AMSI VACATION RESEARCH SCHOLARSHIPS 2019–20

EXPLORE THE MATHEMATICAL SCIENCES THIS SUMMER


# Method towards Precision Medicine

## Predicting Responses to Treatments with Single Cell Data

Samantha Chew
Supervised by Jean Yang
The University of Sydney

Vacation Research Scholarships are funded jointly by the Department of Education and the Australian Mathematical Sciences Institute.

# Abstract

Precision medicine offers the means to customise effective treatments for individual patients through a deeper understanding of human genetic profiles and technological advancements. Treatment customisation encompasses the prediction of patient response towards treatments, which relies on our understanding of machine learning. Along with the emergence of single cell technology, the extraction of "higher definition" cell data has now enabled more precise machine learning models to be constructed. The aim of project involves the identification of the most accurate machine learning algorithm for predicting the response outcome of patients towards treatment and the investigation of the most suitable data transformation for normalising single cell data. Plot comparisons affirmed that gene expression data is most suitable for the construction of the highest-performing random forest model. Simultaneously, the use of gene expression data for logistic regression model classification was challenged as compositional single cell data consistently presented more accurate logistic regression models and can evidently be normalised best using a logit transformation.

# Introduction

Traditional medicine assumes humans are similar and treats patients with similar symptoms using standard treatments. However, patients may not always react well with these treatments and experience side effects. This is because there are many features among cells that make each of them unique despite being known to function similarly across all humans due to their similarities in structure.

Traditional bulk RNA-Seq analysis involves the analysis of averaged cell measurements and may overlook some of these differences as a result. With the recent development of single cell RNA sequencing, expression profiles can now be extracted from individual cells and analysis can be done with less information-loss. Along with machine learning algorithms, the subtle difference between individual cells can be used to build better prediction models to aid with choosing the best treatment for each individual. Such personalised treatment is the concept of precision medicine.

Cancer patients are generally given a combination of surgery, chemotherapy, radiation and immunotherapy depending on the cancer type, size and stage [2]. Precision medicine enables health professionals to customise more effective treatment combinations based on their understanding of the patient's genetic profile. This ultimately reduces the risk of patient unresponsiveness and alleviates side effects that occur due to the one-size-fits-all approach.

# Aim of this Project

Precision medicine involves predicting the response outcome of patients towards certain treatments. With the ultimate goal of improving such prediction model, the aim of this project was broken down into the following three sections:

- To identify the best machine learning algorithm for prediction model construction using single cell data
- To test if compositional single cell data is a better alternative to gene expression data for the purpose of prediction model construction
- To identify the most suitable data transformation for single cell analysis

# Statement of Authorship

The workload was divided as follows:

- Samantha Chew compared the accuracy rates of different machine learning algorithms and investigated if compositional single cell data performed well in comparison to gene expression data.
- Kelvin Liu investigated different metrics, such as normality, to determine which data transformation is the most suitable for single cell analysis.

# Dataset description

The sadeFeldman melanoma dataset used for this project was obtained from the public database GEO, which is accessible at GSE120575 [1]. The dataset was from the "Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma" study done by Sade-Feldman et al. This study aimed to identify factors associated with success or failure of checkpoint therapy. The dataset consisted of 16291 single cell samples that contained 50513 different genes. The cell samples were mapped to the patient ID, patient respond outcome, therapy name and their respective cell types. The cell samples were of 11 different cell types and were extracted from 29 separate patients. The patients were divided into "Responder" and "Non-responder" groups, which were representative of whether a patient responded to the given treatment.

# Approach

Using the sadeFeldman dataset, patient ID was first plotted against cell types to obtain a matrix that demonstrated the count of each cell type for every patient. A compositional cell type matrix was also constructed by calculating the proportions of each cell type in every patient using the count matrix. Using these matrices, prediction models were classified using machine learning algorithms to predict if patients would respond to treatment depending on their cell type count and composition. Among the many machine learning algorithms, logistic regression, decision trees and random forests were chosen to be performed on these matrices as these algorithms were deemed to be the most suitable for predicting categorical binary outcomes [3][4].

To determine the best machine learning algorithm for constructing prediction models using single cell data, the accuracy rates of each prediction was used as a metric for comparison. This is usually done by first training the prediction model using a dataset. The resubstitution accuracy rate or testing accuracy will then be obtained by resubstituting the same dataset or an independent dataset into the model. However, the sadeFeldman melanoma dataset only consisted of 29 patients and an independent test set was not available. Thus, a repeated 5-fold cross validation was performed instead to calculate the accuracy rates of each prediction model [5].

As shown in Figure 1, this was done by first dividing the matrices into 5 folds then sampling randomly from the divided matrix. One fold is set aside as the testing set and the remaining folds are used to train the model. The testing set is substituted into the constructed model to obtain an accuracy score. The accuracy score is stored in a vector and the process is repeated for all 5 folds. Once completed, a resample is taken from the divided matrix and the same process was repeated until there was 50 resamples.

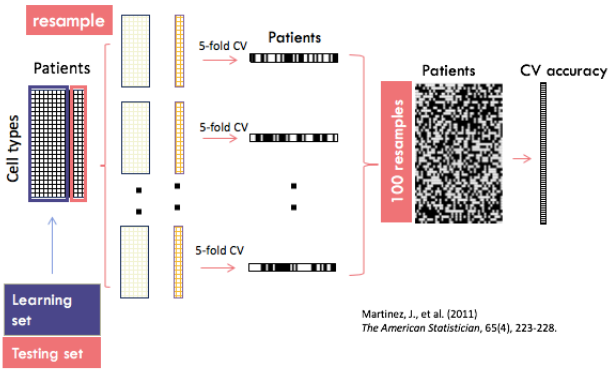


**Figure 1**: Average 5-fold cross validation diagram

Finally, part of the innovation in this project was to compare the accuracy of models constructed using a compositional single cell matrix instead of the traditional gene versus patient matrix (gene expression data) as shown in Figure 2. Thus, the accuracy rates of models constructed with count data, compositional data and gene expression data were compared using summary tables and a boxplot.

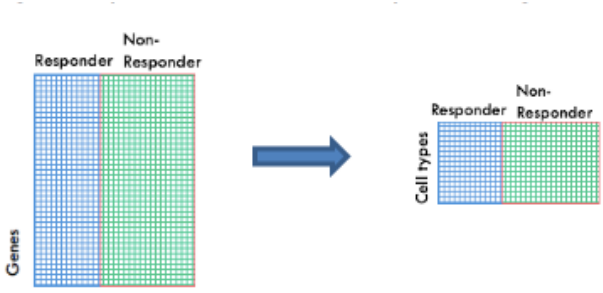


**Figure 2:** Gene expression data versus compositional data diagram

# Results and Discussion

In the box plot shown in Figure 2, it can be seen that logistic resubstitution accuracy rate was perfect across all prediction models. This result led to the realisation that this evaluation score can be quite biased since the model is being evaluated using the same observations used to train model. Otherwise, ignoring the resubstitution rate, it can be observed that random forest is the best-performing machine learning algorithm regardless of the data used. To clarify, the past method mentioned here refers to the accuracy rates calculated using gene expression data, while raw data refers to the accuracy rates calculated using single cell count data.

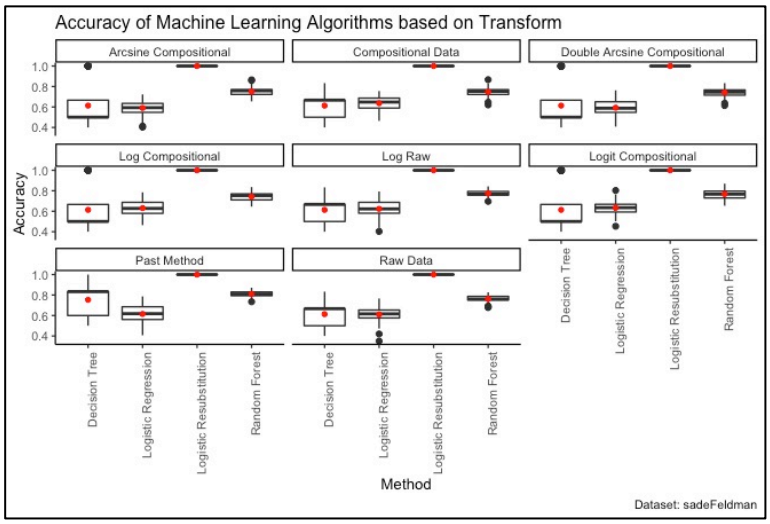


**Figure 2:** Comparison of the accuracy rates of different machine learning algorithms and data transformations

From these summary tables in Figure 3 and 4, it can be inferred that models constructed using gene expression data had the highest accuracy rate for the decision tree and random forest models. This method also provided consistent results considering how its standard deviation is quite small. On the other hand, the compositional data had the highest accuracy rate and the smallest standard deviation for logistic regression. Therefore, it can be concluded that using gene expression data may be a better option for the classification of a random forest or decision tree model, while compositional cell data may be a better option for the construction of a logistic regression model.

| transform | Decision Tree | Logistic Regression | Logistic Resubstitution | Random Forest |
|---|---|---|---|---|
| Arcsine Compositional | 0.69 | 0.60 | 1 | 0.76 |
| Compositional Data | 0.76 | 0.64 | 1 | 0.75 |
| Double Arcsine Compositional | 0.63 | 0.60 | 1 | 0.76 |
| Log Compositional | 0.70 | 0.63 | 1 | 0.74 |
| Log Raw | 0.55 | 0.62 | 1 | 0.77 |
| Logit Compositional | 0.69 | 0.63 | 1 | 0.77 |
| Past Method | 0.81 | 0.63 | 1 | 0.81 |
| Raw Data | 0.58 | 0.61 | 1 | 0.76 |

**Figure 3:** Comparison of accuracy rate means

| transform | Decision Tree | Logistic Regression | Logistic Resubstitution | Random Forest |
|---|---|---|---|---|
| Arcsine Compositional | 0.20 | 0.07 | 0 | 0.04 |
| Compositional Data | 0.17 | 0.06 | 0 | 0.04 |
| Double Arcsine Compositional | 0.19 | 0.07 | 0 | 0.04 |
| Log Compositional | 0.25 | 0.07 | 0 | 0.05 |
| Log Raw | 0.13 | 0.08 | 0 | 0.03 |
| Logit Compositional | 0.13 | 0.06 | 0 | 0.04 |
| Past Method | 0.24 | 0.09 | 0 | 0.03 |
| Raw Data | 0.30 | 0.07 | 0 | 0.03 |

**Figure 4:** Comparison of accuracy rate standard deviation

Finally, it is commonly known that most statistical analysis relies on the normality assumption. However, the distribution of both the count and compositional data used were quite skewed. To tackle this problem, my collaborator, Kelvin proposed 4 transformations for the compositional data, namely log, logit, arcsine and double arcsine to improve the normality of the distributions. He quantified the normality of each distribution after every transformation using the Kolmogorov-Smirnov test then compared them. The value obtained from the test represents how different the distribution is in comparison to the normal distribution. The lower the value, the more similar a distribution was to a normal distribution with the same mean and standard deviation.

As presented in Figure 5, all the transformations did normalise the distributions to some extent. However, the logit transformation had the lowest values for the majority of the cell type distributions, ultimately demonstrating that the logit transformation did best at normalising the distributions. Hence, it was inferred that logit is a great option for normalising compositional data distributions for single cell analysis.

| cell-type | b cell | plasma cell | monocytes | dendritic cells | lymphocytes | exh cd8 cells | regulatory t cells | cytotoxity | exh hs cd8 cells | memory cells | lymphocytes cell cycle |
|---|---|---|---|---|---|---|---|---|---|---|---|
| none | 0.35 | 0.40 | 0.27 | 0.31 | 0.29 | 0.23 | 0.10 | 0.33 | 0.27 | 0.29 | 0.19 |
| log | 0.31 | 0.31 | 0.38 | 0.23 | 0.27 | 0.40 | 0.13 | 0.27 | 0.23 | 0.19 | 0.17 |
| logit | 0.10 | 0.23 | 0.10 | 0.19 | 0.21 | 0.19 | 0.21 | 0.13 | 0.21 | 0.21 | 0.19 |
| arcsine | 0.27 | 0.31 | 0.21 | 0.15 | 0.29 | 0.23 | 0.13 | 0.13 | 0.13 | 0.31 | 0.15 |
| d arcsine | 0.23 | 0.42 | 0.23 | 0.31 | 0.33 | 0.17 | 0.13 | 0.15 | 0.27 | 0.21 | 0.19 |

**Figure 5**: Normality comparison table

# Limitation

Although accuracy was used as the sole metric to evaluate model performance in this project, it is important to note is that have an accuracy rate of 80% for a model does not translate to having an 80% chance of predicting the right outcome for a patient using that the model. The plot in Figure 6 shows the number of accurate predictions out of 100 per patient based on each machine learning algorithm. This plot demonstrates that the model is able to provide the right prediction 100% of the time for some patients and 0% for some other patients, ultimately showing that the model will be biased towards patients of specific cell proportions. Therefore, it is vital to note that accuracy scores are not fully representative of the model performance.

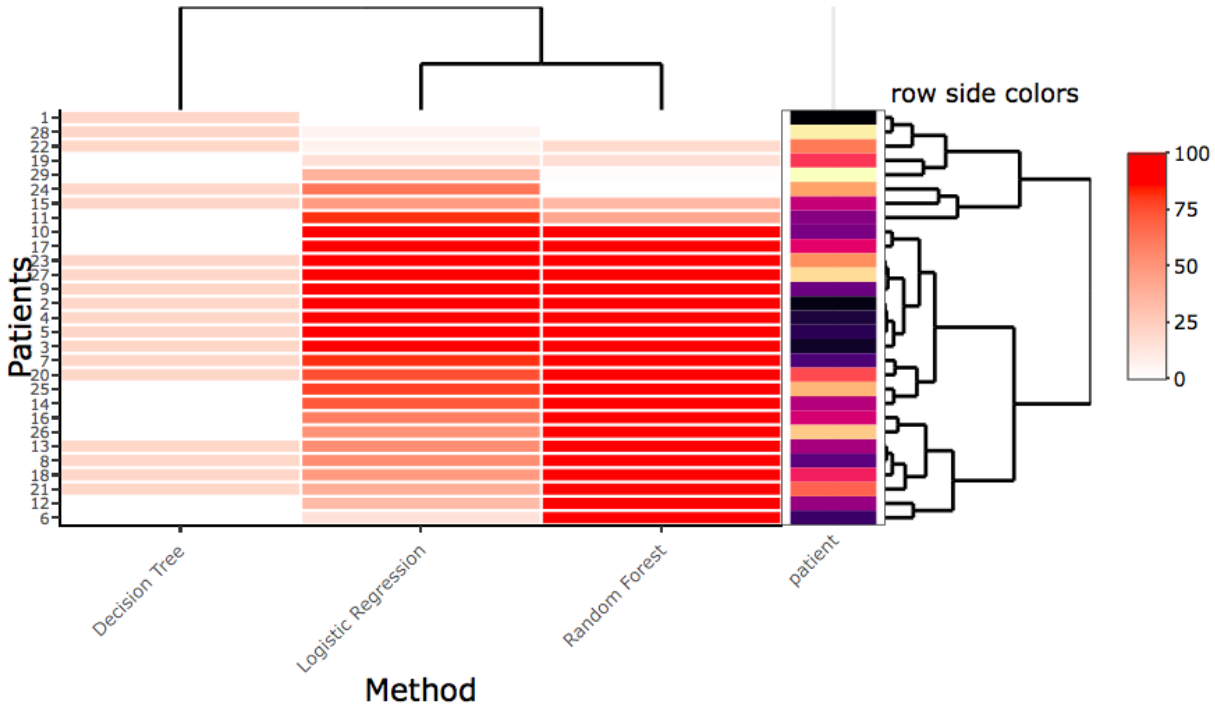


**Figure 6:** Number of accurate predictions per patient based on each machine learning algorithm using raw data

# Conclusion

To summarise, humans are all genetically unique. But due to the advancements in single cell genomics, information from individual cells can now be extracted and the differences between them can be analysed. With machine learning, this information can also be used to construct models that can aid in making better medical decisions. From this project, it can be concluded that random forests provide the highest overall accuracy rate, especially when trained using gene expression data. Compositional single cell data can be normalised best using the logit function and performs best for a logistic regression model.

# References

[1] Sade-Feldman M et al. (2018) Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. URL: https://www.ncbi.nlm.nih.gov/pubmed/30388456?dopt=Abstract

[2] World Economic Forum (no date) How precision medicine will change the future of healthcare. URL: https://www.weforum.org/agenda/2019/01/why-precision-medicine-is-the-future-of-healthcare/

[3] Khandelwal, Renu (2018) Decision Tree and Random Forest. URL: https://medium.com/datadriveninvestor/decision-tree-and-random-forest-e174686dd9eb

[4] Sachan, Lalit (2015) Logistic Regression vs Decision Trees vs SVM: Part II. URL: https://www.edvancer.in/logistic-regression-vs-decision-trees-vs-svm-part2/

[5] Brownlee, Jason (2019) A Gentle Introduction to k-fold Cross Validation. URL: https://machinelearningmastery.com/k-fold-cross-validation/