Recurrent Neural Networks for Classifying Human Embryonic Stem Cell-Derived Cardiomyocytes

Carolina Pacheco and René Vidal

Center for Imaging Science, Mathematical Institute for Data Science, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, USA

Motivations

- Classifying human embryonic stem cell-derived cardiomyocytes (hESC-CMs) is important in cardiac regenerative medicine to reduce the risk of complications with stem cell therapy.
- Adult CMs can be classified based on the characteristic shape of their action potentials (APs), but the understanding of how the shape of hESC-CM APs relates to that of adult CMs remains limited [4].
- Prior approaches to hESC-CM APs classification either discard most of the information contained in the APs or rely on distances that are computationally expensive [3].
- Our goal is to develop a classifier of hESC-CM APs applicable to large-scale datasets.

Clustering Quality Index

- Davies-Bouldin Index (DBI) [1]. The DBI between two clusters \( \Omega_a = \{x_i\} \) and \( \Omega_b = \{y_j\} \) is the ratio between intra-cluster dispersion and distance between clusters:
  \[
  DBI(\Omega_a, \Omega_b) = \frac{\sum_{i} \sum_{j} d(x_i, y_j)}{\min(\Omega_a, \Omega_b)}
  \]
  where \( S_a \) is the mean distance from elements of class \( a \) to the average signal of the same class, and \( M_a \) is the distance between the average signals of both classes.

Contributions

- We propose a new method for classifying hESC-CM APs based on recurrent neural networks (RNNs) with long short-term memory (LSTM) units.
- The learning approach exploits the abundance of labels for adult APs, which can be obtained via simulation of electrophysiological models of the typical adult phenotypes.
- Our semi-supervised approach uses a novel loss function that combines a supervised classification loss for adult APs and an unsupervised contrastive loss for hESC-CM APs.

Problem Formulation

- Set of unlabeled hESC-CM APs \( \Omega_a = \{\mathbf{x}_i\}_{i=1}^{N_a} \); each \( \mathbf{x}_i \) is a time-series of length \( K \).
- Set of labeled adult CM APs \( \Omega_b = \{\mathbf{y}_j\}_{j=1}^{N_b} \); each \( \mathbf{y}_j \) is a time-series of length \( K \) labeled as atrial (\( y = 0 \)) or ventricular (\( y = 1 \)).
- Problem: Assign a label \( y' \) to a new \( \mathbf{x} \), where \( y' \) denotes atrial-like and \( y' = 1 \) denotes ventricular-like.

Classifier Architecture

- RNN with LSTM units as a classifier:
  - Hidden layer: LSTM of dimension \( p = 3 \)
  - Output layer: Sigmoid unit \( \sigma(z) = \frac{1}{1 + e^{-z}} \)
- Computational time for classifying a new sample:
  - 1NN M: 6.74 sec/sample in 2 8-core computer nodes with 8 2.3GHz CPUs per node.
  - Semi-LSTM: < 6 sec for the whole 6940 APs dataset in one 2.2 GHz CPU with 2 cores.

Semi-Supervised Loss Function

The supervised part guides the LSTM to correctly predict labels of adult CMs, while the unsupervised part guides the LSTM to predict the same label for similar embryonic CMs:

- Supervised part:
  \[
  \text{Loss}_{\text{sup}} = -\sum_{i=1}^{N_a} \log \sigma(y_i') + \lambda \sum_{j=1}^{N_b} \ell(\mathbf{y}_j; \mathbf{y}_j')
  \]
  where \( y_i' \) is the predicted label.
- Contrastive unsupervised loss on hESC-CM APs:
  \[
  \ell(\mathbf{x}_i; \mathbf{x}_j') = s_{xy} \cdot \ell(\mathbf{y}_j; \mathbf{y}_j') + (1 - s_{xy}) \cdot \ell((1 - \mathbf{y}_j); \mathbf{y}_j')
  \]
  where \( s_{xy} \) is the correlation between \( \mathbf{x}_i \) and \( \mathbf{x}_j \).

Conclusions

- Experiments confirm the benefits of integrating information from both adult and embryonic cardiomyocytes in a semi-supervised learning scheme for hESC-CMs classification.
- The proposed semi-supervised approach uses the Euclidean metric more effectively than previous methods, outperforming the INN scheme.
- Proposed semi-supervised approach gives results similar to the state-of-the-art (94.75% of agreement) with clear computational advantages when applied to new samples.

Acknowledgments:

The authors thank Dr. Giann Gorospe for insightful discussions, Dr. Benjum Zhin and Prof. Leslie Tong for providing the hESC-CMs dataset.

References: