Chronus Data Analysis Report

By Jayashri V Jagannathan

March 4, 2020

**Goal**

The dataset comprises of multiple signals from isolated pure cells, mixed cells (type1) and mixed\_cells (type 2). The goal of this data analysis is to evaluate the data sets and infer the relationship between the datasets i.e. to use the data from the isolated pure cells to predict how many of each cell type are present in the mixed cell populations.

**Files**

1. chronus\_analysis\_1.ipynb and chronus\_analysis\_2.ipynb – These jupyter notebooks containing data analysis and machine learning models.

2. chronus\_analysis\_3.ipynb – This jupyter notebook contains Decision Tree Model and predictions using the model.

3. mixed\_cell\_type\_1\_result.xls and mixed\_cell\_type\_2\_result.xls – contains the predictions for the classification of the mixed cells for each cell ID obtained from the csv filename .

Note: Since the .ipynb file requires a jupyter installation, I have created the corresponding .pdf file which can be opened on any computer and does not require jupyter installation.

**Data Preparation and Cleaning**

1. Concatenated the data in the .csv files in folders DVV and DV for each pure cell type using the python program given. In this the dv data was scaled by a factor of 10 to match the dvv data. Created a new column called cell\_type to identify the pure cell pertaining to that record in the data. The cell\_type is the dependent variable.
2. Used Python libraries such as pandas and sklearn (scikit\_learn) and jupyter notebook to do this analysis.
3. Looked at the datasets and there was no missing data.
4. Constructed a correlation matrix and noticed that real\_time\_106999 has very weak correlation (between 0 and 0.08 ). So I decided to drop this variable for further analysis.
5. Looked at the output of describe command and the box plot. There does not seem to be an outliers.
6. Used Standard Scaler for preprocessing the independent variables i.e. normalizing each column (x – u)/s where x is the value in a column, u is the column mean and s the column standard deviation.

**Evaluation of different training models**

1. Split the data for pure cells A, B and C into testing and training sets with 30% of the data used for testing.

2. Applied the following Machine Learning Algorithms:

1) Logistic Regression

2) Support Vector Machines

3) Decision Tree Classifier

4) Naive Bayes Classifier

5) K-Nearest Neighbors

6) Gradient Boosting Classifier

3. Tried different values of hyper parameters for Gradient Boosting and K-Nearest Neighbors.

Note: Support Vector Machines did not work at all for this specific dataset. I found online that many people have encountered this same issue.

Ref: <https://datascience.stackexchange.com/questions/989/svm-using-scikit-learn-runs-endlessly-and-never-completes-execution>

4. Evaluated the accuracy score (from sklearn.metrics) of the above models:

|  |  |
| --- | --- |
| Machine Learning Algorithm | Accuracy |
| Logistic Regression | 45.03% |
| Support Vector Machines | Did not work |
| Decision Tree Classifier | 57.65% |
| Naïve Bayes Classifier | 41.16% |
| K-Nearest Neighbors | 57.26% |
| Gradient Boosting Regressor | 53.84% (for testing set was the best) |

5. The Decision Tree Classifier performed the best and so I chose this Machine Learning Algorithm for doing prediction on the Mixed cells Type 1 and Mixed Cells Type 2.

6. Used the Combine\_data.py program for Mixed\_Cells\_Type\_1 and Mixed\_Cells\_Type\_2.

(I kept these separate). Used Standard Scaler on the independent variable and then prediction with the Decision Tree Classifier.

For classifying Mixed Cell Types 1 and 2 into pure cells A, B and C, I used Decision Tree Classifier to predict what cell type is for each data point. Each .csv file is associated with a single cell, and there were multiple data points for each cell. For each data point associated with a mixed cell, the classifier would predict one of the pure cell types A, B, or C. Since there are multiple data points for each mixed cell, there could be different predictions. I used a dictionary data structure to count the predictions of each pure cell type for a specific data type and choose the pure cell type have the maximum count of the predictions. For example, if a mixed cell had 100 data points (say) and 30 data points had predicted cell type to be A, 50 data points had predicted cell type to be B and 20 data points had predicted cell type to be C, I choose cell type to be B because it had the highest frequency of predictions. For detailed information on prediction done on the mixed cell types 1 and 2, please refer to the excel files – mixed\_cell\_type\_1\_result.xls and mixed\_cell\_type\_2\_result.xls.

Actual Numbers of Pure Cells in Mixed Cell Types

|  |  |  |
| --- | --- | --- |
| Description | Mixed Cell Type 1 | Mixed Cell Type 2 |
| Cell A | 1126 | 6839 |
| Cell B | 782 | 1238 |
| Cell C | 406 | 584 |
| Total | 2314 | 8661 |

Percentage of Pure Cells in Mixed Cell Types

|  |  |  |
| --- | --- | --- |
| Description | Mixed Cell Type 1 | Mixed Cell Type 2 |
| Cell A | 48.66% | 78.96% |
| Cell B | 33.79% | 14.29% |
| Cell C | 17.55% | 6.74% |
| Total | 100% | 100% |

**Conclusion**

It will be helpful if we are given information as to how the data is collected such as devices used, and what is the range of the measurements supposed to be so that we can know if the measurements were taken accurately.