Predictive Modeling of Mouse Alcoholism using Biomedical 3D Array Data and Machine Learning

Maha Kanakala

Department of Genetics, Rutgers University, Piscataway, New Jersey 08854

RUTGERS
Douglass Residential College
WOMEN IN SCIENCE AND ENGINEERING

Abstract

This research investigates the application of machine learning algorithms on a 3D array dataset to predict alcohol consumption in mice, using machine learning as a powerful and analytical tool for data analysis. In biomedical engineering, where **time** plays a crucial role in a living organism's development and behavior, this study highlights the importance of temporal information in alcohol research. Using Python-based coding techniques, the project aims to identify patterns with 3D array datasets that can be interpreted by both computer languages and human operators, contributing to improved understanding of alcohol consumption-related behaviors in mice for medical and research applications. The final model predicts the likelihood of a mouse given its 7 features, whether it will become alcoholic or not- alcoholic.

Introduction

In the dynamic realm of biomedical engineering and medicine, where precision and innovation converge, the synergy between cutting-edge data analysis and transformative technology is redefining the landscape of research and clinical practice. The journey towards enhancing human health has evolved into an intricate interplay of meticulous data analysis and advanced technological tools, birthing a new era of discovery. In the pulsating heart of our scientific pursuit lies the recognition of the overwhelming complexities present within biomedical data. As datasets expand in scale and intricacy, the traditional tools of analysis prove *inadequate*, necessitating the integration of cutting-edge methodologies. It is within this context that **machine learning** emerges as a guiding light, with its capacity to decipher hidden correlations and patterns, thereby enriching our understanding of biological phenomena. Our project is an embodiment of this exploration, as we seek to harness the data-unveiling prowess of machine learning to unravel the enigmatic relationships and patterns enshrouding biomedical data.

This study aims to leverage machine learning algorithms for the analysis of a 3D array dataset to predict alcohol consumption in mice. Through the application of advanced temporal data analysis techniques, including LSTM-based modeling and correlation analysis, the project seeks to uncover intricate patterns and trends within the multi-dimensional data. The primary objective is to enhance our understanding of alcohol-related behaviors in mice, contributing to both medical research applications and broader insights into temporal data analysis methodologies.

Each mouse was housed in an individual cage, with food and liquid available ad libitum. For the chronic alcohol drinking experiment, the paradigm of two-bottle free-choice was used (this picture did not show the second bottle). Animals' body weight and food consumption were measured daily on weekdays, while the amount of water or alcohol consumption was recorded daily.

QR Code for Github Repository with code and raw data.

Methods

Data Preprocessing:
Our first step involved the comprehensive preprocessing of the raw biomedical data. This included removing duplicates, null values and organizing the data better for feature extraction. The data was also simplified by normalizing the Mouse category from 4 distinct groups to 2 groups for preliminary model training:

```
# Find the index level corresponding to 'Mouse Category' column
mouse_category_index = df.columns.get_loc(('Unnamed: 0_level_0', 'Mouse Category'))

# Define a function to map Mouse Category to 0 (Non-Alcoholic) or 1 (Alcoholic)
def map_alcoholic_to_binary(row):
    if 'Group-2' in row:
        return 0  # Non-Alcoholic
    else:
        return 1  # Alcoholic

# Create the new 'Alcoholic' column using the map_alcoholic_to_binary function
df['Alcoholic'] = df.iloc[:, mouse_category_index].apply(map_alcoholic_to_binary)
```

Constructing the 3D Array:
We designed a 3D array with dimensions corresponding to the number of subjects, time steps, and features. This multidimensional structure provided a comprehensive overview of the temporal evolution of each feature for every subject. For each mouse, we populated the 3D array with the data from the selected features. This involved iterating through the mouse column and corresponding 4 time steps, extracting the relevant data from the specified columns, and storing it within the appropriate dimensions of the array. The graphic below shows the 3d array, with each mouse, 4 time stamps, and its feature values.

Array Value	Feature
2.43	Daily ethanol intake (g/kg/day)
4.15	Daily total liquid
4.95	Ethanol preference (100*daily ethanol/daily total liquid)
73.45	Daily liquid/weight (1000*ml/kg)
27.00	Daily food intake/weight
209.26	Mouse Category
5.50	Mouse #

Mouse #1 Time Stamp 1: [2.43, 4.15, 4.95, 27.00, 73.45, 209.26, 5.50]
Mouse #1 Time Stamp 2: …
Mouse #1 Time Stamp 3: …
Mouse #1 Time Stamp 4: …

Data Transformation and Enrichment:
Beyond the construction of the 3D array, we recognized the significance of mapping the categorical "Mouse Category" attribute into a binary format suitable for machine learning. By systematically transforming "Alcoholic" and "Non-Alcoholic" categories into binary values (1 and 0, respectively), we enriched the dataset with a target variable amenable to predictive modeling.

Model Definition:
To harness the predictive potential of the enriched 3D array, we defined a predictive model that incorporated both a Bidirectional Long Short-Term Memory (LSTM) layer and a Deep Neural Network (DNN) architecture. This hybrid model aimed to capture both the temporal dependencies and intricate relationships within the 3D array data.

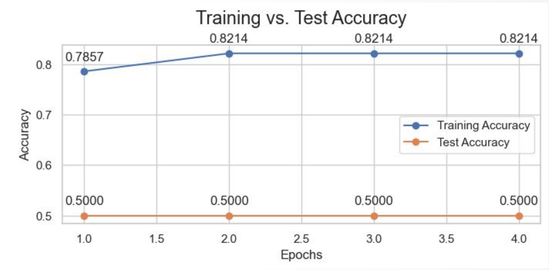
The architecture commenced with an input layer corresponding to the dimensions of the 3D array. A Bidirectional LSTM layer was employed to capture temporal dependencies in both forward and backward directions. Subsequently, a DNN layer was introduced to further distill intricate patterns. The output layer was designed as a sigmoid activation unit, enabling binary classification (0 or 1; alcoholic or non-alcoholic).

Results

The predictive model showcased both promising training accuracy and highlighted challenges in generalization to unseen data, shedding light on the balance between fitting to the training set and attaining real-world applicability.

- **Training Accuracy:** After the fourth epoch, our model achieved an impressive training accuracy of **82.14%**, signifying its capacity to learn patterns from the training data as it increases from each consecutive epoch.
- **Test Accuracy Discrepancy:** However, the model's performance on the test data was notably lower, with a test accuracy of only **50%**. This discrepancy raises concerns about overfitting, where the model becomes too closely tailored to the training data, limiting its ability to generalize to new, unseen data.

The divergence between training and test accuracies illuminates the importance of addressing **overfitting** to enhance the model's real-world utility. This analysis reinforces the significance of evaluating models on distinct datasets to ensure their effectiveness in practical applications.



Conclusions

Through the integration of advanced data preprocessing and predictive modeling, our project exemplifies the potential of machine learning in biomedical research. By navigating the intricate realm of 3D array data structures, we have sought to uncover hidden insights within complex biomedical datasets.

Future Directions

1. Normalization of Data: One potential avenue for improvement is the normalization of data. Presently, the model operates on raw data, but investigating the advantages of scaling all features to a consistent range, such as 0 to 1, could potentially enhance model performance. Normalization has only been applied to the Mouse Category in building the model thus far. This normalization process can help mitigate the influence of feature magnitude disparities, leading to more stable and effective training.
2. Enhancing Code Usability: A critical goal is to make our code adaptable for various datasets. To achieve this, we plan to create a modular and well-documented codebase. Modularity is achieved by separating the Jupyter notebook file into data processing, modeling, evaluation modules where an xlsx file in a standardized format is fed in.

References

1. Brownlee, Jason. "Multivariate Time Series Forecasting with Lstms in Keras." *MachineLearningMastery.Com*, 20 Oct. 2020, machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/.
2. Dr. Niraj Kumar. *How to Use LSTM with 1D, 2D and 3D Array?*, 8 Feb. 2021, supremewinnercom.wordpress.com/2020/05/09/using-lstm-with-1d-2d-and-3d-array/.
3. Sharma, Vaibhav. "Vaibhav Sharma." *Pluralsight*, 17 May 2019, www.pluralsight.com/guides/deep-learning-model-perform-binary-classification.

Acknowledgments

I extend my sincere thanks to Professor Lei Yu for their invaluable guidance throughout this research project. I am grateful to Douglass Residential College, Project SUPER, and Dr. Jo for providing resources and opportunities that have enriched my academic journey.