Thomas Mossington

Detection of Acute Myeloid Leukemia with Supervised and Unsupervised Machine Learning Models

**Introduction**

One of the most important resources for differentiating between diseased and normal cells is Flow cytometry (FCM). Useful as a diagnostic and research tool, advances in technology have allowed the simultaneous measurement of multiple surface and intracellular markers1,2. In this line, Flowcap2, a project to study a range of FCM data, measures the concentration of the protein in each cell3. Data from this project was downloaded, consisting of FCM data for normal and acute myeloid leukemia (AML) cells3.

The dataset examined contained fluorophore-independent data on both the forward and side scattering of light by the cell. Additionally, fluorescence measurements from 29 proteins were included. This transformed dataset is composed of many properties of the distribution of values across all the cells in each sample. These properties include mean, standard deviation, skewness, and kurtosis. In total, 32 measurements are recorded. The data in this table was used to test both unsupervised and supervised machine learning models in the Orange platform, in order to determine the most accurate model and determine if a sample came from a patient with AML.

**Materials and Methods**

*Data Collection*

Data was obtained from the Flowcap2 project and downloaded as FlowCAP2\_AML\_processed.csv. After uploading to Orange v 3.34.0 4, the file was read through a file widget. In total, this file contained 359 instances and 130 features, with no missing values. Feature 1 was skipped, as this column was just the index column. The Class column was chosen as the target column. The data was parsed into an orange workflow using the file widget. Next, a data table widget was attached to the file, to ensure there was no loss of data in the import. A scatterplot was attached to the data table widget to visualize data clustering and determine the most informative projections on the x and y axes.

**Unsupervised Learning**

The data was not split in this workflow as there is no labels. The file widget was attached to a data table and scatter plot. Two clustering methods were used: k-Means clustering and Hierarchical clustering. The k-Means clustering widget used between 2 and 8 clusters, normalizing columns. 1 Re-runs were allowed and a maximum of 100000 iterations were permitted. A PCA widget was attached, showing the first 20 variables. A scatterplot was attached to the PCA widget to visualize the results.

A distance widget was placed between the file and Hierarchical Clustering widget. The distances between rows was calculated, with a Euclidean distance metric and normalized data. For the Hierarchical Clustering widget, complete linkage was utilized, with a maximum depth of 7. The top 5 samples were selected. This widget was attached to a PCA plot and a scatter plot to visualize the results. Finally, a t-SNE widget was attached directly to the data file. Points on this graph were colored by class.

**Supervised Learning**

*Data Splitting*

In a separate workflow, the data was first divided into testing and training data. A preprocessing widget was attached to the data table widget and rows with missing values were removed. Additionally, the features were normalized by standardizing to μ = 0, σ2 =1. Another data table widget was attached to verify normal behavior. Next, a data sampler widget was attached to divide the data into testing and training. 70% of the data was allocated for training, leaving 30% for testing. This data was then attached to a Test and Score widget.

*Clustering*

Four clustering methods were applied to the split data: Naïve Bayes, kNN, Logistic Regression, and Support Vector Machine (SVM). Each clustering method was attached to a preprocessing widget, which either normalized the features (kNN, Logistic Regression, and SVM) or applied an equal width discretization with 5 intervals (Naïve Bayesian). kNN clustering was set to 5 neighbors, using the Euclidean metric and a uniform weight. Logistic Regression was set to use the Ridge (L2) regularization type. SVM clustering used the SVM type, with a Cost (C) of 1.00, a regression loss epsilon of 0.10, and a linear kernel. All 4 clustering methods were attached to the Test and Score widget as learners.

A confusion matrix and scatter plot matrix were attached to the test and score widget to visualize the results. This model classifies data into either normal or AML samples and examines both cross-validation and separate test data. After this, Logistic Regression was changed to Lasso (L1), in order to reduce the number of features and examine the resulting model performance. The value of c was changed until only the desired number of instances were included.

**Results**

*Data collection*

The data shown here has a lot of variations of the same replicates, which will result in a lot of overlapping data points. However, because there is no loss of features in this data file, there is no restriction on the methods used for characterization and classification.

*Unsupervised Clustering*

A total of 3 unsupervised clustering methods were used: k-Means clustering, Hierarchical Clustering, and t-SNE. Both k-Means clustering, and Hierarchical clustering were filtered through a PCA plot before generation of a scatterplot. Generally, none of these methods did a good job of classifying the data. For k-Means clustering, this model overexpressed C1 (Figure 1). Although there should be roughly equal proportions of C1 and C2, this model has a greater presence of C1. Additionally, this model does a poor job of splitting disease state (AML vs. normal) between the two clusters. Unfortunately, Hierarchal clustering does an even worse job than k-Means clustering at accurately classifying the data. There is almost sole expression of C2 (Figure 2). Additionally, this model still does a poor job of differentiating by disease state.

*A picture containing screenshot, colorfulness

Description automatically generated*

Figure 1 – Scatterplot of k-Means clustering and PCA data. The X axis represents PC1, and the Y axis represents PC2.

A picture containing screenshot

Description automatically generated

Figure 2 - Scatterplot of Hierarchical clustering and PCA data. The X axis represents PC1, and the Y axis represents PC2.

In addition to the k-Means and Hierarchical clustering, a t-SNE method was run on the data. Unfortunately, this method does an equally poor job of clustering the data. Although not split by PCA, this method seeks to cluster data points based on disease state. As seen in Figure 3, however, This model clusters all data points together, regardless of disease state.

A group of red and blue dots

Description automatically generated with low confidence

Figure 3 – t-SNE cluster. Red dots represent normal disease state while blue dots represent AML disease state.

*Supervised Clustering*

Of the four supervised clustering methods used, SVM and Logistic Regression had the best performance, although all four methods had similar performance. When examining the AML training data, SVM had the highest Area Under the Curve (AUC) (0.997 vs. 0.952, 0.967, and 0.995). However, SVM tied with Logistic Regression for highest Class Accuracy (CA) score (0.992) and was tied for the highest precision score (Table 1).

Table 1 – Model Statistics for AML Cross Validated Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AML |  |  |  |  |  |  |  |  |  |
| Model | Train time (s) | Test time (s) | AUC | CA | F1 | Precision | Recall | LogLoss | Specificity |
| Naive Bayes | 0.328 | 0.101 | 0.967 | 0.933 | 0.773 | 0.659 | 0.935 | 1.049 | 0.932 |
| kNN | 0.362 | 0.188 | 0.952 | 0.937 | 0.652 | 1.000 | 0.484 | 0.485 | 1.000 |
| SVM | 0.533 | 0.243 | 0.997 | 0.992 | 0.967 | 1.000 | 0.935 | 0.038 | 1.000 |
| Logistic Regression | 0.237 | 0.086 | 0.995 | 0.992 | 0.967 | 1.000 | 0.935 | 0.032 | 1.000 |

Similarly, when examining the Normal data, SVM and Logistic Regression still ranked highest. SVM’s AUC score was 0.997, compared to 0.937, 0.967, and 0.995. Similarly, the CA scores were unchanged (0.992/0.992 vs 0.937 and 0.933). Finally, SVM was again matched for precision, at 0.991 vs. Logistic Regression’s 0.991 (Table 2).

Table 2 – Model Statistics for Normal Cross Validated Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Normal |  |  |  |  |  |  |  |  |  |
| Model | Train time (s) | Test time (s) | AUC | CA | F1 | Precision | Recall | LogLoss | Specificity |
| kNN | 0.362 | 0.188 | 0.952 | 0.937 | 0.965 | 0.932 | 1.000 | 0.485 | 0.484 |
| Naive Bayes | 0.328 | 0.101 | 0.972 | 0.933 | 0.960 | 0.990 | 0.932 | 1.049 | 0.935 |
| SVM | 0.533 | 0.243 | 0.997 | 0.992 | 0.995 | 0.991 | 1.000 | 0.038 | 0.935 |
| Logistic Regression | 0.237 | 0.086 | 0.995 | 0.992 | 0.995 | 0.991 | 1.000 | 0.032 | 0.935 |

Figures 4-7 paint a clear picture of the performance difference between the four models. Logistic Regression does the best job of clustering the AML and normal data points (Figure 4), and SVM performs nearly as well (Figure 5). Naïve Bayes (NB) does a poor job of filtering by color, marking several normal data points in AML (Figure 6). Similarly, kNN labeling does a poor job of filtering by color as well (Figure 7).

A picture containing screenshot, diagram, plot, colorfulness

Description automatically generated

Figure 4 - Scatterplot of Logistic Regression data. Normal data points are represented in red while AML data points are represented in blue. Similarly, x points represent normal data while circle points represent AML data.

A picture containing screenshot, diagram, plot

Description automatically generated

**Figure 5**- Scatterplot of SVM data. Normal data points are represented in red while AML data points are represented in blue. Similarly, x points represent normal data while circle points represent AML data.

A picture containing screenshot, diagram

Description automatically generated

**Figure 6** – Scatterplot of Naïve Bayes data. Normal data points are represented in red while AML data points are represented in blue. Similarly, x points represent normal data while circle points represent AML data.

**A diagram of red and blue dots

Description automatically generated with low confidence**

**Figure 7** – Scatterplot of kNN data. Normal data points are represented in red while AML data points are represented in blue. Similarly, x points represent normal data while circle points represent AML data.

Using the Find Informative Projections function, the best data subset for fitting vs. final evaluation was subset 2, so this subset was used. The goal of this supervised model is to accurately separate disease samples and non-disease samples. The performance of Logistic Regression shows a reasonable performance in this facet. However, because the values in Tables 3 and 4 represent lower values for the testing data than the training data, this indicates a very slight degree of overfitting. However, this model generally does a good job of separating data points on the basis of disease status.

Table 3 - Model Statistics for AML Test Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AML |  |  |  |  |  |  |  |  |  |
| Model | Train time (s) | Test time (s) | AUC | CA | F1 | Precision | Recall | LogLoss | Specificity |
| Naive Bayes | 0.063 | 0.012 | 0.981 | 0.944 | 0.800 | 0.667 | 1.000 | 1.429 | 0.937 |
| SVM | 0.052 | 0.025 | 0.999 | 0.991 | 0.960 | 0.923 | 1.000 | 0.033 | 0.989 |
| Logistic Regression | 0.024 | 0.008 | 0.999 | 0.991 | 0.960 | 0.923 | 1.000 | 0.020 | 0.989 |
| kNN | 0.038 | 0.021 | 0.914 | 0.972 | 0.857 | 1.000 | 0.750 | 0.685 | 1.000 |

Table 4 – Model Statistics for Normal Test Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Normal |  |  |  |  |  |  |  |  |  |
| Model | Train time (s) | Test time (s) | AUC | CA | F1 | Precision | Recall | LogLoss | Specificity |
| kNN | 0.038 | 0.021 | 0.914 | 0.972 | 0.984 | 0.969 | 1.000 | 0.685 | 0.750 |
| Naive Bayes | 0.063 | 0.012 | 0.987 | 0.944 | 0.967 | 1.000 | 0.937 | 1.429 | 1.000 |
| SVM | 0.052 | 0.025 | 0.999 | 0.991 | 0.995 | 1.000 | 0.989 | 0.033 | 1.000 |
| Logistic Regression | 0.024 | 0.008 | 0.999 | 0.991 | 0.995 | 1.000 | 0.989 | 0.020 | 1.000 |

Additionally, confusion matrices were used to examine the rate of false positives and false negatives (Figures 8-12). Overall, SVM and Logistic Regression had the lowest rates of false positives (0 vs 15) and false negatives (2 vs 16). The Naïve Bayes model possessed 15 false positives and 2 false negatives, while the kNN model possessed 0 false positives and 16 false negatives, making these two models inferior to SVM and Logistic Regression.

A screenshot of a cell phone

Description automatically generated with low confidence

**Figure 8** – Confusion Matrix results of Naïve Bayes. False positives are seen in the bottom left corner (15) and false negatives are seen in the top right corner (2).

A screenshot of a cell phone

Description automatically generated with medium confidence

**Figure 9** - Confusion Matrix results of kNN False positives are seen in the bottom left corner (0) and false negatives are seen in the top right corner (16).

A screenshot of a graph

Description automatically generated with low confidence

**Figure 10** - Confusion Matrix results of SVM. False positives are seen in the bottom left corner (0) and false negatives are seen in the top right corner (2).

A screenshot of a cell phone

Description automatically generated with low confidence

**Figure 11** - Confusion Matrix results of Logistic Regression (Ridge). False positives are seen in the bottom left corner (0) and false negatives are seen in the top right corner (2).

To explore the possibility of feature reduction without compromising the performance of the Logistic Regression model, both informative projections and Lasso regularization were utilized. Informative Projects reported that the most optimal instances were CD16-PC5 mom\_2 and CD34-PC5 mom\_2. This was then confirmed by performing Lasso regularization with the Logistic Regression model, and c = 0.018. In total, all features were removed except CD10-PC7, CD34-PC5, and CD16-PC5 (mom\_2 and mom\_3). The confusion matrix for Logistic Regression with Lasso Regularization showed an increase in the false positive rate generated (Figure 12).

Unfortunately, there was a decrease in table performance, as the model had a harder time differentiating on the basis of class. The values for this feature reduction are included in Table 5. As can be seen, there is significant decrease in AUC and CA. Additionally, Figure 13 shows a scatterplot with this model, with decreased performance over the initial model. Therefore, the value of C was increased to 0.40. At this value of C, there was 1 target with 13 instances. The false positive count dropped to 0 (Figure 14). Sample statistics, such as AUC and CA remain at pre-adjustment levels (Table 6), and the scatterplot returned to normal clustering (Figure 15). This model is the greatest feature reduction without compromising model performance.

A screenshot of a graph

Description automatically generated with low confidence

Figure 12 - Confusion Matrix results of Logistic Regression (Lasso, C = 0.018). False positives are seen in the bottom left corner (37) and false negatives are seen in the top right corner (1).

A picture containing screenshot, plot, diagram, line

Description automatically generated

**Figure 13** – Scatterplot showing AML vs. normal tissue for CD16-PC5 mom\_2 and CD34-PC5 mom\_2. Color is indicated by Logistic Regression (Lasso, C = 0.018) and shape is indicated by class.

A screenshot of a graph

Description automatically generated with low confidence

Figure 14 - Confusion Matrix results of Logistic Regression (Lasso, C = 0.40). False positives are seen in the bottom left corner (0) and false negatives are seen in the top right corner (1).

A picture containing plot, diagram, screenshot

Description automatically generated

Figure 15 – Scatterplot showing AML vs. normal tissue for CD16-PC5 mom\_2 and CD34-PC5 mom\_2. Color is indicated by Logistic Regression (Lasso, C = 0.40) and shape is indicated by class.

Table 5 - Model Statistics for AML Train Data, Lasso Regularized Logistic Regression, C = 0.018

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AML |  |  |  |  |  |  |  |  |  |
| Model | Train time (s) | Test time (s) | AUC | CA | F1 | Precision | Recall | LogLoss | Specificity |
| Logistic Regression | 0.323 | 0.137 | 0.983 | 0.849 | 0.612 | 0.448 | 0.968 | 0.572 | 0.833 |
| Naive Bayes | 0.311 | 0.100 | 0.967 | 0.933 | 0.773 | 0.659 | 0.935 | 1.049 | 0.932 |
| kNN | 0.346 | 0.182 | 0.952 | 0.937 | 0.652 | 1.000 | 0.484 | 0.485 | 1.000 |
| SVM | 0.845 | 0.407 | 0.997 | 0.992 | 0.967 | 1.000 | 0.935 | 0.040 | 1.000 |

Table 6 – Model Statistics for AML Train Data, Lasso Regularized Logistic Regression, C = 0.40

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AML |  |  |  |  |  |  |  |  |  |
| Model | Train time (s) | Test time (s) | AUC | CA | F1 | Precision | Recall | LogLoss | Specificity |
| kNN | 0.346 | 0.182 | 0.952 | 0.937 | 0.965 | 0.932 | 1.000 | 0.485 | 0.484 |
| Naive Bayes | 0.311 | 0.100 | 0.972 | 0.933 | 0.960 | 0.990 | 0.932 | 1.049 | 0.935 |
| SVM | 0.845 | 0.407 | 0.997 | 0.992 | 0.995 | 0.991 | 1.000 | 0.040 | 0.935 |
| Logistic Regression | 0.323 | 0.137 | 0.983 | 0.849 | 0.906 | 0.995 | 0.833 | 0.572 | 0.968 |

Using Informative Projections, two important proteins that are important features for differentiating normal and AML cells were uncovered. These were CD16 and CD34, two proteins that have been implicated in AML. CD16, also known as FCG3B, is a receptor for the Fc region of immunoglobulins gamma 5. This protein functions to bind complexed or aggregated IgG and monomeric IgG 5. Natural Killer (NK) cells with overexpressed CD16 are known to have a poor effect on clinical outcome and survival for patients with AML 6.

Similarly, CD34 is a possible adhesion molecule with a role in early hematopoiesis, linked to AML 5,7. Previous research has shown that leukemic cells inhibit CD34+ hematopoietic stem and progenitor cells 8. Because of this inhibition, there is reduced hematopoiesis in AML 8.

**Conclusions**

Examining 3 unsupervised learning models and 4 supervised learning models on this disease vs. normal data provided some telling results on the strengths of each model. The 3 unsupervised models failed to accurately cluster the data, being found unfit for this data type. However, the 4 supervised learning models all performed quite well. SVM and Logistic Regression were the two best performing model, matching each other in performance in nearly every regard. The use of informative projections revealed two features that were crucial for identifying between AML and normal cells: CD16-PC5 and CD34-PC5. When the Logistic Regression model was modified with Lasso Regularization at C = 0.018, these two identifiers (and CD10-PC5) were the only ones remaining.

However, there was significant decrease in model performance at this level, so the value of C was increased to C = 0.40, the last value before a decrease in model performance. This model matched all determined model output levels, while reduced the number of features to 1 with 13 instances, including CD16-PC5 and CD34-PC5. This reduction revealed the 1 feature remaining was from normal tissue. Overall, this fine-tuned model helped identify multiple proteins that are crucial features for deciding between normal and AML cells, therefore succeeding at the desired task.

Reference List

1. Querec TD, Akondy RS, Lee EK, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat Immunol*. Jan 2009;10(1):116-125. doi:10.1038/ni.1688

2. Bendall SC, Simonds EF, Qiu P, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*. May 6 2011;332(6030):687-96. doi:10.1126/science.1198704

3. Aghaeepour N, Finak G, Hoos H, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods*. Mar 2013;10(3):228-38. doi:10.1038/nmeth.2365

4. Demsar J CT, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*. 2013;14:2349-2353.

5. Consortium TU. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*. 2022;51(D1):D523-D531. doi:10.1093/nar/gkac1052

6. Chretien AS, Devillier R, Granjeaud S, et al. High-dimensional mass cytometry analysis of NK cell alterations in AML identifies a subgroup with adverse clinical outcome. *Proc Natl Acad Sci U S A*. Jun 1 2021;118(22)doi:10.1073/pnas.2020459118

7. Mattes K, Gerritsen M, Folkerts H, et al. CD34(+) acute myeloid leukemia cells with low levels of reactive oxygen species show increased expression of stemness genes and can be targeted by the BCL2 inhibitor venetoclax. *Haematologica*. 2020:e399-e403. vol. 8.

8. Jäger P, Geyh S, Twarock S, et al. Acute myeloid leukemia-induced functional inhibition of healthy CD34+ hematopoietic stem and progenitor cells. *Stem Cells*. Sep 2021;39(9):1270-1284. doi:10.1002/stem.3387