Assignment Report

Tanveer Sharma

EE 769

Down Syndrome Prediction from Mice Protein Expression Levels

1 Introduction

Down syndrome, also known as trisomy 21, is a genetic disorder characterized by the presence of an extra copy of chromosome 21. It is the most common chromosomal abnormality in humans, affecting approximately 1 in 800 live births worldwide. Individuals with Down syndrome often exhibit distinct physical features, developmental delays, and intellectual disabilities. Early diagnosis and intervention play a crucial role in improving the quality of life for individuals with Down syndrome.

Advancements in biotechnology and data science have opened up new avenues for diagnosing genetic disorders using molecular data. This project aims to leverage machine learning techniques to predict Down syndrome in mice based on protein expression levels. The rationale behind using mice models is to gain insights into the underlying genetic and molecular mechanisms that contribute to Down syndrome, which could potentially inform research on human cases.

2 Methods Employed

2.1 Data Collection and Preprocessing

The dataset used in this study comprises protein expression levels of mice samples, with features derived from various protein assays. The initial step involved assessing the dataset's integrity by examining the number of samples in each column and identifying data types. The first 77 columns were found to contain float values representing protein expression, while the last two columns contained categorical string data indicating class labels.

To address missing data, a comprehensive data imputation strategy was devised. While one approach could involve replacing missing values with the mean of the entire dataset, this could introduce bias and adversely affect downstream analyses. Instead, we opted for iterative imputation, which takes into account the relationships between features and iteratively predicts missing values based on existing information.

2.2 T-Test on Binary Classification Data

To identify significant variables for prediction, a t-test was performed on the binary classification data. This involved comparing means between two groups: samples predicted to have Down syndrome and samples predicted to be unaffected. Columns exhibiting high correlation were scrutinized further. Among correlated pairs, one variable was retained, and the other was dropped to mitigate multicollinearity issues that can hinder model interpretability and performance.

2.3 Iterative Imputation on Train Data and Test Data

The process of iterative imputation was applied to both the training and testing datasets. Missing values were predicted and filled iteratively based on relationships established within the respective dataset. This approach enhances the integrity of the dataset while minimizing the introduction of artificial bias that may arise from imputing missing values with generic estimates.

2.4 Recursive Feature Elimination with Cross-Validation (RFECV)

Feature selection is crucial to enhance the model's performance and interpretability. Recursive Feature Elimination with Cross-Validation (RFECV) is a powerful technique that iteratively removes less important features while cross-validating the model's performance. This iterative process helps identify the most relevant features, leading to better generalization and reduced risk of overfitting.

RFECV involves the following steps:

- Initially, all features are considered for the model.
- The model's performance is evaluated using cross-validation, often with metrics like accuracy or F1-score.
- The least important feature(s) are removed, and the model is evaluated again.
- This process continues iteratively until the specified number of features is reached or the model's performance plateaus.
- The optimal subset of features is determined based on the point at which performance no longer improves.

In our study, RFECV was employed to identify crucial proteins for Down syndrome prediction. By eliminating non-informative features, RFECV not only enhances the model's performance but also offers insights into the biological relevance of certain proteins in predicting the condition.

3 Results and Discussion

3.1 Classification Metrics

For evaluating model performance, various classification metrics were used. In binary classification, metrics like accuracy, precision, recall, and F1 score were employed. For multiclass classification, accuracy, macro/micro-averaged F1 scores, weighted F1 score, and Cohen's Kappa were used.

3.2 Neural Network and Feature Importance

A neural network with a single ReLU hidden layer and Softmax output was used for classification. Feature importance was calculated using methods such as coefficient analysis, decision tree-based measures, permutation feature importance, and SHAP values.

3.3 Model Comparison

Models were trained and tuned for various algorithms including Linear SVM, RBF kernel SVM, Neural Network, and Random Forest. The impact of feature selection on model performance was evaluated. Comparison of models before and after feature selection was done to observe improvements.

4 Conclusion

The project demonstrated the potential of machine learning in predicting Down syndrome using mice protein expression levels. Different classification algorithms were applied and evaluated based on various metrics. The results highlighted the importance of feature selection for model performance. Further research could focus on refining feature engineering and exploring more complex model architectures.

5 References

1. Prof. Amit Sethi, Electrical Engineering, IIT Bombay.
2. Scikit-learn Documentation: `https://scikit-learn.org/stable/documentation.html`
3. Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. Springer.