then used to classify various genes using one-versus-rest Logistic regression as the classifier. The confusion matrix and the classification metrics were then obtained.

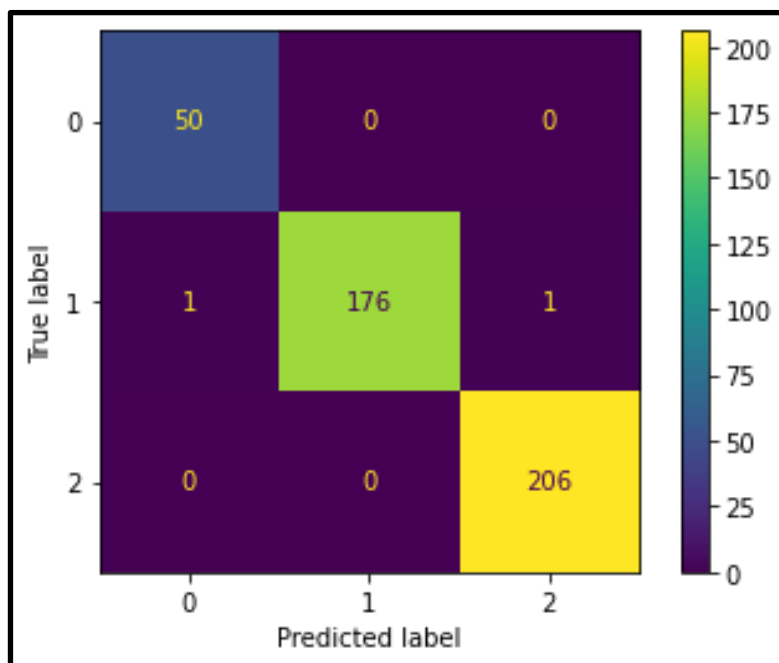


Figure 10: Confusion matrix.

	precision	recall	f1-score	support
0	1.00	0.99	0.99	165
1	1.00	1.00	1.00	45
2	0.99	1.00	1.00	224
accuracy			1.00	434
macro avg	1.00	1.00	1.00	434
weighted avg	1.00	1.00	1.00	434

Figure 11: Classification metrics.

Finally, supervised learning was performed on the given set of labelled data using Logistic Regression.

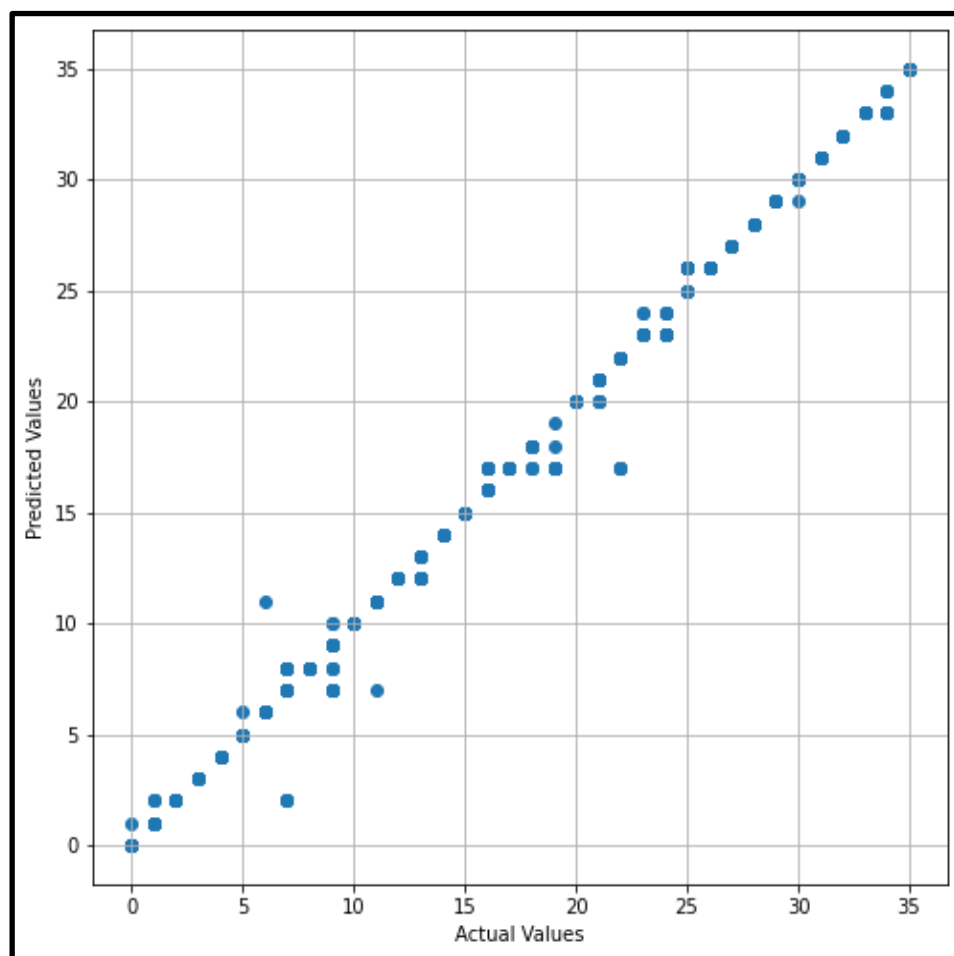


Figure 12: Plot of predicted values vs actual values of the testing dataset.


```
from sklearn.metrics import accuracy_score, mean_squared_error
y_prdct = model.predict(x_test)
print('Mean square error:', mean_squared_error(y_test, y_prdct))
print('Accuracy: %.2f'%(accuracy_score(y_test, y_prdct)*100), '%')
```

```
Mean square error: 0.3700361010830325
Accuracy: 89.80 %
```

Figure 13: Accuracy of the model.

Conclusion:

Principal component analysis (PCA) turned out to be an efficient dimensionality reduction technique with a low turnover time unlike t-distributed stochastic neighbours (TSNE) algorithm which usually requires greater hyperparameter tuning and has higher turnover time. KMeans was implemented on the dimensionally reduced data and it proved itself to be a simple, yet powerful tool for clustering and demonstrated its versatility. The following multiclass classification using Logistic Regression had a negligible number of misclassifications, effectively demonstrating how Logistic regression can be used for such purposes. Taking forward the same model, but with modified hyperparameters designed for regression analysis, we obtained a satisfactory level of accuracy with a trained dataset.

FUTURE TRENDS


The machine learning model can be deployed into production using Flask or TensorFlow. With some modifications designed to handle datasets of varying dimensions, it can truly be an asset to researchers working in the fields of biotechnology and gene therapy. In time, with increased usage and data uploads, the model can make predictions on data using only learning models already uploaded. This benefit will aid in quick analyses of newly constructed datasets.

REFERENCES

1. Likai Wang, Yanpeng Xi, Sibum Sung & Hong Qiao:” A survey of dimension reduction and classification methods for RNA-Seq data on malaria vector”, BMC Genomics, 20th July 2018.
2. Ren Qi, Anjun Ma, Qin Ma and Quan Zou: “Clustering and classification methods for single-cell RNA-sequencing data”, Briefings in Bioinformatics China 2019.
3. Stephane Wenric, Ruhollah Shemirani:” Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies”, Frontiers in genetics 2018.

PLAGIARISM REPORT

In the plagiarism report, literature survey and base paper details are not included.



PLAGIARISM SCAN REPORT

Words	736	Date	June 17, 2021
Characters	4794	Excluded URL	

0%

Plagiarism

100%

Unique

0

Plagiarized Sentences

35

Unique Sentences

Content Checked For Plagiarism

RNA sequencing is a sequencing technique that utilises next generation sequencing (NGS) to determine the amount of RNA in a given sample. RNA sequencing can be used to identify diseases, create biomarkers for clinical indications, find biological pathways for medicinal drug delivery and make genetic diagnoses. This information can be used for personalised disease prevention, diagnostics and therapy for individuals or subgroups. Given that cells store huge numbers of genes, each concerned with the release of various biochemicals within the organism, sifting through and analysing thousands, if not millions of genes is impossible to accomplish through manual means. This is one of the myriad places where machine learning can be used to visualise important data and gleam useful information from. By establishing the hierarchy of the genes, researchers can identify the more important ones, which will enable them to develop better drugs for diseases and can even aid in vaccine development. Single cell RNA-sequencing provides transcriptional (expression) profiling which is a measure of activity (expression) of thousands of genes at once to create a global picture of cellular function. The goal of our program is to analyse the single cell RNA sequencing dataset, which is in the form of a count matrix giving the number of reads in a cell, and come up with a hierarchical structuring for the genes and identify the more important genes. The datasets are all different subsets of a larger single-cell RNA-sequencing dataset, compiled by the Allen Institute. The dataset used contains cells from the mouse neocortex, a region in the brain which governs higher-level functions such as perception and cognition. The single-cell RNA-sequencing data comes in the form of a counts matrix, where each row corresponds to a cell and each column corresponds to the normalized transcript compatibility count (TCC) of an equivalence class of short RNA sequences, rescaled to units of counts per million. TCC entry at location (i,j) of the data matrix can be thought of as the level of expression of the j-th gene in the i-th cell. The unsupervised npy file contains only the counts matrix and the supervised npy contains a labelled dataset with the ground truths obtained by the scientists. This report covers the research, theory, concepts and results of our python program on single cell RNA sequencing. The exploratory data analysis revealed that the dataset has positive skew for the first set entries, followed by negative skew for the last set of entries. This skewness was corrected by $\log_2(x + 1)$ transformation. After this step, principal component analysis was used to reduce the number of dimensions of the dataset from 20,000 to 3. The new 3-dimensional data was plotted in a 3D graph. A visual analysis of the above plot indicates that 3 or 4 clusters exist. To narrow down on the optimal number of clusters for KMeans, the elbow method was used. The preceding analyses of the dataset and the process of finding the optimal number of clusters made it clear that 3 is the total number of relevant clusters. As such, they were plotted on a 3-D dynamic plot. The cluster labels were then used to classify various genes using one-versus-rest Logistic regression as the classifier. The confusion matrix and the classification metrics were then obtained. Finally, supervised learning was performed on the given set of labelled data using Logistic Regression. Principal component analysis (PCA) turned out to be an efficient dimensionality reduction technique with a low turnover time unlike t-distributed stochastic neighbours (TSNE) algorithm which usually requires greater hyperparameter tuning and has higher turnover time. KMeans was implemented on the dimensionally reduced data and it proved itself to be a simple, yet powerful tool for clustering and demonstrated its versatility. The following multiclass classification using Logistic Regression had a negligible number of misclassifications, effectively demonstrating how Logistic regression

can be used for such purposes. Taking forward the same model, but with modified hyperparameters designed for regression analysis, we obtained a satisfactory level of accuracy with a trained dataset. The machine learning model can be deployed into production using Flask or TensorFlow. With some modifications designed to handle datasets of varying dimensions, it can truly be an asset to researchers working in the fields of biotechnology and gene therapy. In time, with increased usage and data uploads, the model can make predictions on data using only learning models already uploaded. This benefit will aid in quick analyses of newly constructed datasets

Sources

Similarity